Automated simple context sheets processing

BOGDAN BOBOWSKI

Departament of Archeology, Ancient & Medieval History, Institute of History, University of Zielona Gora, Poland, bobowski@op.pl

ABSTRACT

Paper present data capture application used to extract hand-printed text from simple context sheets. When completing a form one has to enter information into blank spaces or specially designed fields that make up the structure of the form. This information must then be extracted and processed – forms digitalization. Presented OCR/ICR application Abbyy FormReader offers a user-friendly interface that minimize operational costs. Application interprets data from paper forms many times faster than any professional operator, enabling you to collect data in efficient and secure way. It is noteworthy that the entire process requires only one human operator since all of the stages, except verification, are fully automated. Therefore, automated forms processing is not only much faster than manual data entry but produces much more accurate results. The speed of scanning in this particular case is not very crucial, because packing and loading goods takes a lot more time than scanning orders.

1. INTRODUCTION

Modern computer applications allow users to eliminate the process of manually rewriting information contained in forms. The main task of the software is to collect data from scanned hand-written documents and recognise them in electronic form. Automatically read data are transformed into digital form comprehensible to the storage – analytic applications, mainly the databases. The leader of the machine-print OCR and isolated handprint ICR digital recognition market application is FormReader produced by programmers from Ukrainian Abbyy[1].

2. FIELD ARCHAEOLOGY FORMS

In 1980 The Museum of London Archaeology Service (MoLAS) published the first manual presenting some practical principles of documentation on archaeological excavations. The MoLAS recording system was based on context recordings sheets – illustrating unique activity of the human past (Archaeological Site Manual, 1994.) The context is the physical setting, location and cultural association of artefacts and features within an archaeological site. Forms with blank entry fields for answer to specific questions asked in descriptive text facilitate the objectivization of field records much better than recording them in a blank notebook. The absolute information minimum for contexts recording sheets should contain: context type, site code, context number, grid & square location of context, stratigraphic relations, description of context, sketch, references to finds, photos, drawings and plans, interpretation and opinion (Barker, P., 1994.)
Context recording sheets are a perfect source for computer analysis. The problem is data entry transfer or digitalisation of data from paper questionnaires. Mostly it is a manual process.
Solutions based on direct site data entry process (PDAs database terminal) are not so popular yet, though they are used occasionally (Ryan, N.; van Leusen, M., 2003). In most cases OCR based forms processing is considered complete when the data from all the forms have been captured, verified and saved in a database. There are only two approaches to data collection from paper forms: to involve many people in manual data keying in, or to start using an automatic form input system.

3. MACHINE-READABLE FORMS

Forms mostly consist of one card compilations of blank fields and solid descriptive texts for every field. The size of columns to be filled in is limited by the size of definite compact information. Information can be introduced in descriptive form by means of hand-written text (the written or printed character of letters) as well as in the form of marking a check box, defined closely as the favourable definite positive or negative answer. Forms are a practical solution in the case of assembling information by many respondents. All forms can be used as machine-readable forms. The process of automation of information processing from paper forms aims at acceleration and avoidance of making mistakes during digitalisation of documents. The target object for storage and analyses of data are the databases. A form can be recognised as comprehensible in terms of machine processing when it is characterised by the following features: the location of elements is identical on all copies of a definite kind of form and fields for data input have to be separated clearly from solid descriptions, lines, separators or other background. It is essential at form design step to separate data entry fields from elements which are not important for recognition, such as lines, separators, explanations, background. There are special form design applications with automatic correction of possibly recognisable errors. Reference marks are used to facilitate

1 www.abbyy.com

the matching of form images with templates and to determine the orientation of forms during automated processing. Form identifiers are used for selecting the right template if several templates have the same arrangement of reference marks. A barcode is a set of vertical black and white lines of differing widths used for information encoding.

4. DIGITAL PROCESSING OF PAPER FORMS

We tested Abbyy FormReader application. FormReader offers a user-friendly interface that minimises operational costs and increases throughput by guiding through the key steps of form processing: scanning/file import, recognition, verification of results, validation, and export of extracted data. The software offers image pre-processing, OCR (machine-print text recognition in more then 170 languages), ICR (hand-printed text recognition in more then 90 languages), OMR (recognition of check-marks of various shapes), OBR (recognition of commonly used barcodes), extracting data from semi-structured documents, powerful verification and validation modes. Specific hand-written text differs depending on national features (Wilson, D. R., 2001). The system recognises about 80% of hand-printed letters, 95% of hand-printed numbers and 100% of check marks. Effectiveness is raised by automatic verification of recognition results through built-in dictionaries.

5. OPTIONS AND POSSIBILITIES OF APPLICATION

5.1 CREATING FORM TEMPLATES

For creating machine-readable forms, application uses the integrated FormDesigner module. The information contained in textual fields will permit digitalisation of hand-written data from paper form to the database. To be able to read information on the forms, a form designing module can be used:

- determine the location of form elements – forms must correspond to the form pattern or template, i.e. the location of all form elements must be identical on all forms of the same type
- separate field contents from field borders, text marking, backgrounds, explanatory text, etc.

Design is an important stage of form creation. Scanning, recognition, verification and export depend on the "correctness" of the initial form. The module will check each created form to ensure that it can be successfully processed by the computer. Once you have created a form in Form Designer, you practically get a template that can be used for capturing data from forms. Templates describe form elements and tell data capture applications where to look for particular types of information on the form. Proposed machine-readable simple context recording sheet is designed for recording the most important information (figure 1).

5.2 SCANNING

What is needed for form scanning is the simplest scanner and minimum skill. The matter becomes somewhat problematical in the case of work on thousands of documents over a short period of time, then the best hardware solution is to use automatic feeder special scanner (ADF), which processes scanning at the speed of one hundred documents per minute. The use of feeders makes it possible to scan whole packets of documents. The feeder handles both large size documents (even size A3) and small sizes. The feeder construction prevents the seizure of documents. The possible choice of the runtime mode is on normal or fat paper (above 80g).

5.3 READING AND RECOGNITION

This is an important step of data processing. The application carries out the process of form diagnostics through adjustment of earlier defined templates to the scanned form with reference marks. The next step is recognition of the hand-written text recorded in block form localised in the defined fields.

5.4 VERIFICATION

The user can check the correctness of results manually by comparing letters negatively recognised by the application with the original letters. There is only one manual step in the recognition process. The results of manual verification are automatically recorded by the application, which extends its own base of individual features of hand-written text. The user can use rules to eliminate the verification of the fields that are definitely correct. For each field it is possible to specify how tolerant the application shall be in marking uncertainly recognised characters. FormReader is an effective self-learning system.

5.5 DATA EXPORT

Results of form recognition can be saved locally or exported. Application support exports in main formats, recognised

by databases: .xml, .dbf, .xls, .csv, odbc. It is possible to export graphic formats directly to databases in popular .tiff and .bmp extensions.

CONCLUSIONS

Manual data entry is boring and troublesome and implies many problems such as delays in data capture. Abbyy Formreader is a good solution for automatic data capture. The application interprets data from paper forms many times faster and immensely more accurately than any database operator. Finally, the extracted data look exactly like the data users receive from their current manual system, but faster, and the captured content is more accurate and reliably correct. FormReader processing of specially designed machine-readable context recording sheet (1 page) takes less than a minute, which is much faster than the speed of a professional typist.

REFERENCES

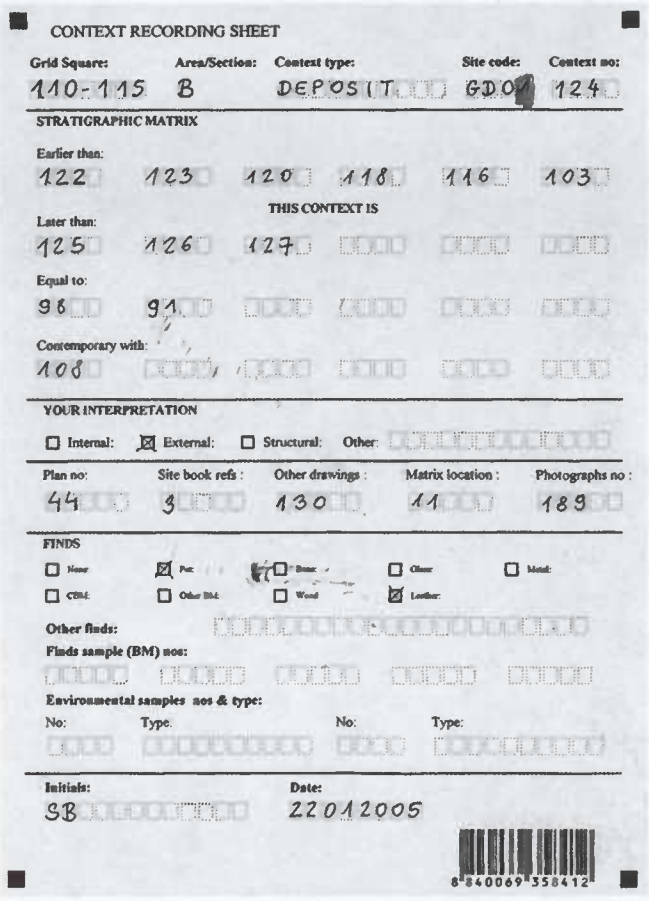
ARCHAEOLOGICAL SITE MANUAL (1994) – Museum of London Archaeology Service – *Third Edition*, London.

BARKER, P. (1994) – *Techniki wykopalisk archeologicznych.* Translated by Z. Kobyliński, Warsaw, p. 79-83.

RYAN, N.; VAN LEUSEN, M. (2003) – *Recent Experiments With the Use of Portable Recording Devices in Archaeological Field Surveys*. CAA 2003 – Computer Applications and Quantitative Methods in Archaeology Proceedings of the 31st Conference, Posters, Vienna.

WILSON, D. R. (2001) – *Applying new technologies appropriately to foreign handwriting recognition difficulties* in Atkinson. T. (ed.), Reflections on ICT, London: Centre for Information on Language Teaching and Research, London, p. 41-52. (www.tomwilson.com/david/mfl/handwriting.doc).

FIGURES


CONTEXT RECORDING SHEET

Grid Square: 110-115 Area/Section: B Context type: DEPOSIT Site code: GDO Context no: 124

STRATIGRAPHIC MATRIX

Earlier than: 122 123 120 118 116 103

THIS CONTEXT IS

Later than: 125 126 127

Equal to: 98 91

Contemporary with: 108

YOUR INTERPRETATION

☐ Internal: ☒ External: ☐ Structural: Other:

Plan no: 44 Site book refs: 3 Other drawings: 130 Matrix location: 11 Photographs no: 189

FINDS

☐ None ☒ Pot ☐ Bone ☐ Other ☐ Metal:
☐ CBM: ☐ Other Bld. ☐ Wood ☒ Leather

Other finds:

Finds sample (BM) nos:

Environmental samples nos & type:

No: Type: No: Type:

Initials: SB Date: 22012005

Fig. 1 – simple machine readeable context recording sheet, ready for use with OCR/ICR application.