WHITEPAPER

AutoCoding (*Finally*) Has its Day

By Sandra Serkes

1.0 Introduction

Used to be a time, not too long ago, when the use of computer software to identify and capture information from a document was considered preposterous, ridiculous, or at least very lousy quality. Labels like “dirty” and “raw” were routinely applied to the automated output, with commensurate disdain.

But times have changed. The automated identification, capture, classification and routing of document and content level information is becoming routine, even in the historically technologically-resistant practice of law.

This paper lays out how and why times have changed, what the changes mean, and how automation is (at last) having its day in the practice and support of litigation and other document-intensive legal matters.

In it, we will cover the origins, growth and future of automated capture technologies, focusing on one of the foremost capabilities now gaining mass-market acceptance, AutoCoding. We’ll briefly review the history of AutoCoding and its application to litigation support, as well as some novel applications of the technology for electronic, paper and “blended” populations. Finally, we will look at the cost-benefit ratios of different types of processes, architecting the best possible solution for different types of document needs.

Author’s Note: The first draft of this white paper was originally written in August, 2008. At that time, the paper focused primarily on the strategic reasons for the paradigm shift towards automation occurring in large-scale litigation document management. Since that time, economic concerns have exacerbated the shift, such that the theories and prognoses laid out in this material have already taken broad market hold 6 months later (February, 2009).

TABLE OF CONTENTS

ABOUT THE AUTHOR

Sandra Serkes is the President and CEO of Valora Technologies, Inc., a technology-based provider for the legal industry. One of Valora’s original founders, Ms. Serkes has been actively involved in Valora since its inception in 2000.

2.0 Glossary & Framework for Discussion

So that we are all on the same page, let's start with a few definitions and commonly understood uses of phrases and concepts discussed in this white paper.

2.1 What do we mean by "Automation" in the context of legal document management?

Automation is the use of computer technology (often blended with manual efforts for setup/training/configuration and quality control) to automate tasks that were historically performed solely by human beings (such as data entry) or previously not easy or even possible by human beings (such as complex pattern-matching calculations). Typically, automated solutions follow these 6 steps:

1. They are developed in advance of their general usage
2. They are custom-fitted, as needed, to a particular purpose
3. They run by themselves, often in high-volume, parallel-processing environments
4. Their results are evaluated for performance, and often,
5. They are re-tooled and run through the first 4 steps again
6. In some cases, "last mile" improvements are made to the automated results to increase final performance

2.2 What is a Document?

A document is an organized set of text, graphics and/or tables, with defined boundaries (start/end), which was created for a specific purpose and has a purposeful organization. A document is distinct from raw *content* (which comprises the document), from an organized *group* of documents (such as an Encyclopedia) or from documents *components* (such as the letterhead, the body or a Distribution list).

There are many definitions of what a document is and is not. For the purposes of this paper, a document[1] can:

1. Be handwritten, printed, typed or generated.
2. Be in any spoken, written or visual language, including computer code, laboratory results or other electronic "languages"
3. Exist on paper, in image form, in electronic form or be generated temporally in response to certain human or electronic stimuli (ex: fax error report, ATM receipt, online travel itinerary, etc.)

[1] This definition of a document is different from, although similar to, the legal definition of ESI (Electronically Stored Information). ESI refers explicitly to how electronic files (and paper documents converted to electronic formats) are stored and managed by litigating parties. ESI definitions concern storage and management of documents, *once they have been created*. Our definition is simply that of *what constitutes a document in the first place*.

The following are all examples of documents, some with attachments (an intended or implied relationship between multiple documents):

1. A handwritten thank-you note.
2. A fax cover sheet (often with additional attached documents being transmitted)
3. A PowerPoint presentation[2]
4. An email message (often with additional attached documents being transmitted)
5. A shipped bill of lading
6. A Cash Flow Statement
7. A Deposition transcript
8. An Agreement with referenced Exhibits. Typically Exhibits are themselves unique documents, with an implied subordinate (child) relationship to the parent Agreement.

The following are also considered documents under the definition given above:

1. An instant message/text message
2. A voicemail
3. A labeled file folder (physical or electronic)
4. An embedded original email "string" that is the origin of an email conversation thread. In fact, each of the subsequent responses and forwards of the conversation is also a distinct document, with a unique Hash value, creation date/time, author, recipients and so on.
5. A passport

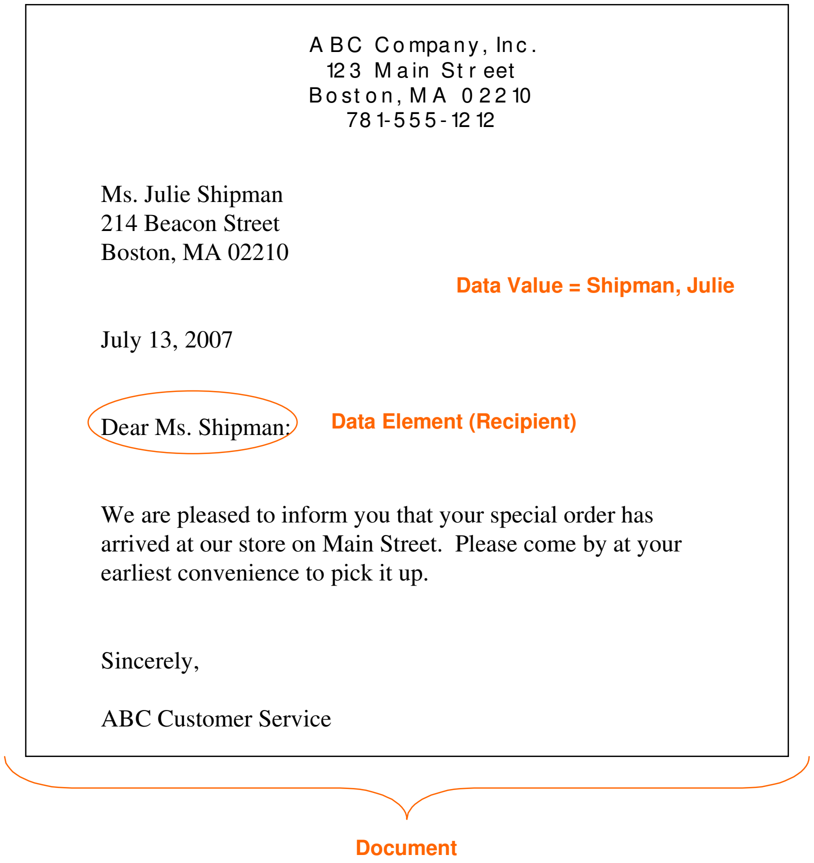
2.3 What is Automated Data Capture?

Automated Data Capture is the practice of using specialized software tools to identify documents, identify data elements within those documents, extract (or "capture") the data values from those data elements and place the values into a meaningful organization per document (typically a database with one document per record).

Pictorially, it looks like Figure 1 on the next page:

[2] The presentation itself is the document. The individual slides are component parts of the document, not documents in themselves.

Figure 1: A Document Being AutoCoded



ABC Company, Inc.
123 Main Street
Boston, MA 02210
781-555-1212

Ms. Julie Shipman
214 Beacon Street
Boston, MA 02210

July 13, 2007

Dear Ms. Shipman:

We are pleased to inform you that your special order has arrived at our store on Main Street. Please come by at your earliest convenience to pick it up.

Sincerely,

ABC Customer Service

Data Value = Shipman, Julie
Data Element (Recipient)
Document

In litigation support, automated data capture is typically called AutoCoding, AutoTagging or AutoReview. The most common form, AutoCoding, is the automated capture of bibliographic (objective) coding fields per document. The typical bibliographic fields captured in AutoCoding are: Document Author(s), Intended Recipient(s), Copyees (CC) and Blind Copyees (BCC), Document Type, Date Created and Subject or Title of the Document. For the document shown above, the correct bibliographic field values are:

Document Author	ABC Customer Service
Intended Recipient	Shipman, Julie[3]
Copyees	<none>
Blind Copyees	<none>
Document Type	Letter
Date Created	07/13/2007[4]
Subject/Title	<none>

More sophisticated AutoCoding engines, also with **AutoUnitization** engines, capture the *relationship* of documents to one another and their logical boundaries. The example above is a one-page document, with no attachments, so its proper Document Boundary and Attachment Range information is:

[3] Normalized format of Julie Shipman.

[4] Again, normalized data, this time standard American date format.

Begin Doc	001-001-0001 (or whatever its Bates or other assigned label is)
End Doc	001-001-0001
Begin Attach	001-001-0001
End Attach	001-001-0001 (no attachments)

More complex documents that have cover pages, attachments, signature pages, certificates of service and other implied relationships are captured and reported in a similar manner, where the attachment range indicates the relationship of the distinct document elements.

2.4 Objective Data Capture Vs. Subjective Data Capture

Thus far we have spoken primarily about objective data capture. That is, data capture that is rules-based, where the rules remain the same for nearly all document populations. For example, the date of the letter above is both obvious and indisputable (typically), the very definition of "objective." In litigation support, objective data capture has come to mean well-understood, simple data capture than can be easily outsourced to others with little to no case matter knowledge.

Subjective data capture means that the value of the data element can change depending upon its context. For example, the exact same letter may become privileged when the author/recipient parties are attorney and client, rather than supplier and customer. This subjectivity based on context forms the basis for most legal document review that occurs in litigation. The prevailing wisdom is that the subjective coding practitioner should be schooled in legal practice (or at a bare minimum in the key concepts germane to the matter) to be able to provide the appropriate subjectivity when evaluating the document.

Again, however, technological improvements coupled with economic realities are shattering that provincial view. Like bibliographic coding, subjective document review can typically be boiled down to a few core concepts at the "top level." The rise and popularity of outsourced "first pass review" services is based on this notion. Today, outsourced personnel (and thus soon to follow automation technology) routinely perform first pass document review tasks, such as: determining presumptive privilege, assessing responsiveness, labeling "hot docs," and identifying key issues and themes. Each of these is readily automatable, and a natural extension of AutoTagging or AutoCoding capability.

Furthermore, there are additional review-ish tasks that are not particularly subjective in nature, and highly automatable. Grouping documents by their relative similarity (also called Near Duplicate Detection), or by their conversational thread nature

A History of AutoCoding

AutoCoding, as a technique for data capture in the legal industry, became technologically and commercially viable when two things happened. First, computer processing power became fast enough and cheap enough for hundreds of thousands of pattern-matching algorithms to run in real-time, parallel-processing. Second, coding of litigation documents had become standardized and mainstream enough that its rules could be codified and the work outsourced.

To be sure, the early results were far from perfect. In fact, in many cases auto-coded output was downright dreadful! However, like most new innovations, there was nowhere to go but up[5].

And up it went. In the early days of auto-coding, performance was generally considered to be at 40% yield, a combination of accuracy & coverage measurements. This meant at least two errors for every document! Today, yield rates exceed 98% in certain circumstances (notably ESI populations), with average yield at about 85-88%, more than double the results of only 8 years ago and still improving.

[5] There have been many studies and analyses of the performance improvement curve of early innovations, from the first automobiles, which used steam to power their engines, to the first cell phones, which weighed over 3 pounds and needed their batteries recharged after 60 seconds of usage! For an excellent discussion on how innovations improve technologically over time, read The Innovator's Dilemma, by Clayton Christensen, HBS Press.

(also called Email Thread Grouping) are techniques to group like-minded documents together to speed up the document review effort. Much of this categorization, classification, and grouping capability has been automated for several years. It is only taking hold now, however, due to the explosion of the number of documents requiring review.

2.5 Custom Data Capture

Finally, the most sophisticated AutoCoding engines have the ability to custom capture almost any kind of data element imaginable. From in-content types of data elements (such as names of parties listed, organizations mentioned, etc.), to line item data capture (listings on a phone bill or credit card receipt, filled-in forms, etc.) to incidences of intent, tone or usage history (presence of hostility, presumptive privilege, change in style or format, etc.) to classification and culling mechanisms (issue codes, presumptive relevance/responsiveness, duplicates, etc.), AutoCoding is today being used in many complex and unique automated data capture situations.

One of the most efficient uses of custom data capture is for records retention and backfile conversion projects. The low cost / high-speed nature of automated data capture is a natural fit for such large document populations. More and more medical, real estate, oil & energy, personnel and financial records are being automatically indexed by high-volume, customizable data capture solutions. The best of these solutions can be initially overlaid on top of an existing manual effort. The manual work is gradually phased out until it is saved only for the most complex or unusual data capture work, saving the bulk of the effort for low cost / high speed automation.

3.0 Big Changes

While AutoCoding technology continued to advance over the last decade, (See Sidebar: A History of AutoCoding), some other sea changes were simultaneously taking place.

3.1 From Paper to Electronic

First, the nature of litigation documents was changing, from primarily paper in origin, to primarily electronic in origin. This meant that the text quality of the documents (typically extracted during ESI processing or embedded into the native files themselves) was much improved over the OCR text from scanned paper documents. Since the text record is the basis for nearly all AutoCoding activities, AutoCoding's yields rose commensurately. In fact, combined with the use of data normalization technology[6], the better text base actually improved AutoCoding results *exponentially* (rather than simply linearly).

[6] Data normalization is a technique where similar data elements are grouped together and analyzed for similarities and probability of association. For example, the colloquial forms of “Bob” and “Rob” are both normalized to the more standard form of “Robert.”

3.2 Document Volume Explosion

The second change occurring outside of pure technological development is the size and scope of today's document populations. Not only are most documents now electronic in origin, the overall *number* of documents has exploded, with the inclusion of electronic communications as discoverable material. Remember, each email and all its embedded strings (Re, Fwd, BCC, and so on) and attachments is a separate document. Because the scope of the problem had magnified so greatly, simple document-by-document solutions were becoming too costly and unwieldy to maintain. Automated approaches to data sorting, tagging and culling were needed. With the problem's origins, so came its solutions – automated methods for understanding and managing such large volumes of electronically-created data.

In short, the marketplace has come to accept two key components of AutoCoding: 1) that automation can and *must* be used to manage such large document populations and 2) that solutions that get most of the way there (as opposed to perfectly there) are *good enough*, particularly given the cost to go the last mile.

This last point is quite significant as it is a radical departure in thinking in the legal industry. Long known for its attention to detail and need to cover every angle, the legal profession exacted tremendous concessions from coding and other document processing vendors in the past. Vendors were made to produce, defend and guarantee performance metrics of all kinds (accuracy, turn time, defect rates, and so on.). Detailed Coding Manuals were created, revised, signed off upon and used as "CYA" fodder in pricing and payment disagreements.

However, the advent of commonplace ESI processing changed the rules of acceptable performance to be far more reasonable and attainable. For example, it is generally accepted that ESI processing will simply yield some portion of "un-processable" documents. Perhaps they were misnamed, perhaps password-protected with unbreakable codes, perhaps files were embedded with viruses, and so on. In any case, it is today expected that a portion of files shall remain unknowable and unobtainable, except under the most exacting circumstances (and sometimes even then). In a similar vein, the courts stess the notion of "reasonable" methodology and expense in managing and producing such vast quantities of material, leaving open the likelihood that there would quite reasonably be some portion of material simply not processable. All of this leads the industry to a gradual acceptance of "good enough" as the new standards bar, rather than "perfect." Once "good enough" became the norm for ESI processing, any number of complementary services were free to (indeed asked to) follow suit.[6]

[6] As a specific case in point, Valora used to not sell its "raw" AutoCoding output to customers, out of consideration of the less-than-perfect performance as a stand-alone result. Today, we do sell our pure AutoCoding, in direct response to many, many customer requests, with the explicit understanding that the standalone results are "good enough." In 8 years, our AutoCoding request rate has gone from 1/1,000 to 1/4. For 2009, we are projecting 1/3 requests will be for AutoCoding.

3.3 Economic Downturn

Finally, amidst growing acceptance of AutoCoding as a “good enough,” low cost/high speed option, an economic upheaval of unprecedented proportions took place. In the last six months, low-cost, high-speed options look increasingly palatable as real-world, optimal solutions. Particularly as more and more law firms and the litigation support vendors who serve them find themselves in a financial squeeze, automation is fast becoming the way out.

The perfect storm of tight credit, exploding volumes of documents, better text to start from and acceptance of good enough quality levels means that at last AutoCoding is having its day.

4.0 Getting Started: *How to Use Auto-Coding and Other Automation Techniques for Large-Scale Document Populations*

There are lots of automation options available for use in litigation support. They include:

AutoUnitization	Identifying physical and logical boundaries between documents and preparing any associated attachment ranges
AutoCoding	Bibliographic coding & custom data capture (coding)
AutoReview	Near duplicates, EmailThreadGroups, Presumptive privilege, Responsiveness, Issue-tagging
Population Analysis	Pre-emptive ESI analysis, based on population fact reporting. Includes reporting on document type distribution, duplicates, issues, names and cost forecasting
AutoPrivLog	Redacting and labeling privileged documents with bibliographic information

So, where to begin? How can you use AutoCoding wisely? What should you expect?

4.1 Crucial Questions

The best place to start is with a quick evaluation of your document population(s) and your goals. In short, ask yourself what you have (or think you may have) and what you want. Here are some examples to ask about what you have:

1. What was the origin of the documents? Are they paper, electronic or a blend?
2. What kind of case is it? IP? Construction? Contracts?
3. Are the documents clean? Are they easily readable? Are they first generation? Are they filled with computer codes or formatting? Are they handwritten?

4. Do you suspect privileged material might be present?
5. Do you suspect there to be significant overlap (duplication) in documents?
6. What kind of budget do you have? What is at stake? I have always found it helpful to ask clients the following question in this regard: if this case matter were a car, are you driving a Yugo (very cheap but drivable), a Honda Accord (best value for the dollar) or a Rolls Royce (top line everything, cost no object)?
7. What timeframe are you working under? Again, do you need everything drop dead immediately (yesterday?), in a reasonable timeframe or not for some time? In all cases, a preliminary assessment of the material using AutoCoding can yield very helpful results for expectations, planning and resource allocation.

Here are some example questions to ask about what you want:

1. Are people important in the matter? Do you need to know who knew what when?
2. Will you need to make a timeline of events? (Will you need Date information for all documents?)
3. Are issues and topics important in the matter? Will you need to group documents by subject matter or specific contents?
4. Will you be using outsiders to help you review the documents? How will you communicate to them what is important and what is not?

4.2 Population Analysis

Once you have answered the questions above, you are ready to proceed with automation. Valora recommends you pull together a sample set of documents, generally between 100-1000 documents or 0.5-1 GB of electronic files. Once the sample is ready, ftp it to your AutoCoding provider and have them perform a preliminary assessment and run the AutoCoding process. It should take less than 24 hours all told.

The AutoCoding results should speak for themselves. Either the AutoCoding is strong, or it isn’t. Either the documents lend themselves to automation or they don’t. A quick conversation with your provider should answer any remaining questions, including how best to customize the AutoCoding to your specific needs.

Sophisticated AutoCoding vendors will also provide you with a **Population Analysis** report to help forecast the contents and likely automatability of the remainder of your documents. Population Analysis (a.k.a. early case assessment or pre-conversion) is an approach to large-scale document litigation that provides detailed, predictive information about documents in order for the litigation team to make smart, informed choices about how best to proceed in managing them. For more information on Population Analysis, see Valora's article Population Analysis: Litigation Support's New Best Friend in the Winter 2009 edition of the CALSM Newsletter: http://www.calsm.org/nlarticle.php?id=11.

4.3 AutoCoding as a Smart ESI Complement

One of the most logical places to look to AutoCoding is in conjunction with ESI processing. Most people use ESI processing to capture the text of electronic documents (a simple form of automated data capture) and have a quick and easy native "link" that launches the original electronic document right from the database record. Although some people also use ESI to create image renderings of the files and sometimes to capture any available metadata, these latter processes are often seen as expensive luxuries, not worth the additional cost.

Consider instead the addition of AutoCoding at this point. Remember that these are electronic-origin documents, which means they will have a strong, natural text base. Adding AutoCoding as a complementary step to ESI processing provides back all the typical litigation fields expected: Author, Recipient, etc. (see list above.) But, data is normalized, making it easy for searching and sorting. Metadata is not normalized and often results in 14-17 variations of key parties! In addition, metadata can often be very weak yielding "Registered User" as the Author and "file.doc" as the Subject. AutoCoding is a low cost way to supplement or supplant metadata.

4.4 Further Automation Options

AutoCoding is but one of a host of automated processing options for paper and electronic documents. Other related Auto services include: OCR, AutoUnitization, StraightThrough™ ESI Processing, SmartSelection™ ESI Processing, AutoPrivilege, AutoNearDupes, AutoEmailThreadGroups, AutoIssues, AutoReview, NamesMentioned and AutoTranslation. For more about Automation in general and Automated Services in Litigation Support, see Valora's website at: http://www.valoratech.com/allauto.

About Valora Technologies, Inc.

Valora is a technology-based provider of automated document analysis and review services for the legal industry. Valora offers services for paper and electronic populations to law firms, government agencies, corporate legal departments and litigation support organizations around the world.

Valora has developed a strong expertise in the processing, management and analysis of both large and small cases with short deadlines. The company's specialty is providing efficiency, organization and cost control. In its profile of Valora, The American Lawyer magazine noted: "[Valora] gets the profession closer to something truly overdue - the semiautomated review of litigation documents."

About Sandra Serkes

Sandra Serkes is a dynamic leader with an extensive background spanning over 20 years in software marketing, product management and corporate strategy, particularly in document processing, computer telephony and speech recognition. Today, Ms. Serkes oversees Sales & Marketing, Finance & Administration, Operations, Engineering and Corporate Strategy.

A graduate of both Harvard Business School and MIT, Ms. Serkes is a frequent industry speaker and panelist. She is an active participant in the Women Presidents' Org., The Commonwealth Institute, the MIT Enterprise Forum, the Massachusetts Software Council and the Network of Harvard Alumnae. Ms. Serkes serves on the boards of several technology and service start-ups. Ms. Serkes was named a 2006 "Woman to Watch" by Womens' Business magazine.