CAPTURING SEMI-STRUCTURED FORMS AND DOCUMENTS: CHALLENGES AND AVAILABLE TECHNOLOGIES

Over 80% of all office forms and documents cannot be captured by traditional software designed to recognize fixed-structure forms. They simply lack a fixed structure that the software could recognize. Is there a technology on the market that can reliably capture such forms? How advanced is it and how does it work? This paper describes the semi-structured forms capture technology that is available to businesses.

The Problem with Semi-Structured Forms

To understand why traditional solutions cannot deal with semi-structured forms, one has to understand what differentiates them from fixed, or structured, forms.

With structured forms, each form has exactly the same layout with fields positioned at exactly the same places on the page. The quantity of these fields per page is also fixed.

To capture information from such structured forms, the templates are used: before starting the capture in production mode, the form is opened in design mode, and the areas to capture are mapped with the mouse. The location of the fields is remembered, and the form is ready for production capture based on the coordinates of the fields to be captured.

The traditional forms processing technologies for structured forms are well established. A wide variety of systems capable of processing many types of fixed forms is now available. Today's advanced systems can accurately capture printed and handwritten characters and process thousands of documents per day.

Semi-structured forms are documents in which the location of the data and of the fields holding the data varies from document to document. For example, the shipping address on the purchase order can be in the top left or top middle or bottom right area of the form, but it still has to be captured as the shipping address. So field coordinates cannot be used for capture any more. The quantity of fields, or quantity of lines per table, or quantity of items per transaction can also differ from page to page and from vendor to vendor. Some fields – or even full columns in the tables – can be optional, present on some documents but absent on other documents of the same type. Even the quantity of pages can be different: while one vendor sends the invoice on a single page, another submits hundred-page-long invoices. Another problem with semi-structured forms is that in most cases, only a few key pieces of information, or certain fields, are truly important. So the challenge is not only to find the important information, but also to understand which information is not important and should therefore be ignored.

Traditional forms processing programs are not flexible and intelligent enough to process documents that lack a strict structure without extensive customization and system training. Until recently, easy-to-deploy, cost-effective solutions for processing such documents as invoices, payment order forms, legacy forms and template-based contracts were virtually inaccessible to a large audience.

Examples of Fixed and Semi-structured Forms

Credit application forms or account opening applications are typical fixed forms. The bank or financial institution usually designs and prints such form in-house, then customers complete the form, often by handprint, and return the form to the bank for processing. Since all the forms are printed together from the same empty sample form, the fields are perfectly at the same coordinates and can therefore be captured very well with regular fixed forms capture products.

Remittance Advice and Explanation of Benefits forms are examples of semi-structured forms. Although we know generally what kind of information these forms should contain, we cannot tell the exact location of the information since the forms are printed by customers, using different software and different printing media and hardware. Plus, we do not know beforehand how many transactions will be performed and how many details a particular customer will include into each transaction record.

When we go to semi-structured capture, there are not only forms but also a great number of semi-structured documents to capture. For example, supply contracts or mortgage agreements. A mortgage note can include several pages and can be supplemented with signed affidavits and the Deed of Trust. The capture process must be able to locate the key information across multiple pages and populate the database or indexes in the document management system with the right values.

Challenges Facing Forms Capture Technologies and Traditional Techniques That Are Used to Overcome Them

There are some problems that are common to both types of forms, and a capture technology, whether fixed or semi-structured, must be able to deal with them.

1. The pages of a form are often skewed or shifted during scanning, which makes recognition of the page very difficult.

2. Sometimes the forms are printed on different hardware, and scaled up or down to fit the printing area of a particular printer, and so they can be stretched or shrunk compared to the original form.

3. Forms are often submitted by fax. Faxed forms often have considerable non-linear distortions. Everyone has seen faxed pages that start going nicely, then suddenly start stretching on a section of the document, and then go back to normal receiving again, so that only some sections are distorted, and not the entire document.

4. Quite often a page comes upside down, or a portrait-oriented form is scanned in landscape to increase the scanning speed.

How are these challenges met in the modern fixed forms capture software?

First of all, it is done with the use of static blocks, such as static text, black lines or corner marks, that function as anchors against the shifts and skews.

Having just four such elements located close to the corners of the page is enough to compensate for a scaled-up or shrunk-down document. But if you receive a form by fax, you should be prepared for non-linear distortions, and so the software should know the direction in which the faxes are received (portrait or landscape), since non-linear distortions always happen in the direction in which the fax comes. Then you add some more static anchors across the page, so that non-linear distortion algorithms have enough information to compensate for faxing defects.

When a page comes upside down, auto-rotation of the pages is applied. To effectively auto-rotate the pages, a fast recognition run is done on the page. If there is no meaningful text read with the page oriented as is, but there is meaningful text if the page is rotated 90 degrees, then the 90 degrees rotation is applied.

Many fixed forms solutions allow for such shrinking, stretching and relative scale change, which is the first step towards flexibility. If the challenges presented by your forms do not go beyond that, then a good modern fixed forms solution will be able to process your forms. Price is also an important factor here, as fixed forms solutions often cost a half or a third of the price of a semi-structured solution.

Advantages of Semi-Structured Capture Technologies

So, what would be the reasons to go beyond fixed forms, if the fixed forms technology is so advanced?

First of all, it is the high business demand that drives that step. According to industry studies, over 80% of all office forms and documents are semi-structured rather than fixed.

In many cases, the customers are looking for a single data entry point: the paradigm that many companies are following is that there should be a single capture solution for all office documents and forms.

Another reason is efficiency. One flexible layout replaces a multitude of fixed form templates that are necessary for each document variation. In fixed forms technology, if each customer sends you his own variation of the form, you have to make a template for each such variation. With a semi-structured solution, there are no more templates. A single flexible layout sets the capture rules for the entire class of documents.

The next reason is flexibility. If you need to capture fields located at various places on the page, to extract details out of a table with a varying number of lines, to read data from transaction printouts with a different number of items, or to capture data that can be located on page 3 or page 7 of a semi-structured document – then you have to use a semi-structured capture technology, since all these challenges are beyond the powers of fixed forms technology.

And last but not least: often it is much easier to tell WHAT we have to capture than to describe exactly WHERE it is – and semi-structured technology allows for that data-centric approach to capture.

Challenges Presented by Semi-structured Forms and How They Are Resolved in the Modern Capture Systems

Challenges facing a semi-structured technology can be summarized as follows:

1. Unlike a fixed form, field location is different on different variations of the same type of form, and so coordinate-based capture can't be used.

2. Some fields are always present on the forms while others are optional.

3. The quantity of lines and columns per table can vary.

4. The quantity and order of lines per transaction can vary.

5. The data can flow across the pages of a multi-page form or document.

One of the solutions to these problems that are used today is sample-based image matching training. Before running in production mode, a software is trained on sample pages, and every time a brand-new image is coming, the software is trained how to capture the fields from that kind of form. This training updates the definition of the layout, and in the future such pages then can be processed automatically. This approach works well for tasks with only a few variations, since image matching for hundreds of images can take a long time.

There is a variation of the image matching principle done on a local scale, for a certain part of the page image. For example, image matching can be used to match a company’s logo with the logos known to the capture system in order to understand from which vendor the invoice or purchase order was received.

Another approach starts with full-page text recognition. Then data fields are located based on anchor elements pointing to the data with some sample relations like "Right of”, "Lower then", "Closest to", and other. For example, if we found the words “Total amount due”, then the number immediately following them should be captured as the “Invoice total” field in the database.

Another approach is locating the fields by their own formatting and content. That works well for dates, e-mail and web addresses, phone numbers and other elements with precise internal structure that is known beforehand. The formatting of that field lets us know what kind of field is it.

Sequential search approach means that fields found on the previous step are used to locate the data on the next step.

Yet another approach is a specialized solution: many companies are creating highly specialized solutions trained to perform a narrow task of capturing selected documents but to perform it very well. Specialized solutions for accounts payables, resumes, or business cards are available. Such a solution cannot be used for generic capture, but it can do a great job in capturing the selected type of the document. So, if 95 percent of your forms processing needs are processing Accounts Payable documents, selecting an AP-oriented solution is the best decision. If most of your forms are resumes, then a specialized resume-capturing product is the right tool for you.

But what if you have a wide variety of semi-structured forms and documents to process, and you don't want to spend a fortune by paying separately per each specialized solution?

In this case, the recommended approach is to select a tool that allows you to use the same package for all the different semi-structured documents and forms that your company processes. An example is ABBYY’s FlexiCapture technology, which is a fully flexible semi-structured capture solution.

Such a solution is a tool that allows VARs, system integrators or IT department personnel to create, test and fine-tune ANY flexible layout with the help of a user-interface-based flexible layout design and testing tool. ABBYY calls this tool FlexiCapture Studio. With such a tool, flexible layouts can be created by people who are not programmers but who simply understand what information has to be captured from the documents and forms. This technology provides a solution for a wide variety of data capture problems by giving the forms capture system a much higher level of intelligence and flexibility. A flexible layout gives a logical definition of the layout of data on the semi-structured form, thus freeing you from the limitations of template-based form matching (such as reliance on the exact placement of fields on the page). The system can find fields anywhere on the page, using any information available: relation to other objects on the page, contents of the field, its size, lines drawn around it, etc.

The flexible layout is then interpreted by the semi-structured forms capture technology incorporated in a recognition product: it serves as a set of rules for the form matching process. The flexible layout enables the recognition system to easily find the necessary fields on the semi-structured form. Once located, the data in the fields can be captured using the OCR/ICR/OMR and barcode-reading functionality of the recognition product.

With such a solution, the document is understood in its entirety: as a combination of static and data elements and relations between them. Any particular component can be missing (an object not printed on the page, or a relation not true for a particular variation of a form), but as long as at least some objects can be found, and some relations are still at place, the location of the fields to capture will be figured out. This allows for additional flexibility in locating data on forms with many possible variations, since any subset of elements can lead to the location of the searched data.

First, the page is recognized in full-page mode, and all text, words, lines, tables and graphical elements are detected. Then they are classified based on the defined set of objects and relations among them. In the result, we know that this text is not just text but, for example, the patient’s date of birth or the total amount of the invoice. Since there can be many similar fragments on the pages, a set of hypotheses can be present, and voting mechanism will determine which combination makes the best sense overall – in other words, which combination will lead to finding the most fields successfully. That combination will be selected for the output data locations.

This technology allows capturing data across pages for multi-page forms and documents. And it is available as both a ready-to-use product and a pluggable technological component that can be used inside of other existing capture solutions that the customer already uses and is familiar with.

Those who wish to implement a technology like FlexiCapture Studio should understand that it is a tool used to develop solutions. It is not a solution by itself, nor is it a technology for a specific task, such as invoice processing. A developer uses this technology to create solutions for capturing a wide range of document types. The result is the creation of a flexible layout. The end solution that is delivered to the customer is the flexible layout and a product that has a flexible layout interpreter incorporated in it.

How a Fully Flexible Semi-structured Capture Technology Works

In theory, capturing data from documents seems to be a relatively easy task. Even when one has to deal with semi-structured forms, at first glance, there is no major problem. The solution seems quite simple: finding anchor objects for the fields for which you are looking, detecting these objects based on their content, and then easily locating the adjacent fields.

Anchor objects

To find a particular piece of information on the form when there is no specific placement of such information, one may rely on field names, lines drawn, etc. For example, if there is a field with a person's address, then normally one should expect the word "Address" before (or, for example, below) the field. These guiding elements on the form, normally designed to simplify its perception by a human, are very useful when searching the form for fields. Such elements are called "anchor objects." The traditional approach is to use an OCR technology to read the contents of the page and then program a procedure that looks for specific fields by using anchor objects. One obvious drawback of such approach is that the wrong anchor objects may be identified, because the form is not analyzed in the whole. For example, it is quite possible that the word "address" is encountered several times on the page. Or there is a word "actress" on the form that is wrongly recognized as "address”. As a result, incorrect data will be captured. A much more powerful technique, utilized, for example, in ABBYY FlexiCapture, is based on the IPA principles, which allow the technology to analyze the form as a whole and consider all the possibilities of data placement to determine the best choice.

In practice, however, the situation can be much more complicated. For instance, if the anchor object is a text string, then it is quite possible that the text is not perfectly readable, and that the OCR may capture only a part of the text, or the text with some mistakes. Also, the same word, or even several words, found in the anchor text could also be written somewhere else on the form. It can also be that on some forms the field is on a single line, while on others it is on several lines, and it is not obvious how to distinguish lines belonging to the field from other text on the form. Sometimes there is no anchor text for a field, and one must rely on borders or lines drawn nearby. In this case, there is a high probability that the line is not uniform due to a not-so-perfect scan or careless handling of the paper.

Therefore, in the real world one has to expect vast variation from the ideal form model. An effective technology must be able to account for all variations and deliver reliable results. The great advantage of a technology like FlexiCapture is that its document model works despite such variations. Because it is based on intelligent technology and the IPA principles (see inset on IPA below), such a technology never relies on any fixed presumptions: you may specify any object or its properties to be tentative. Using the IPA principles, it generates a set of hypotheses based on rules provided by a human operator, and then picks up the best hypothesis for the whole set of objects on the page. This last point is very important: the technology makes decisions not by analyzing each object separately, but instead, by taking into account the relationships between all the objects and the characteristics of each object. Only then does it determine the best match for the entire set of objects. A system that analyzes individual objects one by one – without accounting for the whole and examining the relationships between objects – would fail. If a wrong decision is made for the first object, the system will fail to find all other objects that are related to that first object.

IPA Technology

Recognition technologies utilized by a solution like FlexiCapture are built on the principles of Integrity, Purposefulness and Adaptability (IPA). Unlike other recognition technologies, which focus on recognizing patterns, IPA takes recognition a step further by using artificial intelligence to train the computer to analyze documents in the same way that the human brain would analyze them.

Following the principle of ***Integrity,*** *the technology treats a document as a single object consisting of many "integrated" geometrical parts such as words, lines, pictures and other elements. Each one of these parts may similarly be analyzed as having its own integrated and interrelated parts. For example, a compound element may contain several basic elements.*

Following the principle of ***Purposefulness,*** *the technology, just like the human brain, purposefully generates hypotheses about objects on a document. It performs this function by interpreting the flexible layout, which is the essence of its designer's knowledge about the specific semi-structured forms. Built-in* ***Adaptability*** *allows the technology to more precisely generate hypotheses about specific objects based on the information collected from other parts of the image. With further technology improvements, its adaptability will go even further, making automatic adjustments to improve the flexible layout based on the analysis of the real documents being processed. In other words, the system learns and trains itself over time.*

IPA works for semi-structured forms processing so well because the approach is totally different from the way fixed template matching works. A typical fixed template matching algorithm relies on the fixed placement of static objects, such as lines, crosses or black boxes at the corners of the page, or large chunks of text. This information is matched against known templates, and the best choice is applied. With such an approach, there is no way to account for large variations in the forms' design, other then developing a separate new template for each and every possible variation. Such an approach is obviously expensive, difficult to maintain, and limited to the cases when it is still possible to use fixed templates.

How the Flexible Layout Works

The flexible layout is a set of elements organized in a tree.

An element is a set of distinguishing characteristics of a specific object on the page. When the flexible layout is matched against the particular page, elements will correspond to objects on that page. As the same object on different pages could be different, and even may not exist on some pages, the element should be generic enough to cover all the possible variations of the object that it represents.

A fully flexible semi-structured capture technology provides powerful tools to make such a generic description. You can specify ranges for an object's placement, its contents in terms of regular expressions, substrings or a set of variants, and relations to other objects on the page. You can provide for a possibility that the element will not be found on the page. You can also describe the single object using several elements of different types and then pick the element that was found during layout analysis. This is very useful when the same logical element on different forms has different contents. If the flexible layout is well made, there are always several ways to find the required information, and when one of them fails, another will work.

The element tree contains simple elements as leaves. These elements could be grouped into compound elements. Compound elements provide a level of abstraction: it is much easier to divide the whole form into logical groups of elements than to deal with a large set of separate elements. Any compound element can, in its turn, be incorporated into another compound element. There is no limitation on nesting (see the picture below for an example of the element tree).

When the flexible layout is matched against the page, the system tries different scenarios for setting correspondence between the elements and the objects on the page. Each possible correspondence is called a hypothesis. All the hypotheses generated during flexible layout matching are also organized in a tree. This is because several hypotheses may exist for each element, and as hypotheses for all elements are interdependent, different hypotheses for the related elements are generated. By analyzing the hypotheses tree, the developer can find out the reasons why something occurs during flexible layout matching, identify problems, and modify elements to improve the matching.

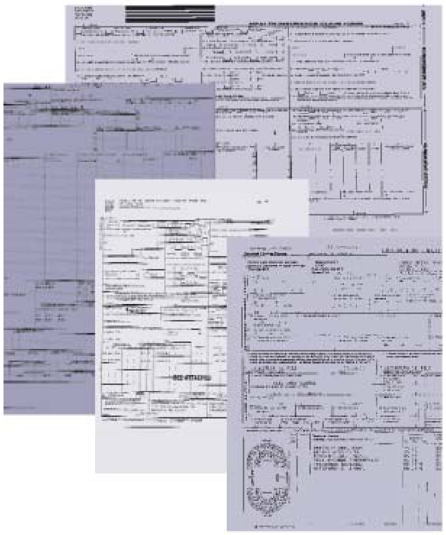
Examples of How a Fully Flexible Semi-Structured Capture Solution Works on Semi-Structured Forms and Documents

A UB-92 form, used in the medical industry, is an example of a fixed form with many variations. It can be processed with a fixed forms solution, but that would require creating a new fixed template for each variation. With the semi-structured capture technology, a single flexible layout is sufficient for all variations; therefore, maintaining a semi-structured solution is more efficient.

Another good example is a Remittance in Process Report form. This is a flexible form with tabular layout, which has repeated lines with a similar structure. There is some information that belongs to the whole page level, and then there is a table below, which can have more or fewer lines with a similar structure. A semi-structured solution is the best approach to capture such forms, because we do not know prior to scanning how many lines each new document will have, plus documents printed by different companies can have quite different layouts for the table.

EOB and Remittance Paid/Denied forms can be used as examples of forms with mixed transactions. There can be different numbers of transactions per page, the transactions can be of different types, the order of data inside transaction can differ, and at any moment the transaction can flow across the page to the following page. Here we also have several levels of nesting in the information hierarchy ladder: first, there are several page-level fields, then many transaction-level fields, and then many transaction detail line-level fields. A form that is so complex cannot be captured with fixed forms technology at all; such form presses against the limits of the capabilities of the latest semi-structured capture technologies.

Multi-page mortgage applications provide another good example of a semi-structured document, with varied indexing data positions, data flow across the pages and a mixture of forms and semi-structured documents created using text-editing programs. And all this mixture has to be processed as a single scanned paper stream, with data captured from different positions in forms and documents across the folder of documents submitted by the agent. That is another great place to apply semi-structured capture technology.



Examples of semi-structured documents

Total Cost of Ownership of Solutions for Fixed and Semi-Structured Forms

In order to correctly access the cost of each of the two types of solutions, one must consider the entire lifecycle of a solution and the total cost of ownership in each case, from implementation and training to everyday production, maintenance and upgrades for the solution. The costs comprising the total cost of ownership are compared in the table below.

	Fixed Forms	**Flexible Layouts**
Product purchase cost	Lower	Higher
Implementation cost	Many Templates	Single Layout
Operating cost	Low (labor-efficient)	Low (labor-efficient)
Maintenance cost	Changes applied to many templates	Changes applied to a single layout
Solution scalability	New templates to draw	Quick adjustment of layout

The purchase price for fixed forms solutions is lower than for semi-structured capture systems. This is the main reason to go with fixed forms – but only for tasks that can be done with fixed forms technology, which is less then 20% of all forms and documents used by businesses and other organizations.

Implementation cost is lower in the flexible layout world, because there is only a single layout to implement instead of a large number of templates.

Operating costs are about the same. In fact, quite often the operator that runs the capture system may not even know what is used to capture – fixed templates or flexible layouts, because for an operator, the information on the screen and the verification process look the same for both types of solutions.

Maintenance cost for flexible layouts is lower for the similar reason why the implementation cost is lower – there is only a single layout to maintain instead of multiple templates. If one more field has to be captured from the invoices (for example, the amount of tax total), then the change is applied to a single layout and not to each variation of fixed templates.

Solution scalability is considerably better with the flexible layouts: instead of drawing a complete new template, we simply have to do a quick adjustment to the existing layout – most likely only to a couple of fields.

According to market analysts' reports, the cost of software can amount to 25-50 percent of the total cost of ownership. This means that careful consideration of all involved costs can help with selecting the best technology for your specific needs.

Selecting a fixed or flexible technology is only one part of the problem that you have to resolve. The other part, which quite often is even more important, is to select a proper capture platform or an element that can be plugged into your capture system.

If you do not have a capture platform yet, it is highly recommended to get one. There are many great platforms on the market, such as Ascent Capture from Kofax, AutoStore from NSI and HP, Enterprise Forms from ABBYY, as well as great platforms from Captiva, Verity and other vendors. Selection must be based on usability, total cost of ownership, and compatibility with the existing solutions that your company uses – such as MFPs on the front end and document management systems and databases on the back end.

But if you already have a capture platform in place, do not reinvent the wheel! If all operators are already trained, the system is working, and all that you need is to extend its capabilities to include semi-structured documents and forms in addition to fixed forms – then get a pluggable technological component that can work seamlessly inside the capture platform that you already employ. This will minimize the initial implementation and training cost and will allow you to utilize your previous investment in the capture platform for many more years.

Conclusion

With new semi-structured forms capture technology, it is now possible to capture the wide variety of forms and documents that cannot be processed by traditional fixed forms software. The new technology also makes it possible to use a single capture platform to process fixed and semi-structured documents.

ABBYY is a leading developer of intelligent document recognition and forms processing technologies. There are several product lines from ABBYY – FineReader full-page OCR, FormReader form processing solution, and FlexiCapture technology available as a standalone product or as a pluggable technological component for other capture platforms. More information about ABBYY is available from www.abbyyusa.com.

BIBLIOGRAPHY

J. Wnek. Automated Data Extraction from Structured Documents. In *Proc. 2003 Symposium on Document Image Understanding Technology*, 31-36.

M. Koppen, D. Waldostl and B. Nickolay. A System for the Evaluation of Invoices. In *Document Analysis Systems II* (World Scientific, 1998), 223-241.

T. Bayer, H. Mogg-Schneider. A Generic System for Processing Invoices. In *Proc. Int. Conf. on Doc. Analysis and Recognition* (IEEE Computer Society Press, 1997), 740-744.

A.R. Dengel, B. Klein. SmartFIX: A Requirements-Driven System for Document Analysis and Understanding. In *Proc. 5th International Workshop on Document Analysis Systems V*, D. Lopresti, J. Hu and R. Kashi eds. (Springer, 2002), 433-444.

J. Wnek. Learning to Identify Hundreds of Flex-form Documents. In *Proc. of SPIE, Document Recognition and Retrieval VI*, D. Lopresti and J. Zhou (Eds.), Vol. 3651 (1999), 173-182.

J. Wnek. Machine Learning of Generalized ocument Templates for Data Extraction. In *Proc. 5th International Workshop on Document Analysis Systems V*, D. Lopresti, J. Hu and R. Kashi eds. (Springer, 2002), 457-468.

T. Breuel. High Performance Document Layout Analysis. In *Proc. 2003 Symposium on Document Image Understanding Technology,* 209-218.

D. Doermann, C. Shin, A. Rosenfeld, H. Kauniskangas, J. Sauvola and M. Pietikainen. The Development of a General Framework for Intelligent Document Image Retrieval. In *Document Analysis Systems*, 1996, 605-632.