



## White Paper: Fixed Versus Semi-structured Forms in ABBYY FlexiCapture

By: Jim Hill  
Published: May 2017



Summary: Anyone new to data capture will be faced with an immediate decision of choosing either a fixed or semi-structured approach to extraction of data from a form. What constitutes a fixed form versus semi-structured and what are some guidelines for distinguishing them in ABBYY® FlexiCapture?

One day, my office phone rang from someone who saw information about our products on our company web site. As a new sales person tasked with selling document capture, I was used to answering these type of calls; but this time the caller threw in an interesting spin. They needed ballpark pricing right now. They then went on to explain that they needed to extract data from a particular document and export it to a database. I began to ask them about their document and how the information was laid out on the page. Was the information always in the same position on the page, or did it move around from document to document? Were there multiple versions of the document and what was the annual page volume that they expected? These qualification questions were required because of the nature of the product I was going to recommend depended very closely upon the type and location of the data to be extracted from the form. Once I determined that they were most likely looking at a form in which the data moved around from document to document I was able to provide them with the ballpark pricing they required. Without that information, a verbal estimate would have been impossible. Why? As you will learn in this article, the structure of document is extremely important in determining the type of technology used to extract data from within the documents.

The solution that I recommended to my caller was ABBYY FlexiCapture. FlexiCapture is very powerful document capture software which automates the extraction of data from documents and makes it very easy to export this business-critical data to back end systems. One of the first decisions that must be made during the design phase of a new capture project is to determine the nature of the document from which the data is to be extracted. In the case of a scanned paper document, including one already available in electronic PDF format, this first decision boils down to the determination of whether the document is a **fixed form** or **semi-structured form**. The ABBYY software makes extraction of data from either type of form a straightforward process through the use of the included wizards. For more complex scenarios, the data extraction process can either be done manually or through the initial use of the wizards which can then be tuned according to your needs.

### Definition of a Fixed Form

A fixed form is, by definition, one in which the data to be extracted is always found in the same **absolute** position on a page. You can think of it this way; data is extracted through the creation of a lens grid containing the pixel values for the absolute locations of the data. The configuration process involves understanding the X and Y coordinates of the position of fields as well as the width and height of the fields.

A fixed form is configured in ABBYY FlexiCapture through the creation of what they call a “document definition.” According to ABBYY’s documentation, a document definition “describes the location of

document elements and indicated fields to be used in data extraction.” In order to optimize the collection of the proper data for any particular field the document definition includes the provision of rules defining the type of data expected in the field including a means to perform a database lookup of possible values with ready-made extraction rules provided.

A fixed form requires sufficient structure in terms of textual or line elements such that the document can be aligned with the document definition. To illustrate the nature of this requirement let’s examine a typical fixed form as shown in the image below.

**FORM 01**
**CUSTOMER'S REFERENCE DATA CONFIRMATION**

Main

Bank guarantees the accuracy of information on the Customer/the Holder specified herein in accordance with the General Terms and Conditions and the procedures established by applicable laws.

---

**PERSONAL DETAILS**

Mr  Mrs  Ms  Miss

First Name: [.....] Middle Initial: [.....]

Last Name: [.....] Nationality: [.....]

City of Birth: [.....]

Birth Date (mm.dd.yyyy): [..]/[..]/[.....] Social Security Number: [.....]

---

**ADDRESS/CONTACT DETAILS**

**FIXED ADDRESS**

Country: [.....] ZIP code: [.....]

State: [.....] City: [.....]

Street: [.....]

House: [.....] Building: [.....] Apartment: [.....]

Home Phone: [.....] Mobile Phone: [.....]

---

**PRESENT ADDRESS**

Country: [.....] ZIP code: [.....]

State: [.....] City: [.....]

Street: [.....]

House: [.....] Building: [.....] Apartment: [.....]

---

**PROFESSIONAL DETAILS**

Employed  Self-employed  Retired  Other

Name of Employer or if Self-Employed, Trading Name: [.....]

Occupation: [.....]

Country: [.....] ZIP code: [.....]

State: [.....] City: [.....]

Street: [.....]


House: [.....] Building: [.....] Apartment: [.....]

Work Phone: [.....] Mobile Phone: [.....]

Gross Annual Income, USD: [.....] Employment Date (mm.dd.yyyy): [..]/[..]/[.....]

---

I Confirm the Correctness of the Information above



6 529141 677079

(Signature)

Stamp

(Employer HR Officer Signature)

Document Date (mm.dd.yyyy): [..]/[..]/[.....]



To the extent that you have control over the design of a fixed form versus inheriting a backlog of forms from which you must extract data, there is a wide variety of design considerations for such a form. This includes the use of constrained fields, proper anchors to center the form, and a means to identify the document. The anchor elements in the image above are the small black squares in each corner of the form. The method used to identify the form in this image is the barcode on the bottom of the page, however a text string such as the form title could also have been used. Accurate extraction of handwritten values requires that constraints be provided on the form, such that the person hand printing the form is forced to print these values into a certain pre-defined grid pattern. In the form above, these are represented as the dot boxes shown to the right of the field names. If insufficient design criteria is specified, you always have the option of using ABBYY's FormDesigner tool, which is included with the software, in order to design a new form which would include the required fixed form elements.

### Definition of a Semi-Structured Form

Semi-structured forms are those in which the location of the data and fields holding the data vary from document to document. This type of document is perhaps most easily defined by the fact that it is not a fixed form. In the document capture industry, a semi-structured form is also called synonymously an unstructured form. Examples of these types of forms include letters, contracts, and invoices. According to a study by AIIM, about 80% of the documents in an organization fall under the semi-structured definition.

The extraction of data from semi-structured documents relies upon the use of business rules to locate the information within the document because the information on the page can move around from document to document. These business rules rely upon the fact that the data to be extracted is always in the same **relative** position to a defining characteristic such as a character string, a character string match to a regular expression, or a defining physical characteristic of the document such as certain line structure layout such as a table grid with column headings.

A semi-structured form is created in ABBYY FlexiCapture through the creation of both a document definition and an underlying entity called a "FlexiLayout." A FlexiLayout is, according to ABBYY, "a formalized description of a set of unstructured documents which enables a data capture application to locate data fields on the documents and extract information from those fields." A FlexiLayout is made up of elements and blocks. Elements are, according to ABBYY, "one or more image objects, such as separators, static text, pictures, etc. An element contains information about the type of the object, its geometric features, its likely location, and relationships to other objects." Blocks "correspond to fields on the documents which data must be captured. A block specifies the type of data that the field may contain and the coordinates of the image area where the field is likely to be found."

To illustrate the unique technical challenge created when the need arises to extract data from these types of documents, refer to the invoice document below.



J. Goldsmith manufacture

Delivery address  
 CROWN LTD

Billing address  
 CROWN LTD

**Invoice no. NT-RC-075873-04**

Reference	Designation	Unit	Qty	Net unit price	%	Total CHF	Sales
	Carried over		74			6'619.4	
AO.612.002.0.SP	WHEEL, CROWN 2080/31023 20030101 Pce: : 10	Pce	10	27.60	35	179.4	220
AO.520.005.0.SP	PINION, WINDING 2080/31120 20030101 Pce: : 10	Pce	10	30.60	35	198.9	220
AO.546.004.0.SP	WINDING STEM 2120/51010 WINDING STEM 20030101 Pce: : 15	Pce	15	15.10	35	147.3	220
AO.628.105.0.SL	ROTOR 2601/20580 20030101 Pce: : 10	Pce	10	15.90	35	103.4	220
AO.610.104.1.SL	COIL 2601/20590 20030101 Pce: : 10	Pce	10	37.00	35	240.5	220
AO.610.111.1.SL	COIL 2610/20590 20030101 Pce: : 10	Pce	10	16.80	35	109.2	220
AO.220.176.0.SL	YOKE SPRING 2610/61100 20030101 Pce: : 10	Pce	10	2.30	35	15.0	220
AO.270.115.0.SL	BRIDLE +, BATTERY 2610/20764 20030101 Pce: : 10	Pce	10	0.40	35	2.6	220
AV.FOURVT.HIS	MAINSRING 5020/20100	Pce	3	9.00	35	1.6	220
AO.234.033.0.SP	MOON PHASE JUMPER 2825/53083 20030101 Pce: : 5	Pce	5	10.00	35	32.5	220
	Carried over		167			7'665.00	

Notice some characteristic of this document which make the use of fixed form extraction process impossible. First, a varying amount of line items can be expected for future versions of this invoice from the J. Goldsmith vendor. Second, the document does not include much in the way of horizontal line data that makes it possible to align the form, and no vertical line elements are provided at all. Also, the form is altogether missing any sort of corner locating anchor elements. Finally, multiple different types of invoice layouts can be expected from the same vendor not to mention the myriads of formats that can be expected from the rest of the vendors.

There are still ways the software can extract data however. Notice the invoice number field, in the case of this document the field **label** ("Invoice No.") is located to the left of the corresponding field value ("NT-RC-075873-04"). Here an easy way to extract the invoice number value would be to include in the FlexiLayout the search string "Invoice No" and the expected relative location to be to the right of this string. There is an easier way. ABBYY has provided a special version of FlexiCapture to deal with invoice type documents, including purchase orders, called ABBYY FlexiCapture for Invoices. Other version of the FlexiCapture product are also available including ABBYY FlexiCapture for Mailrooms, which is designed to classify and extract data closer to the point of origin such that key business decisions can be made more rapidly.

## Conclusion

A fixed form is one in which the data to be extracted is located on the same absolute position on the page. In contrast, a semi-structured form is one in which the information moves around on the page. Data from both types can easily be extracted through the use of ABBYY FlexiCapture. ABBYY provides preconfigured version of FlexiCapture for the extraction of data from invoice and purchase order type documents.

## How User Friendly Consulting Can Help

We have been providing consulting services for ABBYY FlexiCapture, including both the desktop / server end user application as well as the API product, since its inception. Please reach out to us if you would like to have us solve your data extraction challenge or just for help with properly configuring your existing ABBYY FlexiCapture system. We provide a wide range of consulting options including ABBYY trained and certified personnel as well as a wide range of training options to get your employees up to speed on FlexiCapture very quickly.