

aiim® View on:

Content Capture And EMC Captiva Document Capture

Don't Let Sleeping Dogs Lie: A New Look at Content Capture

Among the thousands of acronyms within the IT industry, one of the oldest is GIGO (garbage in, garbage out). The maturity of this acronym indicates its fundamental resonance; it is as true today as it was 50 years ago. Despite all the advances in Enterprise Content Management, if input into the system is flawed, the entire system is potentially flawed, and those errors will not only endure throughout the life cycle of the content, but create many more related problems.

The capture phase (authoring, scanning, etc.) is a basic building block of an Enterprise Content Management (ECM) system. As illustrated in the popularized AIIM diagram depicting the definition of ECM (see Illustration 1), capture is one of five basic components of ECM (the others are store, manage, deliver, and preserve).

Both of these perspectives provide a false sense of security. The capture process maybe “an old dog” in the ECM world, but this is a case where we should not let sleeping dogs lie. Process efficiencies and enhancements can be realized if capture is reexamined and repositioned in many enterprise settings.

Small improvements in technology and processes can bring dramatic increases in process acceleration, minimizing many potential bottlenecks. Moreover, capture can, and often should, be positioned at various points of a process. It would behoove current and potential users of capture technology to scrutinize each instance of capture, and determine the most opportunistic point of capture and approach to the process.

Consider a contract that is being collaboratively developed; information capture occurs multiple times over along period. Consider an insurance claim; capture of the claim form may kick off the process, but throughout the process capture will be used over and over as additional input is required and received. At each step, the process designer should make an informed and strategic decision as to when, where, and how best to capture that input.

That is the focus of this paper: We examine capture alternatives and options to deploying the functionality in a range of choices, from standalone point solution to integrated component of a distributed ECM platform. A new perspective is provided, one in which capture goes beyond that of standalone application to a service or platform available throughout an organization, in a just in time fashion.

Repositioning Capture as a Platform

Information capture may not be an ascent technology, but it still plays a pivotal role in content enabled solutions. Even for organizations that have established capture within their arsenal of ECM tools, there is potential benefit in rethinking the capture strategy. Over the last few years, capture technology has made advances in many capacities, including hardware based speeds and feeds and software based character recognition. Developments in three specific areas, however, are behind the emergence of capture as an integrated value adding enterprise service. These are:

Process integration.

Distributed processing.

Intelligent document recognition.



But although views such as this elevate capture to a high degree of importance, they potentially do capture a disservice as well. The popular opinion, as shown in Illustration 1, is that capture typically occurs at the beginning of the ECM process. This perspective leads many companies to implement capture systems as centralized (sometimes off site), standalone component applications, often running capture in batch mode. Furthermore, the maturity of capture technology has given many professionals a perspective that scrutiny and reexamination of the capture process need go no further than monitoring output quality.

Collectively, these enhancements reposition capture as a service that can be invoked from virtually any location and application, that can be seamlessly integrated into broader applications, and that is capable of expediting content management through automated rules based processing. Capture need no longer be viewed as an application specific silo, but can instead be considered a callable, flexible service, leveraged as an asset across organizational boundaries. This results in streamlining of processes and a reduction in processing costs as standalone approaches are replaced by integrated, distributed, and intelligent processes.

Process Integration

Whereas distributed processing untethers capture, enabling remote, close to the application processing, process integration enables capture to be implemented as an integral part of workflow, business process management, records management, or ERP systems. Capture is not a subprocess, but an integral step that moves the content directly and immediately to its appropriate stage. Capture is thus a seamless step, invoked as needed and automatically returning content in context, in support of a larger process or management system. Process integration positions capture as a highly leveraged enterprise competency, not simply a point solution.

Ultimately, leveraging a service oriented architecture (SOA), capture is positioned as a service that can be invoked at any point as just that, a value adding service. This service becomes callable from sales force automation programs, document management systems, BPM systems, and ERP systems and processes within them, and symbiotically can pass content and triggers to these applications and processes. Taken on a more granular level, capture can be integrated with knowledge management alerts, initiating the issuance of an alert when new content is added to an application or process, a folder, or a record class. Capture thus becomes part of an integrated strategic platform for content management and collaboration, rather than a standalone process of scanning.

Distributed Processing

This enhancement to the capture process directly challenges the notion that capture occurs in a centralized location, typically at the beginning or end of a process. Support for distributed processing is at the very foundation of providing capture as a callable enterprise service that is capable of being used at any time, from any location, by any application. Distributed processing provides a single user interface to any number of physical capture devices or routines. In this approach, the capture software is decoupled from the physical hardware or the program based means of capturing content. This renders a single user interface to “capture,” whether it is being invoked on a remotely located high throughput batch machine or a local desktop single sheet fed scanner. The flexibility of distributed capture systems lies in their ability to support all types of capture devices throughout an organization (including remote standalone offices) from the same interface, integrated across organizational applications.

Distributed processing not only simplifies the interface to capture, but allows centrally administered and defined capture functionality to be deployed locally. This has many benefits. Localized capture eliminates the cost associated with physically shipping documents to the location of the centralized scanner. This in turn increases process cycle times by eliminating the time associated with document transport, and eliminates the potential of loss or damage to the documents in transport. Perhaps more powerfully, distributed capture allows the capture and indexing process to occur at the source within the business process and to be controlled by the individual most familiar with the content being captured. The quality of the index is likely to be higher when those familiar with the content are performing the task. The time burden associated with indexing is also minimized when it is spread across the process.

To appreciate the extent to which such distributed processing can occur, consider the ultimate distributed setting. Imagine working offline, out in the field. Working with a customer, you obtain

content via a series of eforms and paper documentation. Using a laptop and portable scanner, you capture and index the content with the customer present. Upon returning to the office, you connect the laptop to the system, and all of the captured content is automatically synced to the appropriate applications and repositories. Ultimately, the syncing becomes even more powerful when it is linked into appropriate processes, via the integration functionality discussed below.

Intelligent Document Recognition

The most complex step in information capture may well be the identification and tagging of the captured content. One could argue that without this critical step, the powers of integration become severely limited. Integration, as discussed above, typically relies on some level of insight regarding the content—for example, “What type of content is this?,” “To which customer does this content belong?,” or “What do I need to do with this type of document?” Identification and content extraction can represent the first significant bottleneck in the capture process, as well as the first bottleneck in the bigger business process to which it is linked. Once content is physically captured, the necessary steps of indexing, classifying, and extracting it to be stored in other applications must be performed. Until recently, this work had to be done by people—the most expensive form of process. Over the years various forms of document recognition (barcodes and separator sheets, for example) have been used with varying degrees of success to streamline this stage of the capture process and minimize human involvement.

Recent advances in recognition may therefore be thought of as evolutionary rather than revolutionary. An advance among these enhancements that supports the positioning of capture as a “universal service” (as opposed to a siloed or specialized function) is innate flexibility in some capture systems. This flexibility allows users to dynamically determine content identification and extraction, without having to specifically hard code or fine tune the process logistics for each content type or business application. Heuristically and dynamically, intelligent recognition recognizes document types and potentially their relevance to applications and processes.

Content can come in as an XML stream, on paper, in email, in attachments to email, on a fax, or via an eform, and it will be processed in the same manner, with little to no delay as the incoming format or type is automatically recognized and appropriately processed.

Such recognition can trigger further intelligent processing, increasing the value and speed of the recognition process. For example, once identified as an invoice, the captured form can be further intelligently processed to extract billing name and address and amount due.

Identification and extraction can be the single greatest bottleneck cost to capture and possibly the single greatest cost of ECM. Intelligent recognition immediately speeds up the deployment of the capture functionality, and over the long term, it greatly reduces reliance on human labor, the most expensive approach.

It should be appreciated, however, that no approach to intelligent identification and extraction is foolproof. That is why the better capture applications allow users to set confidence thresholds on automated identification and extraction routines. Content that is processed automatically that achieves a value above the confidence threshold is automatically passed to the next stage of the process. Captured content that falls below the confidence threshold is automatically routed to appropriate personnel for review. Such an approach allows an organization to maximize the value of its experts, who can then focus their time on the “difficult” content. Capture solutions that provide such control deploy proprietary algorithms for determining confidence thresholds. These should be scrutinized and rationalized by the potential user of such a system. Striking a balance between deciding not to pass certain information to the human reviewer (to minimize reliance on valuable human resources) and ensuring that what is identified and extracted is correct (to avoid potential serious ramifications) is a crucial part of a capture strategy.

Conclusion

Although most organizations have invested in some form of automated content capture, many fail to realize that there is an economy of scale (cost and performance) that can be realized by establishing capture as a centrally administered, locally distributed service and a corporate competency, integrated into corporate processes. Capture should be positioned as an enterprise capability, readily invoked by any process in a dynamically tailored manner. In this way, the organization can standardize best practices for the capture and declaration of content across the organization.

Capture should be invoked at the most logical point of the business process. It no longer should be viewed as occurring only at the front or back end of the process. Capture software and processes should be decoupled from hardware and distributed throughout the workforce. This eliminates unnecessary specialized or redundant interfaces and integration.

Organizations considering adopting such an information capture model need to view capture not just as a siloed function, but as a subprocess or subsystem easily integrated into other processes and systems. This subprocess should be made up of several subprocesses itself, as shown in Illustration 2.

If the architecture is sound, this subprocess acts as a hub that accepts any form of content (paper, eforms, PDF files, etc.) and processes it in context.

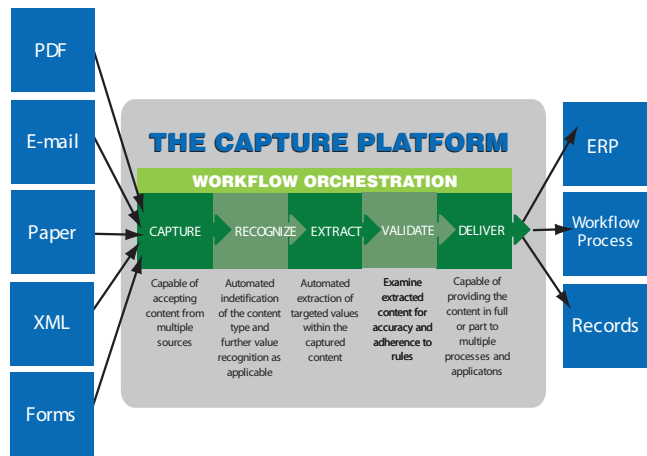


ILLUSTRATION 2

The system should be modular, providing specialized processing capability and flexibility at each stage of the process (capture, classify, extract, validate, and deliver). It should be open at the back end, providing a standardized approach to integration with all business applications. Finally, the system should provide process control in order to automate and orchestrate the processing of the content.

EMC Captiva Document Capture

EMC is one of the information capture solution providers on the market today. The Captiva products are reviewed here as they relate to the issues raised in this AIIM's View On.

The EMC Captiva capture solution is architected as a platform (see Illustration 3), that supports both centralized high volume batch capture and ad hoc distributed capture.

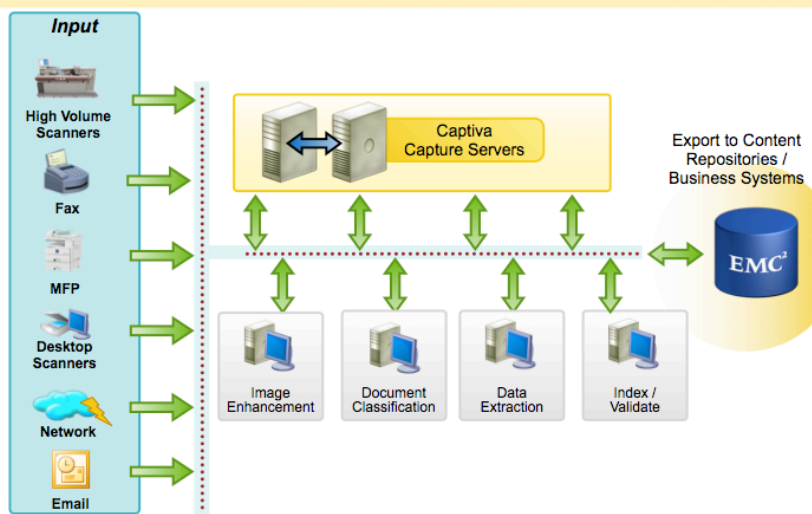


ILLUSTRATION 3. The Captiva Product Architecture

This architecture separates the solution's functionality into three basic capabilities: scanning/operator interfaces, capture processing, and system management. The modular approach to product architecture permeates the product, as discussed in more detail below, which underscores a general flexible and open approach to capture. Capture is positioned as a platform play, not a silo implementation.

Captiva Capture Server

The capture server is the core of the Captiva capture solution. The capture server provides an open platform that manages and controls the document capture process. It handles the processing of content and integration into other business platforms, applications, and systems. The capture server provides process definition functionality, known as capture flow processes. The capture flow processes control how documents are handled, including scanning, image enhancement, document identification, recognition, validation and export. Scanning and validation require some user involvement, but all other tasks are performed by Captiva modules which function as clients to the capture server, which governs the enterprise capture environment, including routing documents and processing instructions to the appropriate client modules and balancing the workload.

Within a centralized and/or distributed environment, the capture server can prioritize processing by levels of urgency, which can be defined by the system using a series of rules. The capture server provides integration with existing document/content management systems, ERP systems, and process models. Captured content can be stored in an appropriate repository and automatically forwarded to the appropriate departments or individuals.

Although it is part of the EMC family of products (which includes EMC Documentum), the Captiva solution has been integrated with IBM/FileNet, Open Text, IBM, Microsoft, and Hyland. The Captiva solution also provides a set of development tools that allow the customer to finetune the integration process to other applications.

Distributed Web Client

The Captiva capture platform supports several ways to enable remote capture. One way is the distributed capture process begins with a Web-based thin client.

Control of the capture process occurs directly from within the browser, where a remote user captures images using either a desktop scanner or browsing a network folder for images that were captured via a multifunction device, network scanner, or other source.

Access to capture is provided through a single Web client interface that can be invoked from any remote site and the client supports more than 300 scanners through the ISIS standard. Scanning can occur both online and offline, with automated syncing to the server upon connection.

The Web client interfaces with an application server, which acts as the middle tier and serves all requests, including handling business logic, and communicating with the Captiva capture server. Remote workers can handle the scanning and indexing of documents in an ad hoc fashion, or take advantage of the capture capabilities of the client and server. For example, a batch of documents can be scanned within the Web client and classified using document identification techniques such as barcodes, or the batch of documents can be submitted to the capture server and classified using Captiva intelligent document recognition routines. In addition to the Web client, distributed organizations can scan directly from a multi-function peripheral (MFP) device to the Captiva capture server utilizing the integration that is available between the Captiva capture platform and eCopy ShareScan.

Document Recognition

Automated identification of documents is supported based on full page image analysis, anchors (e.g. a logo or block of text), keyword analysis, handwriting detection (e.g., handwritten sections are skipped, or tagged and routed for human processing), and OCR-based text matching analysis—ranging from a single word to an entire text phrase. Validation processing of extracted data within Captiva includes quality control measures. For example, if during a capture process the application recognizes captured content as an invoice, it will tally the line items and make sure that they equal the “total” field. Any identifiable discounts will be applied atop the equation.

Confidence thresholds and business rules are supported, which enables documents that are not identified and/or data that is not validated be routed to an operator for review.

Usability and Secure Processing

It is worth noting that the Captiva capture solution provides functionality beyond the scope of this AIIM’s View On. Within the Captiva Web client, captured content is treated as a working form of enterprise knowledge. That is to say, the captured content can be marked up by business users, to increase the usability and value of the content as a business resource. Content can be highlighted and redacted. Annotations can be applied via pop-up text boxes, stamps, and arrows or lines. Each of these forms of annotation does not alter the original content, but rather is overlaid onto it. More importantly, the annotations (including redaction) can be controlled for security.

Finally, the Captiva capture solution provides a secure platform for the capture process and the captured content. Tight audit trails are maintained throughout the capture process. Both client-side and server-side log files are automatically maintained. Multiple versions of an image are maintained across its life cycle. In this manner, the original file is maintained for compliance reasons; annotation and cleanup can occur for readability and processing reasons. Content transmissions via the Web client are secured through SSL, encryption can be used, and there is support for corporate VPNs (virtual private networks).

Conclusion—Transactional Document Processing

The Captiva capture solution allows users to create both a distributed and centralized capture platform within the organization. But it is worth noting that these EMC Captiva products are also part of the bigger EMC family of Enterprise Content Management products. On the one hand, EMC views capture as a platform that can be integrated into virtually any other application. On the other hand, EMC views Enterprise Content Management (ECM) as a modular platform that can address many business challenges and of which capture is just one component.

The company labels ECM deployments targeted at addressing business processes “Transactional Content Management” (TCM). TCM is further explored in Illustration 4.

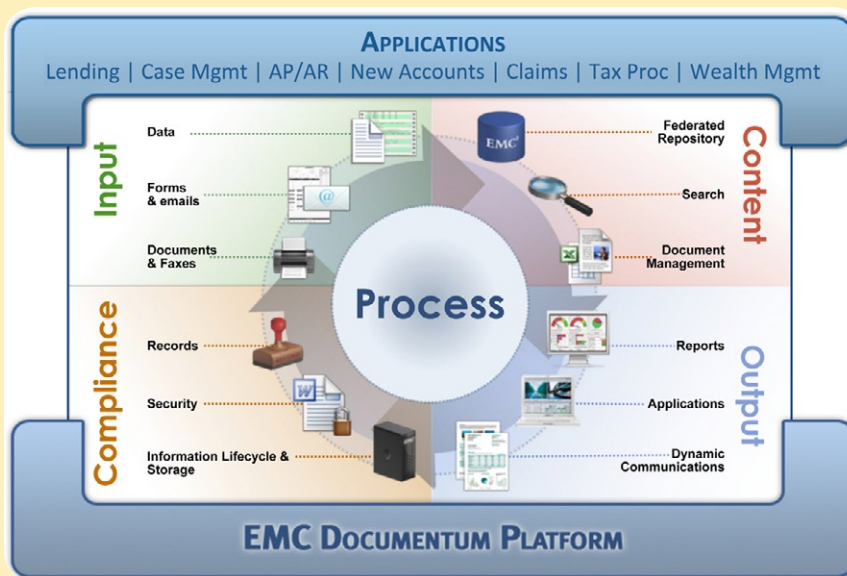


ILLUSTRATION 4. The EMC Transactional Content

TCM is an EMC “solutions framework” and category of applications that enable near-turnkey deployment of business processes that involve transactional content, requiring the unification of process and content. Examples include claims processing, loan origination, case management and invoice processing.

TCM is potentially a valuable integrated platform for addressing these business problems, but the modular design of the Captiva capture products allows them to be deployed individually and integrate with existing systems or processes. They do not have to be tied to other EMC products.

Due to the fundamental importance of document capture and the breadth of its capabilities, EMC Captiva’s products should be positioned as an organizational capability, not a silo solution or specialized function.