

# CALSM NEWS

[www.calsm.org](http://www.calsm.org)

## The Low-Down on Automated Document Encoding

By Sandra Serkes, Valora Technologies

### Introduction

Just when you thought you had document production under control, along comes a new, almost impossibly tantalizing idea: auto-coding of case documents. No more mind-numbing data entry of Author, Recipient, Date, Subject Matter, and so on for the paralegals and associates! No more crossing your fingers as you ship your highly confidential data overseas. Auto-coding's promise is that with the help of a simple software program, thousands of hours of manual data entry will be automated and eliminated.

Is it too good to be true? Well, yes and no. Automation has made tremendous strides in reducing the hours and effort involved in document coding, but we are not ready to replace human effort altogether. The most promising solutions offer a combination of human and machine skill, resulting in ever faster, cheaper and more accurate results.

### Why Auto-encoding is the Holy Grail

Document Encoding has been around for a long time, probably since the first days of discovery and certainly once the first attorney became overwhelmed with discovery documents. Without an indexing technique to control and manage the paper, modern litigation would be impossible. Early approaches to indexing involved stacks of notebooks where assistants entered data for each document into the relevant row and column - rather similar to today's computer spreadsheets. With the advent of databases (of which spreadsheets are really a primitive form), fielded data entry became the standard technique. The combination of ever-increasing volumes of paper and the need to rein in encoding costs has forced the practice to an outsourced model, often involving cheap overseas labor in developing countries.

The promise of a software program that could do everything human beings do and

*continued on page 2*

### Calendar of Upcoming Litigation Support Events

**October 22-27, 2003** — Summation training class to be held by Martin Consulting Group in Minneapolis, MN (October 22), St. Cloud, MN (October 23), and Moline, IL (October 27). Contact Connie Martin at (651) 322-4980 or [Connie@MartinConsultingGroup.com](mailto:Connie@MartinConsultingGroup.com).

**October 29-30, 2003** — Chicago LegalTech Conference. Lela Laurent at Daticon has extra exhibit hall passes available. She can be reached at [llaurent@daticon.com](mailto:llaurent@daticon.com)

**November 5, 2003** — CaseSoft Certification Seminar in Washington, DC. These workshops are for litigation support, practice support and IT professionals, not end users. It is also the key first step for those interested in becoming certified to train or support CaseSoft tools. If you have questions, or wish to enroll, please contact Judy Herndon at [jherndon@casesoft.com](mailto:jherndon@casesoft.com) or (904) 273-5000, ext. 231.

**November 14, 2003** — Electronic Discovery and Records Retention Conference (Chicago) hosted by Glasser LegalWorks. For further information please go to [www.legalwks.com/conferences/ediscovery/index.htm](http://www.legalwks.com/conferences/ediscovery/index.htm). Glasser LegalWorks is offering a \$200 discount to all active and prospective CALSM members who are not already registered. Just mention CALSM when you register.

**December 17** — CALSM Meeting: LiveNote, to be held at McAndrews, Held & Malloy

**April 7-9, 2004** — Daticon Educational Conference to be held at the Mohegan Sun (<http://www.mohegansun.com/index.jsp>) in Connecticut. Agenda is currently being developed. Count on special Ediscovery focus sessions featuring Chuck Kellner and Gary Carignan, Ediscovery industry veterans and experts.

## CHICAGO ASSOCIATION OF LITIGATION SUPPORT MANAGERS

AUTUMN NEWSLETTER 2003

### OFFICERS:

Barbara C. Hanahan, President  
[bhanahan@winston.com](mailto:bhanahan@winston.com)

James B. Salla, Newsletter VP  
[jsalla@jenner.com](mailto:jsalla@jenner.com)

Michael Weiler, Membership VP  
[mweiler@mhmlaw.com](mailto:mweiler@mhmlaw.com)

Alison L. Weinberg, Programming VP

Sarah Mallon, Secretary  
[smallon@sachnoff.com](mailto:smallon@sachnoff.com)

Joni K. Eskridge, Treasurer  
[jeskridg@skadden.com](mailto:jeskridg@skadden.com)

## In This Issue:

### The Low-Down on Automated Document Encoding

by Sandra Serkes . . . . .pg 1

### Showcase Your Legal Department

by Alison L. Weinberg . . . .pg 5

### Calendar of Litigation Support Events

. . . . .pg 1

## New and Noteworthy:

On August 27, Andrew D. Richmond of KPMG made a presentation to CALSM on KPMG's Forensic Technology Services and their web-based solution, Discovery Radar. For those of you who missed the meeting, more information about Cypress Technology Center can be found at [www.cypresstechnologycenter.com](http://www.cypresstechnologycenter.com). Andy can be reached at (312) 665-5397 or at [adrichmond@kpmg.com](mailto:adrichmond@kpmg.com).

more is irresistibly seductive. "Why, if I only had access to this magic program," so the thinking goes, "I could encode every document that was ever created!" Associates or paralegals would run the software each night and by morning all documents would be instantly encoded, at almost zero-cost and with perfect accuracy and judgement. This is about as close to the Holy Grail that litigation support could ever wish.

If only. The flaw behind this type of thinking is not so much the notion of almost zero cost to encode - that part is correct - but rather the notion that one software program could ever replace the judgement and expertise that human encoders employ. Indeed, such a software program would need almost constant revision and training, resulting in much higher costs, and would still be far from what a trained human being could produce.

But don't despair. Although we are not yet at Holy Grail levels of accomplishment, much has been happening in the technology space to combine the efforts of human beings and software programs to create a joint encoding technique - with very promising results. The rest of this article details the science and application of document encoding as it is being conducted today, assuming a combination of programmatic and human skill. Usually called Automated Document Encoding, this paper outlines the technique, its benefits to the legal community and its best applications for success.

### Defining Document Encoding and Approaches to Creating It

Document Encoding is the systematic process of entering descriptive information about each document into a database. By building this database first, documents relevant to the matter at hand can more quickly be identified and retrieved, using a combination of searching and sorting. A document database is often combined with search of the documents' full text (OCR). The database and the full text are complementary: the full text can provide every reference to John Smith, if that's what's needed; the

database can provide only those documents authored by John Smith during January of 1998, if that's what's needed.

Currently there are three basic approaches to performing document encoding: manual, automatic and automated.

### Manual Encoding

Manual encoding is familiar to most people, and is the most dominant form of coding in use today. Manual coding is, in essence, data entry. Trained personnel sit with documents (or images) on one side and enter fielded data into a database on the other. The work is slow, tedious and inherently variable from one person to the next and even between one person's output at one point versus a later point.

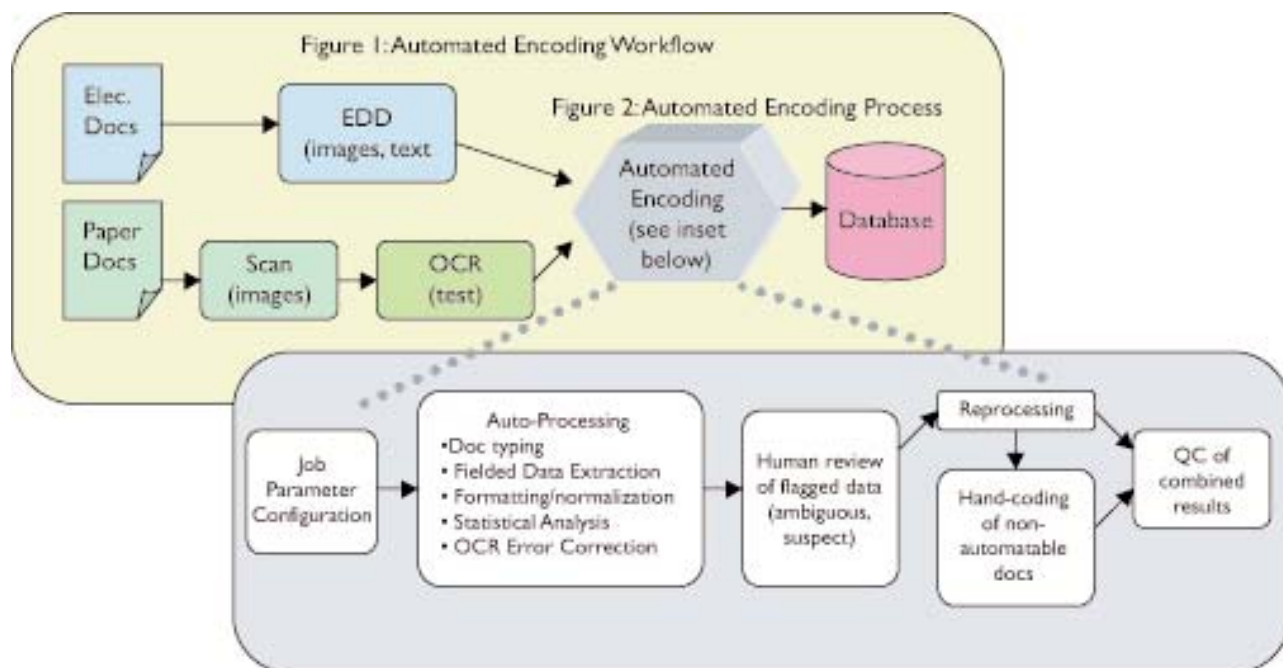
### Automatic Encoding

Automatic encoding uses computer software to accomplish the encoding task to the best approximation possible. Software rules are written to identify and process documents as a person would, with obvious limits to flexibility and judgement. Automatic encoding has almost no human effort, apart from running software. Automatic encoding relies on OCR output to interpret documents and is typically hampered in much the way search engines are by poor or limited OCR. The advantages to automatic encoding over manual encoding are its speed (often measure in hours, not days or weeks) and its cost (typically 75-80% of manual coding costs). Its major drawback is its poor quality (80% accuracy, at best; 30% or lower, as worst).

### Automated Encoding

Automated encoding is a new approach that attempts to blend the best elements of manual and automatic approaches. The goal is to combine the speed, consistency and price point of automatic encoding with the judgement, flexibility and accuracy of manual encoding. The result is a highly accurate database (often more accurate than manually produced databases, due to the emphasis on consistency and normalization) delivered in a fraction of the

FIGURE ONE



time and cost of manual efforts, with few of the quality drawbacks of automatic encoding.

### The Construction Analogy: Digging a Basement

It is often helpful to use a construction analogy to understand the various approaches to document encoding. Suppose you need to dig a hole, say, the basement for a new town hall.[end note:<sup>1</sup>] The manual approach is to hire a hundred men, give them all shovels, and get out there and start digging. Eventually, you'll have your basement, albeit with variations in depth, edging and shape according to the variability of each shoveler's abilities.

The automatic approach is to grab some dynamite, light the fuse, toss it in and - boom! - instant basement. It's fast all right, but not very exact and certainly nothing you would try in a highly populated or nature-preserved ecosystem.

Now consider the bulldozer. It is relatively fast (much faster than shoveling, but slower than dynamite), relatively accurate (and easy to standardize) and relatively safe (operated by a trained technician with limited danger to others).

In a cost-benefit tradeoff, the bulldozer clearly wins. It can be brought in specifically for the task, performed quickly and with expertise, and produces end results that are satisfactory for what is needed.

Unsurprisingly, most foundations are dug by bulldozers today, rather than with shovels or dynamite, although neither shovels nor dynamite have been rendered obsolete by the bulldozer. Instead, the task of digging has specialized into different niches. For your basic basement, the bulldozer is usually the best answer. But for a quick, small hole - perhaps to plant a tree - a shovel works well and is the most cost-efficient technique. And for a large train tunnel in a mountain, perhaps to be dug through rock, dynamite is the only technique powerful enough to even accomplish the task.

And so it is with document encoding. For basic encoding needs - bibliographic fields and more-or-less generic document types - automated encoding is usually the best answer. It is fast, accurate and cost-effective. But for highly specialized material or unusual or subjective field requests, manual encoding will deliver better results. And for project with tiny budgets or immediate time-frames, automatic encoding might be better than nothing at all.

### How Software Processes Documents

There is a certain black-box, smoke-and-mirrors component to technology that can make people uncomfortable. They don't know - or can't see - what is really going on. Software is made up of lines of code, rather like the algebra proofs you had to do back in eighth grade. The code itself is a list of very simple instructions and rules that the computer follows, one after another. Just as document encoders are trained on any particular coding project (given a set of coding rules), so is software trained (given a set of instructions). But the computer instructions are much simpler, and there are many more of them.

It is helpful to understand what it uses as inputs and what it produces as outputs. People use visual cues as inputs when they encode documents. The author is the script name at the bottom of a letter. Maybe it is also given on the letterhead. People work with actual documents, or images of documents.

Software also uses visual cues, but cannot interpret an image as people do. Instead, it interprets text -created by an OCR process or provided natively from electronic documents. By reading the

text and looking for visual and contextual clues, software attempts to determine a document type (or several possibilities). Once it has made a determination of document type, it attempts to uncover the relevant fielded data that is required by that document type. For instance, if the document is an email, then the software will look for a date, a single author, one or more recipients, a possible CC and a possible subject line. Letters frequently end with terms such as "Sincerely" or "Very truly yours" followed by a signature and name. Software can use this to determine the document author.

When the software has completed its analysis of a document, it records the fielded data into a database record and stores it for later export to the litigation support program of choice (Summation, Concordance, Introspect, etc.). This process is repeated for each document until the end of the collection is reached.

The description thus far applies to both automatic and automated encoding. But this is where automatic encoding stops. Automated encoding employs a number of additional techniques, including trained, professional QC staff, to improve the quality of the output.

Automated encoding typically includes additional steps for name normalization, document templating, duplication detection, statistical sampling and confidence reporting.[end note:<sup>2</sup>] These steps are meant to assist QC staff in interpreting and improving the quality and completeness of the machine-produced results. Most automated encoding processes involve several iterations of computer and human interaction, with increasing the quality metrics each time.

For example, software can find duplicates and near-duplicates of a document and group these together. This allows QC staff to check the similar documents together, resulting in faster processing and more consistent output. Similarly, an automated approach can identify anomalous data - data that seems inconsistent with the rest of the database - and identify this to the QC staff, allowing them to find, review, and correct potential errors quickly.

While automatic encoding may lend itself to being packaged into a standalone software product, automated encoding does not. Automated encoding relies on a combination of trained QC effort and QC tools working in conjunction with the back-end encoding software. Furthermore, the back-end encoding software is often being updated and revised, often in response to the needs of a particular document population. Replicating this organic approach is difficult and costly, and best suited only for high volume situations.

### When to Automate

Automated encoding is a good, all-purpose solution for most encoding jobs, but not for all. It is important to understand when to automate and when not.

Everyone would like the highest quality output at the lowest possible cost in the shortest amount of time. These parameters are always in place. But to determine whether a document population is well-suited for automation or not, consider these additional parameters.

### Handwriting

With current technology, handwriting is not suited to automation. Handwriting does not provide useful OCR output and some-

times interferes with typeset content, for example when someone signs through their printed name. Fully handwritten documents, such as notes on a napkin, cannot be automated at all; they must be hand-encoded. However typeset documents with only some handwriting can often be fully or at least partially automated (and completed using human QC as necessary). The typeset letter where the only handwriting is the author's signature is a good example. Four of the five required fields (recipient, date, doc type, subject) are automatable and only the fifth (author) must be entered by hand at QC time.

Unless your population is almost entirely hand-written (and likely predates 1970), chances are good that some or most of your documents are automatable. A quick visual scan of the documents should give you a rough idea of the prevalence of handwriting. A more formal statistical sampling of the documents would answer the question definitively.

### Document Types

Most cases involve documents of a familiar type. There are correspondence documents (letters, faxes, email, memos) financials (SEC filings, budgets, forecasts), business transaction documents (agreements, invoices, bills of lading) and a host of others that are generally recognizable and understood by most litigation support professionals. Such documents are a good fit for automation as they are similar from case to case and within the population itself. As a rule of thumb, consistency bodes well for automation.

It is the documents that have you scratching your head that similarly cause problems for automation. Does it fit into more than one category? Who generated it? How was it used or intended? If these questions are hard to answer, the document will be hard to code, whether by hand or with automation. Another rule of thumb is: If you can easily explain how to code a document, then it's probably automatable. If you might code it one way and your colleague another, then it's probably best to leave it up to human effort entirely.

### Image and OCR Quality

Most imaging and OCR vendors will grade the materials to be processed, partly to determine the price of performing the service. A low letter grade (A or B) is actually a high quality grade, indicating that the pages are clean, with clear text, and in good physical shape. A higher letter grade (C-F) indicates later generation copy, tears and bent corners, and a high incidence of binding materials (staples, folders, paper clips.), which add to the document preparation costs.

Document grading is also useful in estimating fitness for automated encoding. As you might expect, lower letter grades are better candidates for automation, because of the clean resultant image quality. Look at a few documents. Are they easy to read? Can you make things out without magnification? A rule of thumb here, too: if you must magnify document images more than twice to be able to read them, the OCR will likely be poor quality and the document is not a good fit for automation. (Incidentally, you may wish to do this before ordering OCR on your population. If you can't read it after reasonable manipulation, neither will the OCR, so don't waste the money!)

### Fields to Encode

There is an infinite variety of fields that can be encoded in docu-

ment populations, although most people gravitate towards "basic bibliographic" fields: author, recipient, copyees/BCC, date, title and document type. While manual encoding often adds considerable cost to capture additional fields, particularly in-content fields such as Names Mentioned and Key Words, automation can often encode such fields very cost-effectively. Most people would like to have Names Mentioned and Key Words as part of their standard encoding fields, but have been prohibited by cost or time in the past. Automation can often provide these fields with little, if any, additional time delay and moderate cost increment.

### Incorporating Automation into the Document Production Workflow

Now that you know when to encode using automation, here is how the whole thing goes:

### Conclusion

For document populations with the right characteristics, automated document encoding is a compelling and competitive alternative to manual processing. By combining the strengths of computer processing and human judgment, automation reduces costs and turnaround times, and increases quality. The combination approach of human skill and computer programming is a relatively new technique in the legal community, although it is widely used in other industries such as customer relationship management and factory automation. Automated document encoding has been successfully employed by law firms, large and small, litigation support vendors and the American Bar Association.

### About the Author

*Sandra Serkes is the President of Valora Technologies, the leading provider of automated document encoding solutions to the legal community. She is an authority on automating document processing and speaks frequently about the topic. Her undergraduate degree is from MIT, and she holds an MBA from the Harvard Business School.*

*To learn more contact Hendon C. Pingeon of Valora Technologies. He can be reached at: [hpingeon@valoratech.com](mailto:hpingeon@valoratech.com) or 781.642.8806.*

### ENDNOTES

1 Apologies in advance to Virginia Lee Burton, much-loved author of the classic children's book, *Mike Mulligan and his Steamshovel*.

2 So as not to go too far down the technical discussion path, these terms are explained in the Glossary at the end of the article.

Copyright (c) 2003 Valora Technologies, Inc. All Rights Reserved. 09/20/2003

### GLOSSARY

#### *Automated Document Encoding Terms*

**Attachment Range:** the delineation of one or more documents that are linked together, typically by content association or a physical binding mechanism (paper clip, staple, folder, etc.)

**Automated:** a combination of human and machine effort, often with one assisting or operating the other.

**Automatic:** a computerized process that involves machines only (no people) to produce a result.

**Capacity:** a measurement of the total number of documents that

can be processed in a given timeframe. Typically measured in days or weeks.

**Confidence Reporting:** An indication of the likelihood that a particular piece of data is accurate. This can take the form of a numerical value, a letter grade, or an mark for "suspicious" or questionable data.

**Configuration:** the process of setting up a computerized run of software according to specified parameters.

**Coding Manual:** a specification to be used by document coders to ensure conformance with customer requirements.

**Coverage:** Valora's term for the percent of a document population which is automatable.

**D/E:** Data entry. A manual process of inputting data about a document into a database.

**EDD:** Electronic Data Discovery. A loosely applied term covering a host of processes related to the processing of electronic documents, such as email. Generally accepted to mean the production of images, text and metadata for electronic files.

**Encoding:** Valora's fancy word for document coding, also known as indexing

**Global QC:** a process of quality checking that does not look at each record individually, but rather looks at the fielded data as a whole. Typically used for consistency checks.

**Image endorsement:** the process of adding a term or sequence of numbers to a set of images. Often used to Bates stamp a document population.

**Key words:** a field option for document encoding that allows for any word or phrase (from a list) to be collected if it appears anywhere inside a document.

**Load file:** a computer file that matches images and text together for the same document record.

**Normalization:** the process of consolidating multiple versions of a data entity (such as a name or organization) into a single, canonical form.

**Objective coding:** a form of document encoding that does not seek human judgement or a particular point of view to create the fielded data. Examples of objective coding include: document date, author and title.

**OCR:** Optical Character Recognition. A computer process that creates a text file from an image file. Text files can be used for searching and other automated processing.

**Parsing:** a computerized process where software attempts to interpret the content and/or parameters of a document.

**QC:** Quality Control. The process of checking each data record for accuracy, completion and consistency.

**Sampling:** the process of selecting representative documents from a population for further processing and analysis. A statistically valid sample can be used to forecast performance metrics of the entire population.

**Subjective coding:** A form of document encoding requiring prior knowledge about a case and/or the litigation strategy. Examples of subjective coding include: issue coding and coding for privilege.

**Template:** Valora's word for a more granular representation of document type.

**Throughput:** the rate at which documents enter and/or leave a

document encoding process. Typically used in high volume coding situations.

**Turnaround Time:** the elapsed time a project takes to complete.

**Unitization:** the process of separating pages into their logical document boundaries. Sometimes used to mean the process of turning physical document breaks, created at scan time, into logical document breaks.

**WYSIWYG:** What You See Is What You Get. Used in a coding context to mean capturing data verbatim as it appears in each document. The opposite of normalization.

---

## Showcase Your Legal Department

*By Alison L. Weinberg*

So what's the problem? Your legal department can be classified as "world class". You've done all the right things: re-vamped your organization's structure with clearly defined roles and responsibilities, developed a portfolio appropriately distributing workloads to in-house and outside counsel, converged your outside counsel law firms, and identified your preferred network of vendors. You've become the ideal law department, a well-oiled machine from inside out. So why haven't things changed?

Legal Departments are often thought of as cost centers whose primary goal is to manage corporate litigation, but they are so much more. It's time corporations recognize a law department's potential in order to maximize the value the department can bring to the overall corporate strategy. By bringing the legal department in at the onset of corporate transactions, legal can mitigate risk and prevent litigation. The bottom line, help the company save and make money. How can this happen?

External strengthening of the legal department within the corporation will provide the empowerment to become an active member of the organization versus acting as a corporate appendage responsible for damage control. To achieve this, the legal department must be willing to invest in some shameless promotion to market them as an integral part of the organization.

Marketing is defined as the process or technique of promoting, selling, and distributing a product or service<sup>1</sup>. Legal Departments have services to provide and as with any professional service organization, marketing is key to gaining clientele. Although corporate clients automatically exist as a result of active legal problems and litigation, it is the clients who seek advice that legal departments need to target. Legal departments should strive to expand their use within an organization and become a trusted advisor versus a "fire-fighter" and marketing can help them transition to the next level. So what needs to be done?

Through promoting it's strategy, skills, and institutional knowledge, a legal department can become a stronger force within an organization. With all organizations, successful advertising and product sales start with a marketing strategy.

- **Recognize Your Value.** What services can you provide that can help make the corporation reach it's goals? What key resources (i.e., people, processes, technology, etc.) do you have that the company can utilize to accomplish their mission?
- The value a legal department brings to a corporation goes beyond litigation and departmental cost management. Legal departments have the knowledge and ability to assist in making

good business decisions that can help a company save money and generate revenue. By providing legal advice from the beginning to the end of a transaction, legal can identify risks, prevent loss, and even uncover more transactional benefits and advantages.

- Acknowledge the value the department can bring and list the value-added services that are key to the company's success. Next, identify the resources and knowledge that makes those services work. Assess the skills base of those in the department, what are their marketable strengths, who is the person most qualified in specific areas? After you have named these key personnel, designate them as the point person to act as a liaison to the organization. Once you have outlined the value your department can provide to your organization, you are ready to market to your target audience.

- Identify Your Target Market. Who are your clients? What are their needs? What services can you provide to meet their needs?

- Your ability to become a pro-active department rests in your clients' hands. Knowing your clients is key to becoming more involved in corporate business. By understanding who your clients are and performing services that meet or exceed their expectations, you will gain their confidence in your abilities to service their needs. Accordingly, your clients will begin to seek your advice and depend on your services to make sound business decisions.

- Review your organization's structure and identify the personnel that would most likely require your services. Become familiar with their roles and responsibilities and identify the services that they can benefit from. After you establish the needs and wants of your clients and your role in meeting those needs, you can begin to author your marketing communications.

- Putting It Into Words. How can you communicate your marketing strategy to the masses? What format will convey a more marketable unit?

- Because you have established who your indispensable resources are, a simple way to advertise these people is to illustrate a departmental structure. Publishing an organizational chart is a practical way to spotlight the who's who in the legal department.

- Publish a departmental mission statement. Writing a mission statement is effective in delivering a message of who you are and what your intentions will be. Draw upon the consideration that was given to uncover the department's value and target clientele. Consider the standards of excellence in service that you will provide to the organization. The utilization of this information will be the center of your marketing communication. Now that you have completed the creative process, it is time to initiate active visibility and accessibility.

- Increase Your Visibility and Accessibility. How can you become more visible to corporate personnel? What will make you more accessible to you clients?

- Visibility is important to clients. It is easier for clients to identify with visible and accessible entities that just a name. Become actively involved in the organization is essential to gaining visibility. People trust advisors that are able to speak to them on their level. Knowledge is essential.

- Learn more about the business. Become involved in corporate functions and associations. Attend board meetings, request face-time with the CEO and be visible to Stockholders. An active General Counsel is an involved General Counsel and the depart-

ment will reap the benefits as a whole.

- Once armed with more information and current business activity, the legal department will be more proactive versus reactive. The department's function will evolve from "fire fighter" taskmaster to a risk management role. You possess the ability to troubleshoot and forecast threshold issues plus have the available resources on hand to address them.

- Sustain a Trusting Relationship. How can you maintain your client's trust and confidence? What can you do to make sure they continue to come back for more?

- Keep the relationships going. Make the clients happy. Once you gain their confidence don't lose it. Maintain the relationship by keeping open lines of communication. Be sure to solicit feedback on your department's performance to ensure continued quality of service. Distribute quarterly newsletters as means of continuing the marketing momentum and longevity of superior service.

One of the most valuable assets to a corporation is a department that has a vested interest in helping the company succeed.

Becoming an active member

- The Legal Department's initial sales pitch will open the door to corporate recognition and opportunity. The ability to maintain a trusting relationship will get the corporation to move forward and embrace the legal department as an integral pro-active force.

---

*Alison L. Weinberg is a Masters of Science candidate in Information Technology Law at The John Marshall Law School. Alison can be reached at [5weinba@stu.jmls.edu](mailto:5weinba@stu.jmls.edu).*