69 Lura Lane
Waltham, MA 02451 508-816-3650 aliusdoc.com

White Paper

Automated Data Capture 202 – Beyond the Basics

By: Paul Traite, CTO/Co-founder of AliusDoc LLC

Automated Data Capture 202 – Beyond the Basics

Automated Data Capture has been around for over 50 years, and has progressed from the first magnetic ink character recognition (MICR) check readers, to sophisticated systems solving complex data capture tens of times faster than unassisted manual processing. This micro-course "202" goes beyond the well-known basics of a "101" course, to describe the abilities of these modern solutions.

Contents:

About This Course Offering

This micro-course was developed by Paul Traite, CTO / Co-founder of AliusDoc LLC.

Paul has 25 years of experience in providing data capture solutions to organizations, starting with one of the first large-scale capture solutions of US medical claims for Blue Cross/Blue Shield, and helped create solutions since then for the US Patent Office (USPTO), Kansas Department of Revenue, and the Columbian 2010 census.

AliusDoc (www.AliusDoc.com) provides both pre-packaged solutions, and its development platform AD-SCI™, for automating complex document identification, indexing, and full field data capture.

So... What is a Document Anyway?

Just about anything which needs to be transactionally processed in an organization can be considered a document. Traditional examples include things like:

- Tax forms
- Invoices
- Medical records
- Change-of-address letters
- Mortgage paperwork
- Account applications

Note: Transactional processing can include just document identification, indexing and archiving; or fuller data capture and additional business processing.

In addition, anything that can be searched for useful business information can also be considered a document. For example:

- Spreadsheets (such as Excel™ files)
- Engineering drawings
- Web Sites (such as XML or JavaScript)
- Emails (and their attachments)

Complex Inputs that Require Modern Solutions

Document Source -not just scanned images
Back in the old days, filled in paper documents were mailed to a central location, which scanned them into images to be processed. Modern solutions need to handle documents coming in as paper, as well as in various electronic formats such as pre-scanned images in many formats, PDF (embedded images w/ w/o text), MS-Word documents, spreadsheets, and emails.

Document Paper Source -no such thing as a "fixed" layout
Even when you design your own forms for paper documents, the paper document you receive back will vary because it can come from your print shop, or user's printing a

PDF version, and then in either case possibly photo-copied and/or faxed back. The resulting mix of changed margins, skewed documents, scaling, and color/shading transformations can still pose a challenge for some modern automated data recognition solutions.

Semi-structured Documents -infinite variety of layouts, single common set of data fields

Probably the best way to describe semi-structured documents is to compare them to structured documents. A US Federal tax form is one of the oldest forms of structured document. However, a "W2" wage reporting document is semi-structured. The set of fields it has are mandated by the IRS, but locations of the fields are not. This results in a (nearly) infinite variety of layouts of paper W2 documents. Good modern solutions should be able to easily find the data fields from any layout, without having prior knowledge of each layout.

Unstructured Documents -I'm moving to a new address

Unlike semi-structured, identification of the specific type of unstructured document is the biggest challenge. The amount of actual field data is often small. While the data can often be automatically extracted, the manual keying is also usually trivial; so little savings may be achieved. The codification of document ID provides real labor savings, often from staff that are more highly skilled than basic data entry staff.

What You Should Expect a Modern Solution to do During Production

Document Prep -Patch Codes and Separator sheets are passé

As automated indexing / field data capture eliminate more and more keying, the document preparation (prior to capture) has become a bigger share of the remaining expense. Well-functioning modern solutions eliminate the need for inserting patch codes or separator sheets before/after each document prior to scanning, saving significant time and expense.

Minimal manual intervention

No automated solution is perfect. They all will at least occasionally:

- Fail to identify a document
- Incorrectly identify a document
- Fail to find any data for an index or field
- Find the incorrect data for an index or field

In all of these cases, modern solutions should require only just a bit of manual intervention. After manual intervention, well-designed modern solutions will automatically continue the automated process.

Flexible Keying

- Modern solutions use a combination of automatic and user-selectable options to optimize the keying experience to accommodate:
- Differing physical dimensions and shapes of computer screens
- Mimicking the layout of fixed structured documents
- Setting the keying order to take advantage of "auto-skipping" subsequent fields.
- Fields sparsely located across multiple pages of a document
- Presentation of multiple potential values for a field, possibly across multiple pages.
- BPOs -each customer needing a different view (i.e. subsets of fields) for the same document (ex. Invoices, POs, medical claims).

Modern solutions are easy to configure so users and/or their supervisors can easily adapt the interface and the solution automatically applies the appropriate interface.

Setting Up for Production

All modern solution platforms provide a set of tools for setting up specific documents for processing.

What separates the good solutions from the rest is:

- Can you do the setup without paying the solution provider for extensive professional services, and
- How much of the setup your document content experts can do, with minimal support from your IT staff.

A well implemented and documented solution platform should require just a few days to a few weeks of vendor training for you to perform the following setup tasks:

- Software Installation
- Workflow setup
- Document ID, indexing, and data extraction setup
- Integrating with appropriate interfaces such as look up tables and export modules
- Setting up keyer interfaces
- Applying additional business logic to the capture process to maximize automation and provide greater accuracy and efficiency.

Modern Solutions allow all of this at a fraction of the effort and cost of older solutions and decrease the client's dependence on the solution provider.

Following are more details on some of the setup tasks outlined above.

Easy integration of business logic into the solution

Well-implemented modern solution platforms allow document experts to use "point-and-click" common business rule building blocks to describe valid and invalid values of both individual fields, as well as combinations of fields.

In addition, the solution platforms all should also supply IT specialists with the ability to extend the business logic with existing internal company modules, or custom logic, using industry-standard languages and tools.

When people refer to an "intelligent capture solution," it refers to the amount of logic the software applies to conduct complex capture. The easier it is to educate your solution, the less expensive it is to improve its intelligence, saving time and resources of IT professionals and users alike.

Easy integration of the solution into your larger organization

Older solution platforms often had complex proprietary formats for passing data into their solution, and getting results out of them and into your organization for transactional processing, data mining, and archiving.

Correctly implemented modern solutions have minimal dependence on costly 3rd party packages requiring even more IT expertise. Instead, good solutions use simple, non-proprietary, and easily readable interfaces such as XML to pass data into and out of their components.

Configurable components -Using only what you need

A well-planned modern solution platform provides the flexibility to use only the components needed, with clear boundaries and easily managed interfaces to each, so you can integrate your existing system components to any component.

In one common scenario, documents may be identified prior to entering the solution (for example, by using a post-office box #, or via earlier Content Management processes). Good platforms would make it easy to pass in the "pre-IDed" information and by-pass the solution's ID functionality.

In another scenario, you may have an in-house keying platform in place. Again, a modern solution would make it easy to import the results of the automatic ID, indexing and field data capture directly into your existing keying system.

More Information

AliusDoc LLC provides AD-SCI™ solution platform for solving complex document sorting (ID or digital mailroom), indexing, and full data capture needs.

For more information, visit www.AliusDoc.com, or for an immediate response, you can call Fatali Karimi, directly at (508) 816-3650, or email us at InfoRequests@AliusDoc.com.