British
Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Optical Character Recognition (OCR) Testing

NBIC Project

Internal Report IR/06/066

BRITISH GEOLOGICAL SURVEY

NBIC PROJECT

INTERNAL REPORT IR/06/066

Optical Character Recognition (OCR) Testing

Eleanore C J Corry and Emily A Swain

Keywords

Optical Character Recognition, OCR, Software Testing, Text Capture.

Bibliographical reference

CORRY, ELEANORE C J, AND SWAIN, E A, .2006. A geological background and planning for any area. *British Geological Survey Internal Report*,IR/06/66. 11pp.

Keyworth, Nottingham British Geological Survey 2006

BRITISH GEOLOGICAL SURVEY

The full range of Survey publications is available from the BGS Sales Desks at Nottingham, Edinburgh and London; see contact details below or shop online at www.geologyshop.com

The London Information Office also maintains a reference collection of BGS publications including maps for consultation.

The Survey publishes an annual catalogue of its maps and other publications; this catalogue is available from any of the BGS Sales Desks.

The British Geological Survey carries out the geological survey of Great Britain and Northern Ireland (the latter as an agency service for the government of Northern Ireland), and of the surrounding continental shelf, as well as its basic research projects. It also undertakes programmes of British technical aid in geology in developing countries as arranged by the Department for International Development and other agencies.

The British Geological Survey is a component body of the Natural Environment Research Council.

British Geological Survey offices

Keyworth, Nottingham NG12 5GG
☎ 0115-936 3241 Fax 0115-936 3488
e-mail: sales@bgs.ac.uk
www.bgs.ac.uk
Shop online at: www.geologyshop.com

Murchison House, West Mains Road, Edinburgh EH9 3LA
☎ 0131-667 1000 Fax 0131-668 2683
e-mail: scotsales@bgs.ac.uk

London Information Office at the Natural History Museum (Earth Galleries), Exhibition Road, South Kensington, London SW7 2DE
☎ 020-7589 4090 Fax 020-7584 8270
☎ 020-7942 5344/45 email: bgslondon@bgs.ac.uk

Forde House, Park Five Business Centre, Harrier Way, Sowton, Exeter, Devon EX2 7HU
☎ 01392-445271 Fax 01392-445371

Geological Survey of Northern Ireland, Colby House, Stranmillis Court, Belfast BT9 5BF
☎ 028-9038 8462 Fax 028-9038 8461

Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire OX10 8BB
☎ 01491-838800 Fax 01491-692345

Columbus House, Greenmeadow Springs, Tongwynlais, Cardiff, CF15 7NE
☎ 029–2052 1962 Fax 029–2052 1963

Parent Body

Natural Environment Research Council, Polaris House, North Star Avenue, Swindon, Wiltshire SN2 1EU
☎ 01793-411500 Fax 01793-411501
www.nerc.ac.uk

Foreword

This report is the published product of a study by the British Geological Survey (BGS) for the assessment of mainstream Optical Character Recognition (OCR) software packages for their suitability to increase the speed of borehole data capture.

Acknowledgements

We would like to thank Kenneth I. G. Lawrie and William Lyall who provided assistance.

Contents

TABLES

Summary

The tests identified Scansoft OmniPage 15 as the package most suited to the project requirement. This report describes how we tested different Optical Character Recognition (OCR) packages for use on borehole scans. **ScanSoft OmniPage 15**, **ABBYY FineReader 8**, **Readiris Pro 10** and **TextBridge Pro 11**, were analysed against certain criteria to determine which one would be more beneficial by increasing the speed of borehole data entry.

1 Introduction

Digital Borehole logs are available at the BGS from a database of scanned images. These need to be converted to text format for entry into a database of borehole details and for use in 3D lithospheric models. Capturing required text from each borehole will mainly be done manually. Where borehole logs are hand written this is the only method but it maybe possible to increase the speed of capture of typed records by automating the process with OCR software.

1.1 AIMS OF THIS REPORT

The aim of this report is to identify which package is the most suited for capturing text from scanned borehole logs for the National Borehole Information Capture (NBIC) project.

The project is part of the Information Products programme for which digital data is being transformed into easily accessible and more readily understandable formats for both geologists and non-specialist users.

2 Methodology

Four packages (ScanSoft OmniPage 15, ABBYY FineReader 8, Readiris Pro 10 and TextBridge Pro 11) were tested against the following criteria:

i) Ease of use

ii) Visual appearance

iii) Speed

iv) OCR Processing

Initially the packages were tested with a familiar set of borehole logs scans that are discussed in ‘Report on BGS Downhole record types for the National Borehole Information Capture Project’ (Swain et al, 2005) to give an overall idea on ease of use, additional complications and visual appearance, speed of reading and speed of manual correction. The attributes that we found affected the effectiveness of OCR in the scanned records included:

i) Bold fonts

ii) Small fonts

iii) Underlined text

iv) Numbers

v) Strikethrough text

vi) Lined background

vii) Skewed text

viii) Speckled and damaged scans

A table of suitable criteria from which to rate the packages was built (see Table 5) and each package was rigorously assessed against it to produce an overall rating of suitability.

3 Software summaries

3.1 SCANSOFT OMNIPAGE 15

Ease of use:

- Clear with three stages, very easy to use
- Saves settings

Visually:

- 4 panels: commands and past images selection to the left, a preview of image and text editor. Basic statistics at the base of screen, such as …
- Doesn't show where on the document it is reading, but it is possible to zoom
- There is a window highlighting suspect words

Speed:

- Moderate OCR time
- Fast to correct with suggested words

OCR processing:

- Remembers words and uses this training data on next scans if you want it to, good for geological terms

Other Features:

- Save to wide selection of software
- Images can be despeckled, deskewed
- Likes to classify areas as graphics so avoids the character recognition; using the draw text zone rather than draw process zone can stop this

3.2 ABBYY FINEREADER 8.0

Ease of use:

- Very easy and simple to use, with four stages:
    - Open image
    - Read
    - Spell check
    - Save

Visually:

- This programme is visually the strongest. As well as the image and text boxes, at the bottom is the appropriate zoomed in part of the image you are looking at thus making it very clear
- Easy to zoom in and out with the choice of numerous percentages zooms

Speed:

- Moderate to fast OCR
- Quick and easy to correct words as it gives word suggestions

OCR processing:

- Very good accuracy but it struggles to read joined up handwriting

Other features:

- Automatic save to wide selection of software

3.3 READIRIS PRO 10

Ease of use:

- Not clear how the process works and couldn't work out how to make it read only in one text area.
- All setting are selected each time you start up
- Two stage process

Visually:

- Not cluttered: 2 columns to left, commands and past images selection. Basic statistics at the base of screen, e.g
- One document preview but no facility to zoom. Can't see where it is reading from on the preview
- Doesn't clearly show a zoom of the suspect words

Speed:

- Fast OCR time
- Slow to correct as it wants each character corrected
- No suggested words

OCR processing:

- Couldn't divide lines of handwritten text, suggested a single B for three lines of text
- Can learn by entering into new dictionary

Other Features:

- Automatic save to wide selection of software
- Can detect and correct orientation of badly scanned images
- Can select and sort text box windows
- Fine adjustments can be made when rezoning documents

3.4 TEXTBRIDGE PRO 11.0

Ease of use:

- Easy to use three stage process:
  - Load file
  - Perform OCR
  - Export results
- Very simple

Visually:

- Document manager column down the left hand side, which is visually poor when looking at the details and one has to scroll along to see them all.
- The rest of the screen is made up of two boxes, the original image and text editor. You can switch between these different views or have the choice of changing their positions however overall it is visually poor.

Speed:

- Fast OCR time
- No suggested words

OCR processing:

- Couldn't read handwriting, suggests numbers and dots instead of actual words
- Doesn't remember words that have been altered

Other features:

- There is the option of a fast scan or accurate scan e.g. for borehole SE33SW1:

*Fast scan = 9.19 seconds with 81.73% accuracy

*Accurate scan = 10.51 seconds with 89.41% accuracy

4 Software analysis

The following tables represent a break down of the different software packages and their abilities. The data is taken from the data management tool within each program. The main column of interest is the accuracy column. The accuracies for the different packages have been compared along side each other thus allowing us to determine which is the most accurate program. Accuracy was measured using the following formula: (words – suspect words/words) * 100.

No table of analysis was produced for Readiris Pro 10. It was disregarded because it didn't meet the initial ease of use criteria and it was a challenge to assess the other features.

4.1 SCANSOFT OMNIPAGE 15

Bore ID	Suspect words	Characters	Words	Words/minute	Reject characters	Recognition time	Acquisition time	Accuracy %
NT27SW 94	3	693	121	2004	0	3.622	0.344	**97.52**
SE33SW 1	35	226	44	119	0	22.000	0.531	**20.45**
NY70SW 1	10	465	78	1446	0	3.236	0.328	**87.18**
SD70SE 28	77	747	139	291	0	28.000	4.625	**44.60**
SD70SE 28 (despeckled)	77	747	139	291	0	28.000	4.625	**44.06**
SE33NE 36	80	749	143	296	0	28.000	4.625	**0.00**
SE33NE 36 (only text)	4	53	4	60	0	3.968	3.482	**2.30**
NZ61NW 4	85	666	87	129	0	40.000	3.482	**57.69**
NY00NW 4 (page 1)	11	111	26	589	0	2.647	0.984	**14.29**
NY00NW 4 (page 2)	6	60	7	146	0	2.864	0.421	**3.17**
NY00NW 4 (page 3)	61	320	63	243	0	15.000	0.520	**7.69**
NC14NE 1	24	141	26	218	0	7.137	1.259	**96.26**
NS66NW 6	8	1203	214	2203	0	5.827	0.346	**11.32**
NY70SW 9	47	282	53	254	0	12.000	4.863	**16.67**
SK89SE 40	31	185	34	233	0	8.727	0.466	**8.82**
SD33NW 311	2	337	76	1265	0	3.603	3.645	**97.37**
NS66NW 119	5	29	7	184	0	2.276	3.130	**28.57**
SE12NW 568	44	233	52	451	0	6.907	0.457	**15.38**
TQ27SE 511	3	452	125	3654	0	2.052	0.513	**97.60**
NZ36NE 62	48	305	68	590	0	6.904	0.343	**29.41**
SO99SW 8 (page 1)	53	822	126	549	0	13.000	0.608	**57.94**
SO99SW 8 (page 2)	6	30	6	131	0	2.739	2.794	**0.00**

Table 1 Scansoft Omnipage 15 statistics

4.2 ABBYY FINEREADER 8.0

Borehole	Uncertain characters	Total characters	Accuracy %
NT27SW 94	0	806	**100.00**
SE33SW 1	244	393	**37.91**
NY70SW 1	94	599	**84.31**
SD70SE 28	810	1029	**21.28**
SE33NE 36	655	760	**13.82**
NZ61NW 4	2	23	**91.30**
NY00NW 4	51	61	**16.39**
NC14NE 1	55	1418	**96.12**
NS66NW 5	263	380	**30.79**
NY70SW 9	9	9	**0.00**
SK89SE 40	135	620	**78.23**
SD33NW 311	4	4	**0.00**
NZ36NE 31	100	160	**37.50**
NJ27SW 1	0	0	**0.00**
NS66NE 119	0	0	**0.00**
SE12NW 568	5	229	**97.82**
TQ27SE 511	186	266	**30.08**
NZ36NE 62	327	1025	**68.10**
S099SW 8	251	358	**29.89**

Table 2 ABBYY Finereader 8.0

4.3 TEXT BRIDGE PRO 11.0

Borehole	Suspect words	Characters	Words	Words/minute	Reject characters	Zones	Recognition time (s)	Acquisition time (s)	Accuracy %
NT27SW 94	26	695	121	4318	0	1	1.681	0.406	**78.51**
SE33SW 1	25	143	47	492	19	1	5.729	0.5	**46.81**
NY70SW 1	14	236	54	2121	1	1	1.527	0.344	**74.07**
SD70SE 28	191	550	207	725	53	1	17.123	2.5	**7.73**
SE33NE 36	12	342	73	458	13	1	9.553	1.156	**83.56**
NZ61NW 4	4	35	9	399	3	1	1.351	1.031	**55.56**
NY00NW 4	3	18	9	189	1	1	2.849	0.969	**66.67**
NC14NE 1	63	1226	216	4141	0	1	3.129	0.407	**70.83**
NS66NW 5	28	93	33	309	15	1	6.407	2.282	**15.15**
NY70SW 9	17	88	23	955	12	1	1.445	0.328	**26.09**
SK89SE 40	73	516	122	1565	9	1	4.676	0.844	**40.16**
SD33NW 311	11	18	11	589	6	2	1.12	3.031	**0.00**
NZ36NE 31	6	402	54	1578	18	1	2.052	1.219	**88.89**
NJ27SW 1	0	0	0	0	0	1	1.071	1.187	**0.00**
NS66NE 119	42	198	58	1366	38	1	2.547	0.515	**27.59**
SE12NW 568	10	523	187	7753	0	2	1.447	0.578	**94.65**
TQ27SE 511	47	232	74	1716	18	3	2.586	0.375	**36.49**
NZ36NE 62	69	564	213	1394	41	1	9.163	1.031	**67.61**
SO99SW 8	62	322	111	976	43	1	6.82	2.563	**44.14**

Table 3 Textbridge Pro 11.0 statistics

4.4 ACCURACY SUMMARY

	OmniPage 15	*TextBridge Pro 11.0*	*ABBYY FineReader 8.0*
Bore ID	**Accuracy %**	**Accuracy %**	**Accuracy %**
NT27SW 94	**97.52**	**78.51**	**100.00**
SE33SW 1	**20.45**	**46.81**	**37.91**
NY70SW 1	**87.18**	**74.07**	**84.31**
SD70SE 28	**44.60**	**7.73**	**21.28**
SE33NE 36 (only text)	**2.30**	**83.56**	**13.82**
NZ61NW 4	**57.69**	**55.56**	**91.30**
NY00NW 4	**14.29**	**66.67**	**16.39**
NC14NE 1	**96.26**	**70.83**	**96.12**
NS66NW 5	**11.32**	**15.15**	**30.79**
NY70SW 9	**16.67**	**26.09**	**0.00**
SK89SE 40	**8.82**	**40.16**	**78.23**
SD33NW 311	**97.37**	**0.00**	**0.00**
NS66NW 119	**28.57**	**88.89**	**0.00**
SE12NW 568	**15.38**	**0.00**	**97.82**
TQ27SE 511	**97.60**	**27.59**	**30.08**
NZ36NE 62	**29.41**	**94.65**	**68.10**
SO99SW 8	**57.94**	**36.49**	**29.89**

Table 4 Accuracy summary

5 OCR rating

The table below shows how we chose which package is most suited for the use of reading borehole scans. It represents what criteria we used to rate each software package.

	ScanSoft Omni Page 15	**ABBYY FineReader 8.0**	**IRIS Readiris 10**	**Text Bridge Pro 11.0**
EASE OF USE				
Overall	10	10	4	8
Complications	6	10	10	8
VISUAL				
Overall	10	10	6	6
Prove reader dialog box	10	9	2	4
SPEED				
Reading	6	7	8	8
Manual correction	10	8	2	6
OCR PROCESSING				
Bold	*1*	*-1*	*-1*	*0*
Small fonts	*0*	*0*	*-1*	*0*
Underlined	*1*	*-1*	*-1*	*-1*
Numbers	*0*	*0*	*-1*	*0*
Strikethrough	*1*	*0*	*-1*	*0*
Lined background	*1*	*1*	*-1*	*0*
Skewed	*0*	*0*	*-1*	*-1*
Speckled and damaged	*1*	*0*	*-1*	*1*
OTHER				
Text zoning selection	7	9	8	9
Reading the text box	10	10	2	10
Page orientation	10	10	10	10
Options/settings	10	5	4	9
TOTAL	94	87	48	77
RATING	**1**	**2**	**4**	**3**

Overall Scale 1-10 with 10 being the best

OCR Processing scale -1 to +1

Table 5 OCR rating table

Readiris became very frustrating and virtually impossible to use for the selected text we were using. For this reason it scored a succession of –1 due to its’ inability to work satisfactorily. No further tests were made with this application.

6 Conclusion

In conclusion we rated ScanSoft OmniPage 15 the best OCR package.