Optical Character Recognition and Its Use in Various Forms of Information Retrieval

William J. Ray, Ph.D.*

Keywords: optical, character, OCR, conversion

INTRODUCTION

Very large amounts of genuinely useful data remain locked in analog only format. Books, magazines, newspapers and even filing cabinets contain something like 80% to 90 % of all known useful information[1] and these data are only slowly being made available in digital format. The purpose of this missive is to discuss practical analog to digital domain conversion via optical character recognition and use of such converted data in the digital domain. Note that we are concerned *only* with the manufacturing process, e.g. the conversion of the data, and not the retrieval of data *per se*, however, how we make the data available to the data store dramatically influences the options that the database operation has.

PROBLEM SETUP

In the best of all possible worlds (to borrow from Voltaire) we would fully convert and clean up all existing information. This is the optimum choice as it allows us to use full text searching and indexing (forget the problem of shear volume and expanse of the search – this is, after all, only an exercise) and, potentially, deep tagging for intent and classification. This, of course, needs essentially infinite resources and does not take into account the relative value differences of information.

Industrially it is possible to classify data that are information into three or four classes. The first of these are those data that require and deserve *complete conversion.* By complete conversion we mean not only capturing and accurately re-rendering the correct glyphic representation of the text in question (along with any pure image data) but also capturing the typesetting elements associated with the document. Two subclasses can be identified in this category, these being the *facsimile conversion* process where the document is rendered in

*Group InfoTech, Inc.

exactly the same way to the page break as the original and the *re-flow conversion* process where the typesetting elements are captured but the pagination of the document is allowed to format based upon the required output geometry.

Another general category of conversion is associated with database creation. Such conversion, known as *tagged, unformatted conversion*, generally results in an SGML tagged product that is used as a searchable database. Legal and some medical databases are built in this manner. This technique amounts to a complete conversion without capture of the typesetting elements

Incomplete conversion schemes are used for certain types of data. Such schemes generally use the image as the user deliverable while allowing the database to be searched on the converted data. This scheme uses converted, unedited data as the search space and is known as a *dirty ASCII conversion.*

Finally there is the *hybrid conversion* scheme. This is similar to the dirty ASCII approach but the resultant OCR output is parsed to remove non-word tokens and stop words. This provides a search file that avoids the potential for combinatorial explosion in large search spaces.

Non-OCR conversions, known as *facsimile reproduction*, are also common but search keys need to be entered into whatever data store is used in order to retrieve such image data.

WORKING SOLUTIONS

Three user types have been identified for *large scale* conversion. The traditional user is the legal industry due to its' reliance upon and need for massive volumes of searchable case law data. This industry is a *tagged unformatted* conversion user.

A more recent user type is the book publishing industry. These users are *complete conversion* or *facsimile reproduction* consumers. Typically, the reproduction users deal in centralized print on demand (POD) manufacturing and employ the reproduction process as a digital version of the analog "opticopy" process. The conversion user is more interested in digital delivery to the end customer via E-book, local POD or other internet based delivery schemes.

Finally the newest large scale user is the data repository operation that deals in research or library science. This is a new user category as the value of digital delivery of such data had not been very clear to this segment until the overwhelming impact of the internet was felt. Typically these operations will not

be able to afford or require *complete conversion* nor do they actually need to deliver a converted document. What is required is a very through capture of key data in order to build deep searching capability. Such users fall into the *dirty ASCII* or *hybrid conversion* schemes.

The *complete conversion* schemes are well documented elsewhere[2] and we will not deal with those approaches here. However, the *dirty ASCII* and *hybrid schemes* – which we will refer to as *incomplete conversion* -- have not been well documented and will be discussed at some length.

Incomplete conversion is not a new technique. Khevgas[3] described such an OCR engine which was coupled with an image display subsystem in the early nineties. Subsequently products such as Adobe Capture have appeared that allow the user to display a document image and search on the OCR result in the background.

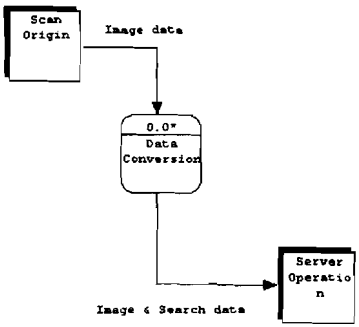
The problem associated with this approach is the proliferation of bad keys due to misspelled words. This is significant as search time for large amounts of data grow very rapidly with the volume of data (assuming that you are doing a full text search) and easily become almost useless with the large increase in the number of spurious word tokens associated with existing OCR engine technology – e.g. the search results in a combinatorial explosion. Attempts have been made to improve the quality of the OCR by using a broken token algorithm[4], however, these have not been uniformly successful.

The problem is to both increase the accuracy of the OCR process and to clean the resultant file of bad search tokens in order to provide the smallest search space possible associated with capturing all of the search tokens possible.

A new manufacturing scheme is proposed whereby the higher quality OCR conversion provided by Marvel is used and the resultant markup of bad word tokens is used in a "smart parsing" of the resultant text file to both reduce bad word tokens and to eliminate stop words. It is possible, as well, to provide certain minimal editorial checking of potential proper names in order to define a global name index local to the converted file (or, perhaps, even to provide a global name index across a serial publication that identifies commonly used names).

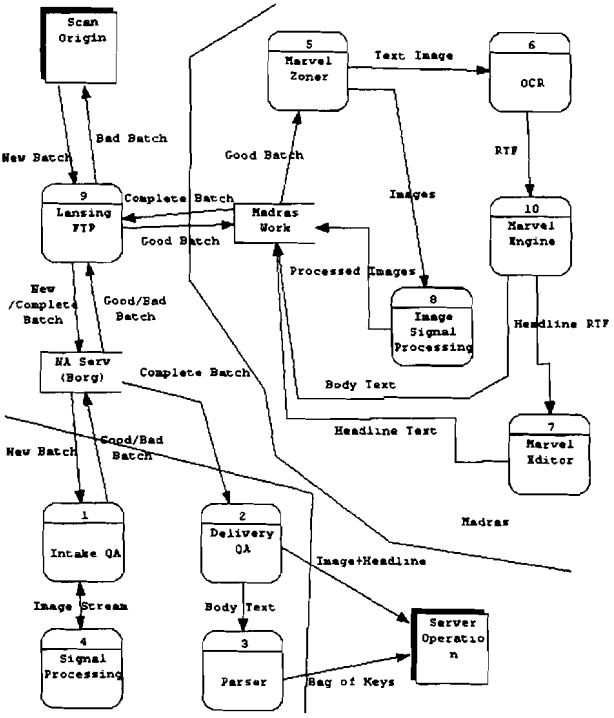
THE PROPOSED MANUFACTURING PROCESS

Figures DFD::0:Hybrid Conversion and DFD:hybrid.DFD:DataConversion are data flow diagrams illustrating the proposed manufacturing workflow. The context diagram (0) simply illustrates the gross data flow.



Hybrid Conversion Processing

The second diagram is a high level flow of the manufacturing process. Note that the process illustrates manufacturing locality (e.g. Madras, Mexico City with the US operations being unmarked).



Data are taken in at the external entity Scan Origin (this is the UMI microfilming plant) and shipped to the Lansing FTP server process and its' data store NA Serv Borg. This is a push dataflow from the customer. Data are then pulled by the Mexican operation into the Intake QA process. Here the image data are reviewed and accepted or rejected on a batch basis (one batch in this case being one issue of a newspaper).

Image data are always accepted or rejected on a batch basis for data control reasons. The microfilm scanning process is subject to a number of errors that the customer cannot afford to detect -- thus the intense QA step.

Accepted batches are sent to a Signal Processing process where the image data are de-skewed and de-speckled. The data are then classified into an initial Nartker score for image quality. Form the scoring step the data are sent to one of several image processing and enhancement algorithms and then the data are re-scored.

The image data are then pushed back to the NA SEV (Borg) store where they are available to be pulled by the Chennai/Madras operation. Note that an E-mail is also generated to production in Chennai calling notice to the availability of a batch. The user may also track flow via a web tracking tool similar to that used in the B&N production scheme.

Madras/Chennai pulls the image data to a local store (noted as Madras Work Server) where the work is dealt out as batches to Marvel Zoner operators. It is important to note that one batch must be sent to one operator – no more and no less. This is so as the zoning event is done as a threaded operation, e.g. an article can appear on one or more pages and the zone for that article needs to link those page zone elements together.

Halftone, line art and other pure image data are sent to one of several processing steps illustrated b the Image Signal Processing process. Text image data are passed through the OCR and Marvel Engine processes. Body text along with all of the text images are then complete and sent to the Madras Work Server store while headline text is sent to the Marvel Editor process for cleanup. These data are then returned to the server store upon completion and validation.

An E-mail message on batch completion is generated to Mexico City notifying that data have been pushed to the NA Serv Store. Mexico City then pulls the data for the Delivery QA process and either delivers the complete data or provides the body text to the Parser process for subsequent delivery.

CONCLUSION

The proposed manufacturing system provides to the end user a clean image – either page image or zoned element data – and an automatically processed key word set for a search engine.

This approach provides a cost effective manufacturing tool for previously financially unapproachable conversion of deeply searchable image data.

REFERENCES

[1] ISRI, 1995
[2] TAGA Proceedings papers
[3] ISRI
[4] ISRI