Large Scale Conversion of Book Backlists for Formatted, Digital Delivery: A Case Study

William J. Ray, Ph.D.*

Keywords: OCR, Workflow, XML, SGML, HTML, POD

Abstract: Large scale OCR conversion of out of print books has become both economically viable and desirable with the advent internet distribution and the general availability of both print on demand (POD) systems and the e-book.

This paper describes a new manufacturing and embedded tagging process associated with the conversion of data from physical pages to tagged electronic files. The paper specifically explores the systems workflow integration used in the manufacturing system across three countries and two continents.

INTRODUCTION

Large scale conversion of backlist or out of print books to be delivered to either print on demand (POD) facilities or for delivery of content via the internet as e-books has become a pressing book industry issue. Something on the order of 60 million English language book titles have been printed in the last 150 years. Only a tiny fraction of these titles are available in digital format.

Several public domain or industry efforts, such as Project Gutenberg, have produced on the low order of thousands of book titles in some form of digital format. Formats vary – particularly in the more recent e-book efforts – and often as not lack the ability to interchange. Further, not all formats are able to support output to either paper or PDF that retain the typesetting ergonomic elements that readers are accustomed to. Finally, the conversion from analog to digital format is often much less than ideal, thus presenting the reader with text that has many times the number of character and formatting errors than that found in traditional typeset documents.

*Group InfoTech, Inc., East Lansing, MI, wjr@groupinfo.com

It is the purpose of this paper and a subsequent paper (see the "XML Europe 2000 Proceedings") to present a general manufacturing system technology specifically designed to convert large numbers of formatted pages into tagged, digital files that either retains the format of the original material or is easily and automatically reflowed into document structures that are not orthogonal to the original page.

METHODS

Figure 1 schematically illustrates the manufacturing workflow associated with large scale document conversion. This scheme has been well defined and is in large scale use for non-formatted document conversion.

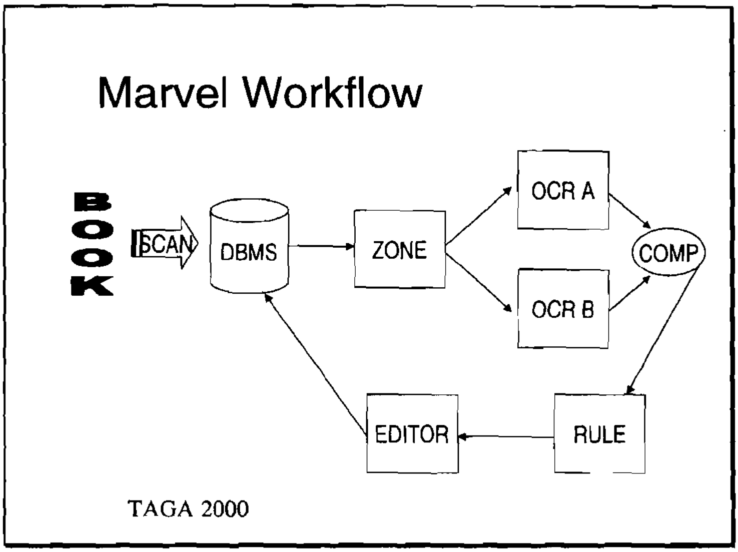


Figure 1: Marvel workflow

Physical pages are scanned into a database management system that is the manufacturing repository. The documents are then zoned into data type (e.g. line art, halftone, text, etc). Text data are sent to two OCR engines – OCR A and OCR B – which use different recognition technology (thus having different basic recognition error patterns). The resultant dual stream of data are then compared (COMP) for error type and likelihood of error based upon the recognition technology. Error markup and automatic correction are applied at this stage.

The compared data are then sent to a rule based system (RULE) that parses and analyses the data as word and sentence elements. At this stage a language independent grammar (LIG) process is applied to the converted text and error markup and correction occur.

Finally, error marked up data are sent to a dedicated editor step for error correction.

The primary use of Marvel (and its predecessor MagicBook) has been for unformatted data that will be used in targeted database applications such as legal data. Such data are "marked up" in SGML -- e.g. tags are applied in line to text data that define the searchable elements. These embedded metadata elements allow rapid keyword searching and Boolean association searches.

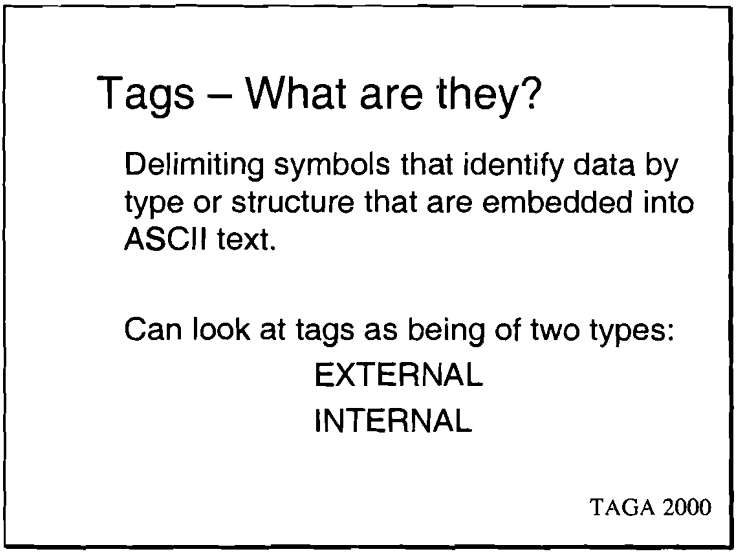
The MagicBook / Marvel technology was originally designed for formatted output applications, however, until the impact of the World Wide Web, the need for conversion of analog text data to formatted ASCII digital data was not well understood by the publishing industry. Furthermore, neither MagicBook nor the industry anticipated the advent of the numerous output formats required for the various forms of e-books and POD manufacturing systems. In particular the advent of the XML movement in recent months has both added to the number of output formats and complicated the question of what is acceptable output.

What is also interesting and, frankly, a problem is the fact that for the most part the early e-book providers were technologists rather than being drawn from the publishing industry. Therefore, basic techniques of reader ergonomics and text quality were, often, neither appreciated nor well understood. This has led to not only significant quality problems with certain converted documents but also has resulted in the inability of at least some of the available devices to get much better in terms of text and format quality.

With this in mind the Marvel conversion system was redesigned to meet the reality of the existing condition so that the best quality output for multiple targets could be generated from a single referenced database document file.

A NEW APPROACH TO TAGGING

As illustrated in Figure 2, tags are essentially delimiters that, when referenced to a definition table (the DTD), identifies certain elements or any subjects within a document text stream. In the academic setting tags traditionally do much more than identify document elements. One can commonly see tags for irony or other types of external metadata that can describe virtually any condition in the document.



Tags – What are they?

Delimiting symbols that identify data by type or structure that are embedded into ASCII text.

Can look at tags as being of two types:

EXTERNAL

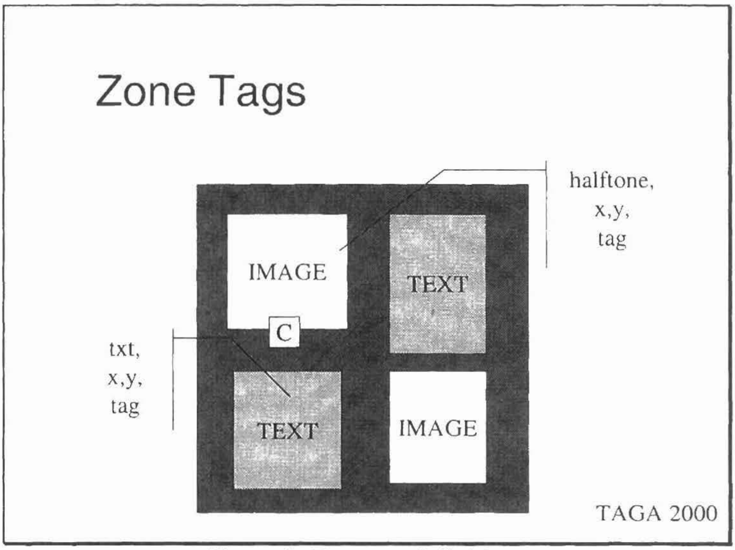
INTERNAL

TAGA 2000

Figure 2: Tags

However, we are concerned with manufacturing books rather than building a database of author intent. So, we are interested, for the most part, in things that can be externally described (e.g. author, image, caption and the like) and only in a few internal elements such as footnote coordination.

In the manufacturing process described above a zone step is required to identify elements of a scanned page image. This step traditionally separates image data (for later image processing) and text data that are sent to OCR. Figure 3 schematically illustrates this process. Note that this process has been modified in such a way as to allow marking not only the data type but also the coordinate (on the given page) and a tag for the zone type.


Zone Tags

TAGA 2000

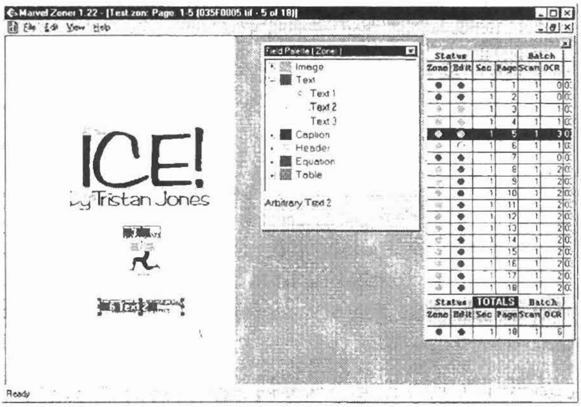
Figure 3: Zone tag definition.

Figure 4 illustrates the actual application process of tagging. Note that the live area (to the left) would show an outline in different colors depending upon the tag type specified. The center dialog box illustrates the DTD data element selection palette. The user selects a data type (a heading, body text, caption or the like) and draws a box around the relevant data type in the live area. The box "snaps to" the edges of the data selected and automatically applies the tag to the saved image file header. These tags "follow" the data through the manufacturing process and automatically apply themselves to the converted, cleaned ASCII.

Note that the right hand selection box in Figure 4 is a view of the production database for the scan batch being worked upon.

An important element in the zone tagging approach is the ability to change the definition of the tag types or to alias one tag set to another. Multiple DTD's can be supported by the application but it is recommended that the a metatag approach be used in high volume manufacturing. A metatag set is a superset of all possible data elements that can occur in a document.

Marvel Zoner



Status				Batch		
Zone	Edit	Sec	Page	Scan	OCR	
●	●	1	1	1	0	0:
●	●	1	2	1	0	0:
◇	◇	1	3	1	1	0:
◇	◇	1	4	1	1	0:
◇	◇	1	5	1	3	0:
◇	◇	1	6	1	1	0:
●	●	1	7	1	0	0:
◇	●	1	8	1	2	0:
◇	●	1	9	1	2	0:
◇	●	1	10	1	2	0:
◇	●	1	11	1	2	0:
◇	●	1	12	1	2	0:
◇	●	1	13	1	2	0:
◇	●	1	14	1	2	0:
◇	●	1	15	1	2	0:
◇	●	1	16	1	2	0:
◇	●	1	17	1	2	0:
◇	●	1	18	1	2	0:

Status	TOTALS			Batch	
Zone	Edit	Sec	Page	Scan	OCR
●	●	1	18	1	6

Figure 4: Zoner application and DTD tag set.

Metatags are arbitrary delimiters that can be mapped to multiple actual tag sets that are used in some form of output. As an example, a metatag set built into a

Meta Tags Provide:

Quark tags and styles.

HTML tags by dialect.

XML tags with or without style "trail" file.

SGML tags with DTD mapping.

Direct to PDF via Distiller or geometry program such as Quark or InDesign.

PDF tags for indexing.

TAGA 2000

Figure 5: Possible metatag fates

database table can "point at" HTML variants, XML, SGML or Quark tags (with an associated style sheet "trailor file" for complex formatting). So, from a single base ASCII set we can bind the appropriate output tag set on the fly. Figure 5 illustrates some possible fates that can be associated with a metatag set.

Thus we are able to apply something on the close order of 90% of the tagging needed for the average book without changing the workflow or the amount of labor required for processing the document through the use of the external tagging metaphor.

For republishing more academic material the ability to internal (or nonlinear) tags is required. This requires a post OCR editorial step that associates elements internal to the document to each other and (possibly) to the page of occurrence. This is accomplished by a modification of MarvelEditor (the EDITOR component of Figure 1). This more complex and less productive task and is described in the XML2000 paper.

REAL WORLD APPLICATION

The large bookstore chain Barnes & Noble has implemented a system to convert large numbers of physical books for digital output. The distributed manufacturing system spans three countries and parts are twelve time zones distant. Figure 6 illustrates the physical layout of the system. Note that the New York component is the job start point with scanning in Mexico City and conversion being done in Quezon City (a borough of Manila) in the Philippines. Note that, at the time of this writing, the primary production server resides in Michigan on the floor of the Michigan State University Computing Center.

Note the POD operations in Figure 6. These represent either existing or planned warehouse based digital printing operations or the planed use of in store printing fulfillment. Not illustrated are outputs through B&N.com that are digital delivery routes.

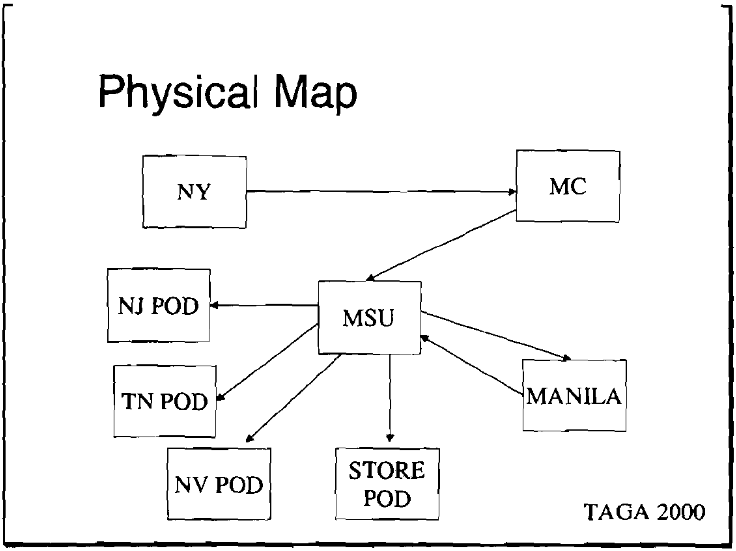


Figure 6: Physical topology of the B&N production system.

Figure 7 illustrates the logical workflow and intermediate data storage elements in the B&N operation. Note that the MC data store is the intermediate manufacturing store in Mexico City while the EL component is the master distribution store for the operation. QC is the Philippine production store.

This manufacturing system has been operation less than a year at this writing and continues to "ramp up" production. It is, however, producing typeset quality books at low cost in high volume. Further, this system is demonstrating that image object zone tagging is both possible and cost effective for reproduction with both the average trade book and the more complex academic style book.

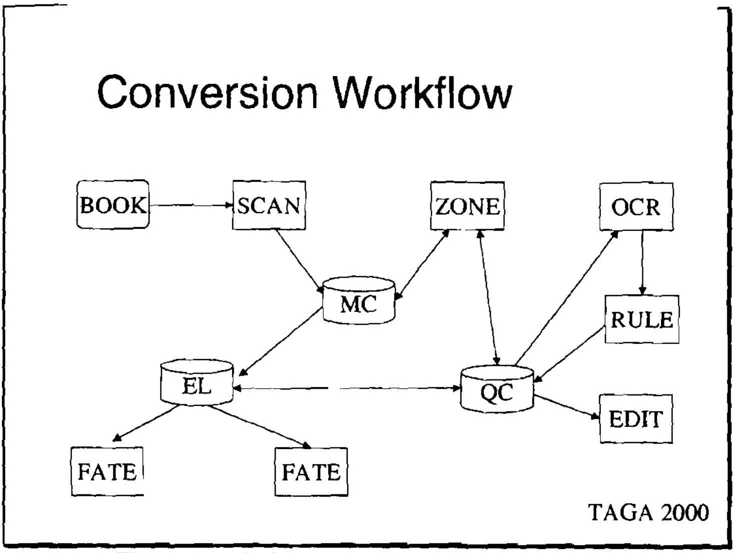


Figure 7: B&N production workflow topology.

CONCLUSIONS

Zone metataging with multiple output fates is both possible and cost effective when intimately combined with the OCR manufacturing process. However, not all tag systems are equal in terms of the formatted delivery quality to the end user. XML and HTML tags, while having a place, do not hold kerning (a key user ergonomic) and are thus more cumbersome in use.

Large scale manufacturing using these techniques is unequivocally possible and cost effective.

REFERENCES

William J. Ray. OCR, MagicBook and MagicMath: a Description of the Technologies and an Explanation of their Industrial Application. TAGA Proceedings, 1995.

William J. Ray. Large Scale Conversion of Book Backlists for Formatted, Digital Delivery: a Case Study, Part 2. XML2000 Europe Proceedings, June, 2000.

Stephen V. Rice, Frank R. Jenkins and Thomas A. Nartker. The Fourth Annual Test of OCR Accuracy. UNLV Information Science Research Institute 1995 Annual report, April, 1995.