OCR: i progetti di digitalizzazione e il riconoscimento ottico dei caratteri

di Markus Brantl e Tommaso Garosci

Come è noto, i programmi software che traducono automaticamente un file immagine di testo nel corrispondente valore ASCII o UNICODE[1], normalmente contenuto in *files.txt*, vengono chiamati OCR (Optical Character Recognition). La maggior parte di questi programmi lavora utilizzando logiche *fuzzy* che confrontano la forma delle lettere costituita da una matrice di punti bianchi e neri con immagini che sono stati istruiti a riconoscere. Quando viene stabilita una corrispondenza, la lettera prescelta viene memorizzata nel testo. Se la corrispondenza si presenta a un basso livello di confidenza, cioè di probabilità di essere stata riconosciuta correttamente, il programma propone comunque la lettera segnalando l'incertezza. Il software che svolge l'operazione di riconoscimento si chiama motore. I programmi ne possono avere uno o più. Nel secondo caso il loro software è integrato da un meccanismo di voto che sceglie quale lettera presenta il più alto livello di confidenza. Normalmente questo sistema minimizza il potenziale di errore dei programmi che si basano su un solo motore[2].

Gli OCR hanno alle spalle una lunga storia[3] e la loro precisione nel riconoscimento di caratteri latini a stampa ha raggiunto una elevata affidabilità[4]. Per questa ragione sono considerati una tecnologia ormai matura, sebbene continui la ricerca per migliorarli sotto molti aspetti. In larga misura si tratta di software integrato allo scanner, ma esistono prodotti più sofisticati acquistabili separatamente dall'hardware. Il loro prezzo può variare significativamente, così come sono diverse le funzionalità che essi offrono. Per questo la scelta del programma OCR

MARKUS BRANTL, Münchener Digitalisierungszentrum, Bayerische Staatsbibliothek, Monaco di Baviera, e-mail markus.brantl@bsb-muenchen.de
TOMMASO GAROSCI, BESS e Ires, Istituto Ricerche Economico Sociali del Piemonte, Via Nizza 18, 10124 Torino, e-mail garosci@ires.piemonte.it
Ultima consultazione siti web: 27 ottobre 2008.

1 <http://www.unicode.org>.

2 <http://www.tei-c.org/About/Archive_new/ETE/Preview/duggan.xml>, Electronic textual editing: effective methods of producing machine-readable text from manuscript and print sources [Eileen Gifford Fenton (JSTOR) and Hoyt N. Duggan (University of Virginia)].

3 <http://en.wikipedia.org/wiki/Optical_character_recognition> traccia una breve, ma informata storia della loro evoluzione.

4 Michael Lesk, *Understanding digital libraries*, 2. ed. Amsterdam: Morgan Kaufmann, 2005. Cita a p.55 i test dell'Università del Nevada, di cui parleremo più avanti, per segnalare i progressi realizzati dal software in questo campo soprattutto per testi chiari su carta di qualità elevata.

è assai importante sia in termini di qualità finale del prodotto sia di efficienza dei processi produttivi[5].

Le buone pratiche e l'OCR

Quasi vent'anni fa Michael Lesk sosteneva che per testi chiari e uniformi gli OCR disponibili all'epoca erano adeguati, anche se la loro velocità era ovviamente assai inferiore a quella dei prodotti odierni[6]. Oggi la manualistica e la letteratura sul tema del riconoscimento ottico dei caratteri specificamente rivolte ai bibliotecari sono singolarmente avare di indicazioni, perché l'affidabilità dei programmi attuali è, come si è detto, considerata oramai acquisita. Un'altra ragione probabile per trascurare l'argomento è che di norma il software è incorporato nell'hardware. Esso non viene acquistato separatamente e rimane, per così dire, nascosto nel processo o comunque non oggetto di scelta di acquisto. Sempre più spesso, e particolarmente per i progetti di un certo rilievo, tale software viene invece acquistato appositamente e quindi è necessario che le sue prestazioni e la sua adattabilità alle necessità vengano valutate esplicitamente.

Le informazioni fornite da istituzioni che esemplificano o illustrano le pratiche della digitalizzazione tendono a concentrarsi sugli standard che è opportuno rispettare a monte e a valle dell'OCR vero e proprio, cioè la scansione/riproduzione fotografica e il salvataggio dei file ASCII che ne derivano, insieme ai metadati che li accompagnano. Si vedano, ad esempio, i siti web della Library of Congress[7], dell'IFLA[8], dell'ICCU[9] e di MINERVA[10], o l'*Handbook for Digital Projects on line*[11] che riporta osservazioni molto generali a cura di Eileen Gifford Fenton (Jstor – University of Michigan), la quale rinvia alle esperienze dell'Università del Nevada di cui parleremo più avanti. Anche un manuale assai valido perché molto preciso e documentato come quello di Jacquesson e Rivier[12] non entra in dettagli: definisce i risultati di

5 Ad esempio, lo storico progetto Gutenberg tedesco (<http://gutenberg.spiegel.de/>), che ad oggi ha messo in linea più di 4.000 volumi, utilizza appositamente una versione del software *Finereader* speciale per i caratteri gotici.

6 Michael Lesk, *Image formats for preservation and access. A report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access* (July 1990): «...Unfortunately, despite many advertisements of OCR (optical character recognition) programs, it is still rather difficult to go from image to character representation. The programs now on the market are adequately fast (10-50 characters per second) for a job that is relatively easy to read (e.g., clear, uniform text), but they are not accurate or versatile enough to handle non-standard type and faded images that are characteristic of old books», <http://www.ifla.org/documents/libraries/net/lesk.txt>.

7 <http://www.loc.gov/library/libarch-digital.html>.

8 <http://www.ifla.org/II/index.htm#2>.

9 Studio di fattibilità per la Biblioteca digitale italiana commissionato dalla Direzione generale per i beni librari e gli istituti culturali alle società Unysis ed Intersistemi di Roma alla fine del 1999 e consegnato al Ministero a dicembre 2000, <http://www.iccu.sbn.it/upload/documenti/aggSDF_pt-7.pdf>. Descrive i formati di output dell'OCR.

10 <http://www.minervaeurope.org/guidelines.htm>: le buone pratiche di alcune iniziative.

11 *Handbook for digital projects*, Sitts editor, Andover (Massachusetts) : Northeast Document Conservation Center, 2000, http://www.nedcc.org/oldnedccsite/digital/dman.pdf.

12 Alain Jacquesson – Alexis Rivier, *Bibliothèques et documents numeriques : concepts, composantes, techniques et enjeux*, Paris: Editions du Cercle de la Librairie, 1999, p. 85-87.

OCR a 300 dpi eccellenti per i documenti a stampa recenti. Cita Jstor e il progetto Tulip che forniscono l'immagine insieme al testo ASCII non corretto. Jstor utilizza un OCR che garantisce (un po' ottimisticamente, a nostro avviso) una correttezza del 99,95%.

Più ricchi di informazioni sono i siti di iniziative specialistiche o di progetti di digitalizzazione cooperativa che non hanno particolari preoccupazioni formali o di principio. Ad esempio Distributed Proofreader, che ha digitalizzato in modo volontario e cooperativo circa 12.000 titoli, è un esempio di approccio pragmatico all'uso e alla valutazione dell'OCR[13]. Così è anche per il Million Books project[14].

I prodotti

Sul mercato sono oggi disponibili un certo numero di prodotti le cui caratteristiche e il cui prezzo possono variare significativamente; tuttavia la maggior parte dei pacchetti attualmente commercializzati si situa in una fascia di prezzo abbastanza ben definita[15]. Un'eccezione è PrimeOCR, che si basa su un motore multiplo e offre maggiori informazioni/statistiche riguardo al livello di confidenza con cui vota il riconoscimento dei caratteri. Inoltre esso fornisce le coordinate delle parole riconosciute consentendo di evidenziarle al termine dell'operazione di *retrieval*. Le informazioni riportate in tabella sono basate su un report di MEDLINE aggiornato da indicazioni reperibili da un sito specializzato nella vendita di hardware e software per la digitalizzazione e da una consultazione diretta presso i siti dei produttori[16].

13 <http://www.pgdp.net/c/faq/scanning.php#12>: «Do I have to use Abbyy Finereader ? No, of course not. The scanning guidelines are heavily skewed toward using that simply because there are more people who have been using that package involved in the site, so there are more people familiar with it to answer questions. Two other packages that people have been successful with are: OmniPage Pro 10 & 11 and Textbridge Millennium Pro. They both have good recognition rates and similar functionality as far as automating the scanning process. The details differ but reading the help files should get you on the right track. OEM software that comes free with scanners CAN be used... just be aware that accuracy is typically much worse, AND be prepared to do a lot more saving and formatting manually».

14 <http://www.archive.org/about/about.php>: nel sito si può, per esempio, leggere una valutazione di Finereader 8.0 pubblicata il 7 dicembre 2007 a cura di Mark Bromley.

15 L'alternativa (futura) più nota e promettente ai prodotti a pagamento è il progetto Tesseract. Originato da un motore di riconoscimento progettato negli anni Ottanta dalla HP, è attualmente in fase di sviluppo e ospitato presso il sito di Google. Per chi volesse cimentarsi con il suo utilizzo: <http://code.google.com/p/tesseract-ocr/>. Ad oggi è possibile utilizzarne una versione gratuita di semplice installazione (FreeOCR.net ver. 2.4 aprile 2008) dal sito: <http://softi.co.uk/freeocr.htm>. È possibile scaricare il motore, una maschera grafica e aggiungervi il dizionario italiano. La qualità del motore appare più che buona. C'è da sperare che il lavoro di sviluppo continui, consentendone l'utilizzo all'interno di un pacchetto dotato delle funzionalità base: il riconoscimento automatico delle aree e la gestione multipagina.

16<http://archive.nlm.nih.gov/pubs/thoma/mars2001_4.php>. Il Progetto Gutenberg, nelle FAQ di Distributed Proofreader (un progetto di correzione di bozze: <http://dp.rastko.net/faq/scan/scanfaq.php#7>), indicava anche alcuni prodotti gratuiti. Tuttavia nel 2008 solo un prodotto tra quelli elencati (SimpleOCR) era disponibile a titolo gratuito presso il sito <http://www.scanstore.com/> con l'avvertenza: «SimpleOCR is useful for those who just need to convert a few pages to text or MS Word to avoid retyping. SimpleOCR is a lifesaver for anyone faced with retyping a file from a hardcopy. However, if your documents have multi-column layouts, pictures or other formatting you wish to preserve, you should try a commercial product».

Prodotto e produttore	Prezzo per una licenza	Indirizzo produttore
PrimeOCR (Prime Recognition)	1.500,00 - 4.500,00 $	http://www.primerecognition.com
OmniPage (Nuance)	110,00 €	http://italy.nuance.com
Cuneiform (Cognitive Enterprise)	70,00 $ - 120,00 $	http://www.ocr.com/
PdfCompressor (CVista)	500,00 $	http://cvisiontech.com/
Finereader (Abbyy)	120,00 €	http://www.abbyy.com/
Readiris (Iris)	130,00 €	http://www.irislink.com/
TypeReader Desktop (Expervision)	395,00	http://www.expervision.com/

Come valutare gli OCR

La stima dell'accuratezza dell'OCR non è un esercizio semplice né standardizzabile. Gli obiettivi possono essere diversi: controllo del rispetto della qualità da parte di un fornitore esterno; definizione delle caratteristiche ottimali dell'immagine (definizione, profondità ecc.); misura del successo del recupero dei descrittori; verifica delle prestazioni di diversi software OCR rispetto agli originali da riprodurre ecc. C'è però un motivo determinante per cui tale verifica è necessaria. Gli OCR sono intrinsecamente dei programmi imprevedibili: diversamente da un foglio elettronico o da un *wordprocessor*, il loro output non è definibile a priori e deve essere misurato empiricamente[17]. Gli esempi che seguono possono offrire un'esemplificazione riguardo ai possibili metodi utilizzabili.

Nel 1998 la School of Information dell'Università del Michigan ha svolto un test per misurare l'accuratezza del progetto Making of America, che all'epoca aveva digitalizzato 630.000 immagini TIFF[18]. I testi ottenuti con il programma PrimeOCR sono stati campionati con apposita procedura. PrimeOCR fornisce automaticamente una stima della correttezza del riconoscimento misurata su una scala di confidenza (Prime Score) che va da 100 (totalmente inaccettabile) a 900 (privo di errori). Tale punteggio è stato correlato con il risultato di una verifica manuale per verificare la ragionevole accuratezza dell'OCR. Il risultato dell'esperimento ha indicato che sopra al punteggio 880 di confidenza (77,6% del campione di pagine) non era necessario alcun intervento correttivo manuale, in quanto l'accuratezza media percentuale dei caratteri era 99,86%, corrispondente a 99,45% per le parole. In conclusione, il test ha mostrato l'affidabilità della rilevazione automatica della qualità del programma e la percentuale assai ridotta delle pagine ove poteva essere necessario procedere ad un controllo manuale.

Un diverso approccio è stato seguito dal Library Digital Initiative Team della biblioteca dell'Università di Harvard[19], che è partita dalla considerazione che l'obiettivo dell'OCR non è la fedeltà all'originale di per sé, ma la possibilità di usare efficacemente il file testuale per effettuare ricerche di descrittori. Ha messo a confronto un indicatore del tasso di recupero di informazione misurato empiricamente con la presunta qualità dell'OCR, indicata dal punteggio di confidenza del testo (Prime Score). Allo scopo sono state selezionate circa

17 «Unlike a word processor or spreadsheet program, the behavior of an OCR system is complex and unpredictable. Like other pattern recognition systems, an OCR system is " trained" using a set of data. Its performance when processing other data is not known a priori, and must be measured empirically». Stephen V. Rice, *Measuring the Accuracy of Page-Reading Systems*, Las Vegas: University of Nevada. Department of Computer Science, 1996.

18 Measuring the accuracy of the OCR in the making of America : A report prepared by Douglas A. Bicknese, <http://quod.lib.umich.edu/m/moagrp/moaocr.html>.

19 Measuring search retrieval accuracy of uncorrected OCR: findings from the Harvard-Radcliffe Online Historical Reference Shelf Digitization Project. Harvard University Library LDI Project Team. Redazione: Stephen Chapman. 1. August, 2001, <http://preserve.harvard.edu/pubs/ocr_report.pdf>.

40.000 pagine per svolgere circa 2000 ricerche (4,7% totale della popolazione), ottenendo un successo del 96,6%. Scopo del test era verificare l'eventuale correlazione positiva tra i due indicatori. Il risultato non ha confermato tale presunta correlazione, in quanto il tasso di successo non variava significativamente in relazione al punteggio di confidenza.

Al di là della conclusione per cui l'indicatore di confidenza non potrebbe essere utilizzato per stabilire a priori la qualità dell'OCR, l'esperimento può indicare la strada verso lo sviluppo di una metodologia economica di misura della qualità.

Un progetto rilevante che ha condotto un test è quello di MEDLINE[20]. In questo caso la verifica è stata condotta confrontando sei prodotti commerciali[21] con un campione di 20.000 caratteri contenuti in 15 pagine di articoli biomedici scansionate a 300 dpi. I giornali sono stati prescelti in quanto presentavano specifici problemi di conversione (testo molto compatto e caratteri di corpo molto ridotto). L'output è stato misurato in base a tre indicatori: caratteri o parole corretti, ma segnalati come dubbi dall'OCR (falsi allarmi); caratteri o parole errati e segnalati come dubbi e infine caratteri o parole errati e non segnalati. Tabulando questi tre indicatori il programma che complessivamente emerge come più vicino al testo corretto è Prime OCR che, non a caso, è un prodotto *multi engine,* cioè con più di un motore[22]. Infine il prodotto prescelto presentava funzionalità considerate di particolare utilità: coordinate dei caratteri (utili nella presentazione dei risultati di ricerche per stringhe), livello di confidenza, dimensioni e attributi dei font ecc.

Più empirico/descrittivo è un altro esempio di valutazione offerto da un sito commerciale che fornisce software per dislessici[23]. Sebbene l'utenza di riferimento siano i soggetti dislessici, il suo pregio è tentare un semplice confronto diretto tra quattro diversi prodotti. Il test ne ha misurato le prestazioni rispetto a quattro diverse tipologie di testo a stampa in termini di: caratteri/parole errate; riconoscimento delle aree di stampa (testo, tabelle e grafici) e ricostruzione del layout originale. Un lavoro simile è stato svolto da un sito commerciale di servizi per soggetti con disabilità[24].

Il toolkit dell'Università del Nevada

In questo quadro si inserisce il lavoro svolto dall'Information Science Research Institute (Isri). L'Isri è stato istituito nel 1990 presso l'Università del Nevada, a Las Vegas, allo scopo di sviluppare tecnologie per la lettura automatica di documenti. Nel corso della prima metà degli anni Novanta l'Istituto ha condotto un'attività sistematica di valutazione dei principali OCR (ISRI Annual Test of Page-Reading Systems) e a tal fine ha sviluppato una serie di programmi automatici per misurarne la qualità dell'output. Il pacchetto comprende diciassette programmi divisi in tre gruppi per misurare: a) il riconoscimento dei caratteri; b) il riconoscimento delle parole; c) la precisione nella definizione delle aree (testo, tabelle, imma-

20 Automating the production of bibliographic records for MEDLINE. George Thoma. National Library of Medicine, Bethesda, MD USA. An R&D report of the Communications Engineering Branch. Lister Hill National Center for Biomedical Communications National Library of Medicine. September 2001. <http://archive.nlm.nih.gov/pubs/thoma/mars2001.php>.

21 Prime OCR, Wordscan, TextBridge, Omnipage, Cuneiform, Maxsoft-Ocron.

22 Altri fattori presi in considerazione: la presenza di un *proofreading* assistito da immagini *bitmap* (una *utility* ormai standard), la presenza di un dizionario medico e l'accessibilità tramite altro software gestionale.

23 How successful is Optical Character Recognition software? A cura di Ian Litterick. Pubblicato il 4 ottobre 2005. <http://www.dyslexic.com/articlecontent.asp?CAT=Reviews&slug=85>.

24 <http://www.microlinkpc.co.uk/news.php?ID=7&CurrentPage=1>. L'articolo dedicato a scanner e OCR è datato 19 maggio 2006.

gini). Per quanto concerne i caratteri, i programmi sono in grado di misurare non solo il numero degli errori, ma anche di riportare, presentandoli uno accanto agli altri, i caratteri corretti in corrispondenza di quelli erroneamente riconosciuti o ignorati. È possibile effettuare la verifica a livello di singola pagina e successivamente di tabulare il totale sommando i risultati delle singole pagine. Un altro programma consente di collocare gli errori nel contesto della pagina per una verifica più puntuale. Altri programmi forniscono alcuni indicatori statistici. In modo analogo operano i programmi che misurano la correttezza delle parole. In più, questi ultimi consentono di utilizzare una lista di *stopword* in modo da circoscrivere la verifica delle parole corrette a quelle significative. Infine la sezione relativa al riconoscimento aree identifica tre tipi di errori che richiedono una correzione. Se un programma di riconoscimento non rileva un'area di testo, è necessario inserire il testo mancante; se identifica un'area grafica come testo, bisognerà eliminare il testo inserito erroneamente; infine, se determina erroneamente l'ordine di lettura di un testo, sarà necessario spostare tale blocco di testo. Il programma misura il numero di correzioni necessarie. Il software per effettuare le verifiche è di utilizzo elementare, ma richiede la disponibilità di un *ground truth*, cioè del testo corretto con cui confrontare l'output da misurare. Nel 2007 l'Isri ha pubblicato il pacchetto dei programmi (OCR Performance Toolkit)[25] sotto licenza Apache, Ver.2.0. In tal modo oggi è possibile per chiunque valutare la qualità dell'output di un OCR rispetto al set di pagine già predisposto dall'Isri o a un campione scelto a piacere.

La verifica empirica

Nel corso del 2008 si è misurata l'accuratezza di tre programmi commerciali OCR largamente diffusi in Italia e all'estero i cui motori sono frequentemente inclusi nel software in dotazione a diversi modelli di scanner: Abbyy Finereader (ver. 9.0), Scansoft Omnipage (ver. Pro.14)[26] e Iris Readiris[27]. A tal fine il *toolkit* dell'Isri è stato applicato ad un campione di pagine estratte da un database creato nel corso del 2006/2008 a Torino[28]. Il campione è stato creato partendo dalla disponibilità di circa 15.000 pagine, risultato del progetto di scansione appena concluso. I volumi oggetto della verifica sono stati tutti pubblicati tra il 1950 e il 2000 in Italia e presentano caratteristiche standard di impaginazione (caratteri di corpo medio, salvo per alcune note a piè di pagina, distribuzione del testo a una o due colonne, limitata presenza di tabelle, grafici e immagini, lingua italiana con occasionali nomi in inglese, tedesco e francese)[29].

25 <http://www.isri.unlv.edu/ISRI/OCRtk#Analytic_Tools>. Dal sito si può scaricare una "cassetta degli attrezzi" che, oltre all'Analytic Tool (Gzipped Tar file di programmi, immagini delle pagine e *ground truth*, cioè testo ASCII verificato), contiene un manuale di istruzioni e la tesi di Ph.D. di Stephen V.Rice già citata, che illustra in dettaglio la logica dei programmi.

26 Precedenti versioni sono già state oggetto di valutazioni, non solo dall'Università del Nevada:«A review in 2002 by ZDnet India (Pardawala and Kantawalla, 2002 [*non più raggiungibile n.a.*]) rated Omnipage at 99.29% and ABBYY Finereader at 99.05%». M. Lesk, cit. p. 56. Si fa notare che al momento dello svolgimento del test era già sul mercato la versione 16 di Omnipage.

27 Non sono riportati i risultati di quest'ultimo prodotto, in quanto sono apparsi da subito di qualità drasticamente inferiore agli altri.

28 Il database è costituito da 37 volumi a stampa pubblicati tra il 1950 e il 2000 e digitalizzati cooperativamente dal gruppo di biblioteche BESS grazie alla consulenza tecnica dell'Istituto Superiore Mario Boella di Torino e al finanziamento della Compagnia di San Paolo di Torino.

29 Si segnala che le pagine contenenti tabelle sono in misura limitata: l'obiettivo dell'esercizio non era infatti la valutazione del loro riconoscimento, che continua ad essere un'area critica per ogni OCR, nonostante i continui miglioramenti di tali programmi.

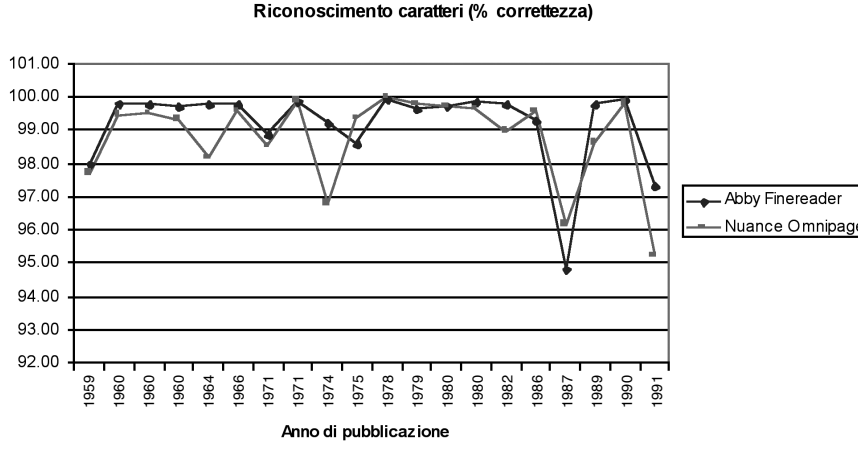
Le immagini sono state riprese tramite una fotocamera Canon EOS 5D con sensore CMOS da 12,8 megapixel, spazio colore RGB e obiettivo 50mm f/2.5 compact-macro. È stata utilizzata un'esposizione automatica non corretta. Nessuna elaborazione è stata applicata in post produzione per correggere il contrasto, ripulire eventuali *bleeding through* o aumentare la *sharpness* dei caratteri. Il colore non è stato eliminato e le uniche modifiche apportate sono state lo scontornamento e il raddrizzamento della pagina.

Per selezionare il campione si è utilizzato un sito[30] allo scopo di scegliere casualmente 20 libri sul totale della banca dati considerati più che rappresentativi in termini di carattere tipografico, impaginazione, contenuti e vocabolario. Per ogni libro sono state scelte casualmente con lo stesso metodo tre pagine e scartate solo quelle prive di testo o costituite esclusivamente da tabelle. Le uniche correzioni apportate hanno riguardato spazi bianchi, a capo, alcuni simboli non alfanumerici ed eventuali errori nel riconoscimento delle aree. Queste ultime infatti non erano tra gli obiettivi della valutazione. Il campione così costruito è risultato costituito da 159.420 caratteri e 22.758 parole. I risultati complessivi sono riportati in tabella.

Precisione del riconoscimento dei caratteri[31] (lettere, numeri, segni di punteggiatura e spaziatura)

	Totale errori generati	**Accuratezza** (% sul totale di 159.420 caratteri)
Abbyy Finereader 9	1.235	99,23
Nuance Omnipage 14	1.894	98,81

Esaminando in dettaglio la concentrazione degli errori, emerge una correlazione evidente tra i due prodotti: entrambi mostrano maggiori difficoltà per le stesse pagine, sebbene, data la ridotta dimensione del campione, si evidenzino scarti apprezzabili. A titolo di esempio si riporta un grafico che confronta le pagine campione riconosciute da Finereader (insieme alle corrispondenti riconosciute da Omnipage) rispetto alla percentuale di accuratezza dei caratteri:



30 <www.random.org>.

31 Generare la lettera "c" dove il carattere corretto è una "e" viene definito dal programma una confusione e determina un errore, poiché si tratta di una sostituzione. Se viene generato "rn" al posto della lettera "m" vengono computati due errori, poiché questa "confusione" richiede una sostituzione e un cancellazione.

Dove e perché si concentrano le confusioni/errori? Il campione utilizzato non è stato sottoposto a particolari operazioni di pulizia, né si è provveduto ad aumentare la definizione ove il corpo dei caratteri era particolarmente ridotto. Ne consegue che, come intuibile, le cause principali degli errori risiedono nella qualità degli originali. I risultati avrebbero potuto migliorare significativamente se gli originali fossero stati sottoposti ad operazioni di pulizia.

È utile segnalare che nel caso di Omnipage le cause principali di confusione risiedono nel non corretto rilevamento dell'a capo (107 istanze) che impedisce la ricostruzione corretta delle parole, la quale, in ultima istanza, è lo scopo dell'operazione di riconoscimento dei caratteri. Le due tabelle riportano gli errori più frequenti rilevati nel campione di pagine selezionato per l'esperimento:

OMNIPAGE (errori più frequenti)	
	{Corretto} - {Generato}
107	{}-{-}
38	{,}-{.}
38	{}-{-a capo}
21	{e}-{c}
16	{}-{a capo}
12	{d}-{(l}
11	{o}-{o}
10	{l}-{l}
8	{d}-{el}
7	{r}-{n}
6	{ch}-{ell}
5	{i}-{l}
5	{s}-{e}
5	{y}-{v}
5	{ì}-{i}
5	{ú}-{ù}
4	{1}-{l}
4	{l}-{1}
4	{d}-{cl}
4	{l}-{i}

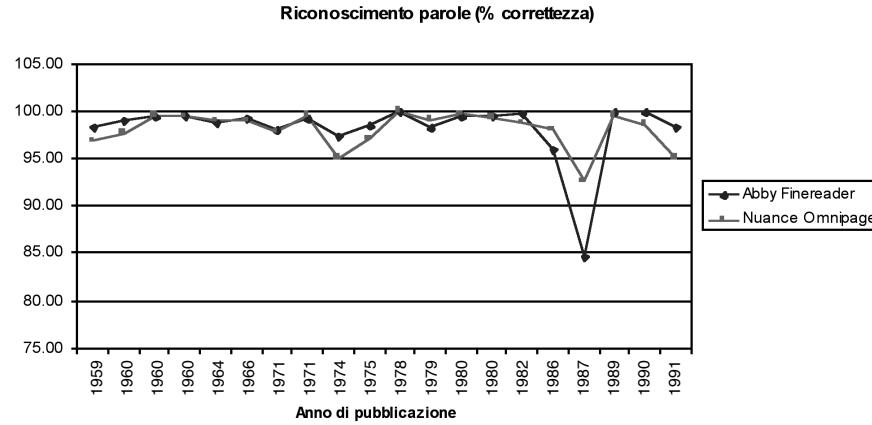
FINEREADER (errori più frequenti)	
	{Corretto}-{Generato}
71	{}-{ }
55	{1}-{7}
41	{ú}-{ù}
35	{a capo}-{ }
24	{1}-{l}
20	{,}-{.}
19	{ì}-{i}
14	{í}-{i}
13	{e}-{é}
11	{l}-{i}
9	{i}-{l}
9	{e}-{c}
8	{n}-{a}
7	{c}-{e}
7	{9}-{)}
7	{e}-{a}
6	{ll}-{U}
6	{m}-{a}
5	{o}-{ò}
4	{b}-{o}

Precisione nel riconoscimento delle parole

	Non riconosciute	**Accuratezza** (% sul totale di 22.758 parole)
Abbyy Finereader 9	348	98,47
Nuance Omnipage 14	409	98,20

Il grafico relativo al riconoscimento delle parole mostra la correlazione nell'efficacia dei due programmi, analogamente a quanto osservabile riguardo ai caratteri. Anche in questo caso i campioni delle pagine sono stati ordinati in base alla precisione. Come si può arguire dal grafico, l'80% degli errori si concentra nel 20% del nostro campione costituito da originali di bassa qualità tipografica, da corpo tipografico di dimensioni particolarmente ridotte (con conseguente bassa risoluzione)

e da nomi propri stranieri. È opportuno aggiungere che in questo caso risulta ininfluente il peso delle tabelle, che pongono, come si sa, problemi di non facile soluzione per tutti gli OCR.



Osservazioni conclusive

Se la precisione nel riconoscimento dei caratteri e, soprattutto delle parole, è la chiave principale di valutazione degli OCR, è opportuno non dimenticare che altri aspetti non secondari vanno presi in esame quando si deve selezionare il programma più opportuno. Si sono riassunte in ordine sparso in tabella alcune considerazioni essenziali:

Lavoro	Note
Acquisizione	
Tempi	In generale sono una funzione dell'accuratezza del riconoscimento. Possono costituire un problema quando per l'eccessivo numero di pagine o la difficoltà nel riconoscimento si rischia di bloccare il programma e perdere il lavoro già fatto. Spesso è consigliabile dividere il lavoro a blocchi non superiori alle 100 pagine o anche meno in caso di presenza di tabelle numeriche o di originali di bassa qualità.
Tolleranza	Al di sotto o al di sopra della definizione ottimale di 300 dpi (nel caso di corpo tipografico non troppo piccolo) si evidenzia l'insorgere di gravi difficoltà per alcuni prodotti. Almeno in alcune istanze, sotto i 72 dpi si è verificato che Omnipage, ad esempio, non accetta il lavoro.
Riconoscimento	
Riconoscimento aree	È sicuramente un aspetto critico, in particolare per quanto riguarda impaginati complessi come giornali o opuscoli o tabelle. Le tabelle, specialmente quelle con strutture complesse con colonne o righe annidate, costituiscono una sfida. Sia Finereader che Omnipage offrono funzionalità per la correzione manuale, ma richiedono un notevole dispendio di tempo e il risultato non sempre è all'altezza.
Confidenza/Correttezza	Il pacchetto dell'Università del Nevada presentato in questo contributo fornisce un approccio tra i più completi per misurare questa funzionalità. Alcuni software forniscono però anche un'indicazione del livello di confidenza del proprio lavoro che aiuta a valutare immediatamente il successo nel riconoscimento.

Informazioni/Statistiche	La presenza delle coordinate delle parole (Prime OCR) consente una più rapida individuazione delle stringhe identificate con le operazioni di ricerca.
Controllo	Molto spesso l'output dell'OCR viene utilizzato solo per la ricerca di stringhe in *fulltext*. Per questo la precisione normalmente offerta dai programmi è sufficiente. Se invece l'output deve essere controllato, è importantissimo verificare le funzionalità delle opzioni di controllo e la comodità dell'interfaccia. Gli autori di questo contributo considerano il "desktop" di Finereader 9 superiore a Omnipage 14.
Salvataggio	
Formati supportati	Sia Finereader che Omnipage supportano una più che sufficiente gamma di formati di immagine (BMP, GIF, JPEG, PDF, TIFF, ecc.) che di salvataggio (DOC, RTF, XML, PDF, HTM, PPT, CSV, TXT; XLS, DBF, ecc.). Non è scontato che invece altri prodotti possano essere più carenti sotto questo profilo

In conclusione ricordiamo che gli aspetti sopra richiamati, la questione dell'affidabilità e i possibili approcci per la sua verifica empirica, sono solo alcuni degli aspetti da valutare. La scelta dell'OCR dipende anche da altre importanti e più generali considerazioni. Alcune di queste sono state elencate da Eileen Gifford Fenton e da Hoyt N. Duggan del progetto JSTOR[32]. Sulla base dell'esperienza diretta di JSTOR, essi sottolineano che la scelta dell'OCR non può essere fatta indipendentemente dalle caratteristiche del progetto di digitalizzazione. Almeno sei fattori devono essere presi in considerazione[33]:

1. Il testo elettronico deve essere realizzato avendo presenti caratteristiche ed esigenze del pubblico destinatario del progetto.
2. Le caratteristiche del materiale a stampa sono determinanti nel definire la qualità del risultato dell'OCR.
3. Le misure di controllo della qualità dovrebbero utilizzare un adeguato supporto statistico.
4. Il software deve adeguarsi alla dimensione presunta del progetto.
5. Ponderare la localizzazione del lavoro. La scelta di un fornitore esterno non può essere fatta solo per ragioni economiche: è importante mantenere il controllo sulle competenze.
6. Programmare con attenzione i tempi e i costi.

Non è difficile immaginare che in futuro gli OCR punteranno progressivamente a diversificarsi e specializzarsi. Per i grandi progetti di digitalizzazione sembra prevalere la tendenza a renderli parte integrante del software più complesso che governa il flusso di lavoro a partire dalla scelta dell'originale fino alla sua pubblicazione online e che di norma è parte dell'hardware dedicato[34]. Per le iniziative di dimensione più limitata, i prodotti qui considerati presentano oggi un valido rapporto qualità/prezzo e non sembrano, quantomeno a breve termine, insidiati da alternative *freeware* come il progetto *Tesseract*.

32 Roger C. Schonfeld, *JSTOR : a History,* Princeton: Princeton University Press, 2003.

33 Le osservazioni della Gifford Fenton sono raggiungibili nel sito della Text Encoding Initiative: <http://www.tei-c.org/About/Archive_new/ETE/Preview/duggan.xml>.

34 Olivesoftware (http://www.olivesoftware.com/), ad esempio, offre OCR integrati per la gestione di originali con layout di stampa particolarmente complessi come i quotidiani o i periodici.

OCR: digitisation projects and optical character recognition testing

by Markus Brantl and Tommaso Garosci

Today programmes that perform the translation of a bitmapped file (typically a .jpg or .tiff) into a machine-readable text (OCR, Optical Character Resolution) are deemed to offer a more than satisfactory degree of confidence. Hence they are considered a mature technology and the assessment of their performance levels does not represent a priority in the broader research field of information retrieval and text manipulation. Still, particularly for librarians, their performance assessment is not a useless or easy exercise. The purpose of such test depends on the scope and scale of relative digitization projects. There is also a more compelling reason why such evaluation is important. OCR's are unpredictable. Their output is not a-priori foreseeable and has to be empirically measured. Accurate empirical evaluations have been carried out by the University of Michigan, Harvard and by Medline on English texts.

For a few years now a set of programmes to test "page readers" (OCR) has been made available online free of charge by the Information Science Research Institute (Isri) of the University of Nevada at Las Vegas. The so-called "frontiers toolkit" (ftk) contains both the control programmes and the bitmapped image files and ground truth control texts necessary to carry out the job.

The article, after reviewing the available literature on the subject, reports on the results of an assessment performed with the help of the University of Nevada toolkit. To this end it was created a sample of 300 dpi bitmapped images of pages randomly selected from a dataset of about 15.000 pages in Italian published between 1950 and 2000 and digitized in Turin in 2006/2008.

With the purpose of carrying out the test a ground truth text version of the sampled pages was manually construed. The tested programmes: Abby Finereader (ver. 9.0), Scansoft Omnipage (ver. Pro.14: ver.16 is the current product) and Iris Readiris, were chosen because they are the default choice in Italy and in fact were used by the digitization project carried out in Turin. Only the character and word accuracy of the programmes were evalued. Results show how Finereader and Omnipage basically have the same performance footprint, whereas Readiris is well below an acceptable threshold (this is why this latter is not included in the report).

The text is complemented with some general remarks about possible future developments as regards free OCR's and a checklist of considerations to bear in mind when choosing an OCR.

MARKUS BRANTL, Münchener Digitalisierungszentrum, Bayerische Staatsbibliothek, Monaco di Baviera, e-mail markus.brantl@bsb-muenchen.de.

TOMMASO GAROSCI, BESS and Ires, Istituto Ricerche Economico Sociali del Piemonte, Via Nizza 18, 10124 Torino, e-mail garosci@ires.piemonte.it

Bollettino **AIB**, ISSN 1121-1490, vol. 48 n. 4 (December 2008), p. 331-341.