An OCR Case Study

Eileen Gifford Fenton JSTOR, University of Michigan

What is OCR?

Optical character recognition, or OCR, is the process that converts the text of a printed page to a digital file. This is accomplished by using an OCR software package to process a digital image of the printed page. The software first analyzes the layout of text on the page and divides the text into *zones* that usually correspond approximately to paragraphs. Next, the order of the paragraphs is determined and then the analysis of the characters begins. Most OCR applications work by looking at character groups, *i.e.,* words, and comparing these to a dictionary included with the application. When a match is found, the software prints the appropriate word to the text file; when a match cannot be made confidently, the software makes a reasonable assumption and flags the word as a low confidence output. Where a word or character cannot be read at all, the default character for illegible text is inserted as a placeholder.

Accuracy of OCR packages varies widely.

OCR is an effective means to read modern typeface captured in high quality page images. Though OCR software has improved significantly over the last decade, OCR does not yet deal effectively with non-Arabic characters or non-modern type and frequently struggles to translate small print, certain fonts, and complex page layouts. The accuracy of OCR packages varies widely among applications and across different source materials.

JSTOR and OCR

JSTOR, an independent not-for-profit organization headquartered in New York, NY, has the large-scale undertaking to convert and maintain digital versions of the backfiles of journals and to develop access tools that allow searching of both full text and indexed components within each issue. To date, JSTOR has converted over 4 million pages from over 100 journal titles. Over 500 academic libraries in North America and abroad have signed on as institutional participants. JSTOR began digitizing journal back runs in the fall of 1994 with only minimal staff devoted to production activities. Since those early days both productivity levels and staffing have increased. Currently, JSTOR prepares approximately 200,000-250,000 new pages for digitization each month. The JSTOR production staff has grown to a group of 20 distributed between operations at the University of Michigan and Princeton University. Several other units at JSTOR including Library Relations, Publisher Relations, User Services, Technology Support and Development, and an administrative group complement the work of the production group.

Each journal page digitized by JSTOR is processed by an OCR application, and the resulting text files are used to support the full text searching offered to JSTOR users. In order to ensure that search results are as accurate as possible, each OCR text file is manually reviewed and corrected to a targeted accuracy level prior to being added to the database. Eliminating this manual review could reduce production costs. However, it has proven to be an essential

step for assuring both the overall quality of the database and the accuracy of scholars' full-text searches.

Key Points When Considering OCR

Digital projects vary widely in content, aim, and scale, and OCR may not be the right solution for all. When considering OCR, it is useful to weigh the following.

1) Select technology that will enhance your ability to meet the objectives of the project.

Manual review has proven essential.

If the project goal includes delivering converted text files to the user, you will want to think very carefully about using OCR. No OCR product is perfect. Text errors will be present in files displayed to users. As a result, you will want to thoughtfully determine the OCR accuracy level required to meet particular goals. If you are using the text files only to support searching, and they will not be displayed to the user, you may be able to tolerate lower accuracy. Decisions about accuracy should take into account the characteristics of the source material. Non-English text, mathematical or chemical symbols, and other special characters are not successfully translated by OCR applications, and their presence should be factored into your decision.

2) Scale matters -- a lot.

The appropriate approach for generating text files is affected dramatically as you move from a 20,000-page project to a 200,000-page project to a 2,000,000-page project, even if the goals of the projects are the same. Similarly, the costs generated by text file production also change dramatically with scale.

3) There is no right answer.

Solutions will be driven by the goal of the project. However, it is difficult to generalize from one project to another even when project goals may be similar. Very specific characteristics such as the nature and quality of the source materials, the available budget, and the time allotted for the project will significantly impact decisions.

4) Costs will be higher by more than you expect.

Even the most carefully planned projects including OCR will experience surprises. Initially selected software may not perform on actual data as it did on test data. You may find processing limitations in the full production phase that were hidden during the pilot phase. Expanding an application's dictionary to include specialized terms may prove to be more difficult than originally anticipated. Any number of unexpected developments may impact production timeframes and therefore budgets. It is helpful if an allowance for these unexpected developments can be built in from the beginning of the project.

5) The answer that is right for today may not be right in the future.

OCR software capabilities have developed significantly over the last five to ten years and improvements continue to be made. The dynamic nature of this technology means that projects of more than just a few months' duration may benefit by continuing to evaluate new products as they become available to determine if greater cost-benefit possibilities have developed.

Sources

Rice, Stephen V., George Nagy, and Thomas A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier.* Kluwer Academic Press: part of Kluwer International Series in Engineering and Computer Science Secs 501. 1999.

Until 1997, the Information Science Research Institute at the University of Nevada, Las Vegas, conducted an annual assessments of selected OCR products. Information on their Technology Assessment Program is available at http://www.isri.unlv.edu/info/technology/ Website: www.jstor.org Readers will find information about JSTOR's mission and history, a description of the contents of the database, and information on institutions participating in JSTOR's work. Also available is a description of our production process, technical information of general interest, and a link to a demonstration of the database.