ABBYY®

Insights into ABBYY® FlexiCapture™

White Paper

ABBYY® FlexiCapture™

Contents

Executive summary

For over 20 years ABBYY has been developing software technologies for automated text recognition, document capture, analysis, and classification. One of the most successful outcomes of this development, ABBYY FlexiCapture system offers powerful and intuitive instruments to streamline time-consuming and labor-intensive tasks associated with paper-based processes.

What makes ABBYY FlexiCapture stand out from other document capture software on the market is the exceptional flexibility, superior accuracy and transparency of its technology – three qualities that drive the development of our software.

This overview reveals key aspects of ABBYY's versatile technologies that allow easy transformation of paper documents of any type and complexity into business-ready data.

Data vs. paper

Paper is an inescapable part of business. Despite predictions, paper-based processes are still common for the vast majority of companies, which are constantly trying to get control over the floods of paper.

Document analysis and recognition technologies have advanced significantly. Optical character recognition (OCR) is now an everyday practice. Automated processing of fixed forms such as application forms, questionnaires or ballot papers is also widely accepted. And what about processing of other types of documents? Forms without a fixed layout, such as invoices, purchase orders, mortgage applications, explanations of benefits, legal documents, correspondence, contracts, patient records, etc., comprise about 80% of all business documents. The spectrum of critical document processing tasks can vary from simple image capture and indexing to highly intelligent data extraction with real-time integration into business applications.

In many cases, performing these tasks turns out to be quite time-consuming because the document types we mentioned have distinctive features that make it challenging to process them.

Fixed forms (or structured documents) are documents that always have exactly the same layout, which means that a particular data field is always located in the same place. This makes automation of form processing tasks quite easy, with the industry standard approach here based on the creation of templates. Each type of a structured document requires creation of one template, which in the case of a simple form can take a matter of several minutes.

Invoices are classified as semi-structured documents because the types of data they contain are generally similar, but the exact layout of different data fields may vary from one vendor

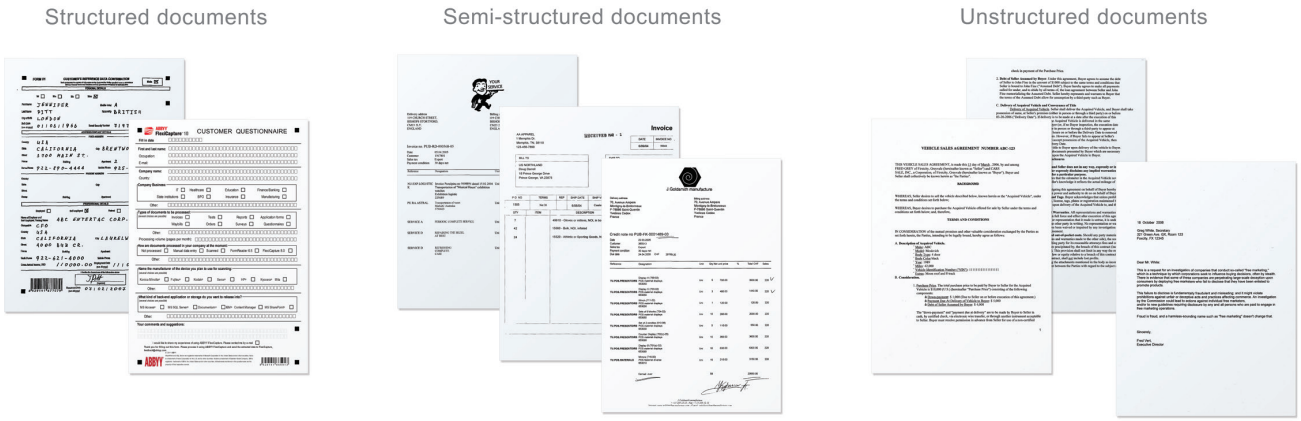

Figure 1. Document types based on the layout

to another. From invoice to invoice the number of line items changes as well as the number of columns, pages, etc. The same applies to purchase orders, explanations of benefits, price lists and so on.

Correspondence and contracts are considered unstructured documents because it is impossible to predict what kind of information they contain. Some key data like date or address can always be present, but the text of a contract or a letter is different in each case.

The unique characteristics of semi-structured and unstructured documents complicate processing and require more sophisticated document recognition methods as compared to fixed form processing.

Based on its extensive experience and scientific research, ABBYY has developed its own techniques for managing semi-structured and unstructured documents, which have been incorporated into ABBYY FlexiCapture software.

Think technology

ABBYY document capture and classification technologies are based on the principles of Integrity, Purposefulness and Adaptability (IPA) that imitate the way humans recognize objects. The same principles form the basis of the renowned and award-winning ABBYY FineReader OCR and ABBYY ICR technologies.

Invoice

AA APPAREL
1 Memphis Dr.
Memphis, TN, 38118
123-456-7890

RECEIVED MAR - 8

DATE	INVOICE NO
6/09/04	99044

BILL TO	SHIP TO
US NORTHLAND Doug Darrell 18 Prince George Drive Prince George, VA 23875	US NORTHLAND Doug Darrell 18 Prince George Drive Prince George, VA 23875

P O NO	TERMS	REP	SHIP DATE	SHIP VIA	FOB	PROJECT
1505	Net 30		6/08/04	Courier		

QTY	ITEM	DESCRIPTION	RATE	AMOUNT
7		49910 - Gloves or mittens, NOI, in boxes	150.00	1050.00
42		15560 - Bulk, NOI, inflated	275.00	11550.00
24		15520 - Athletic or Sporting Goods, NOI	317.00	7608.00
			TOTAL	20208.00

Figure 2. Example of numerical content of an invoice

Imagine you need to distinguish an invoice from other documents in the pile and then find the key data on it: invoice number, date, total amount, vendor address. How would you do that? Probably first you would look for specific words like “invoice” or “invoice number” which allow to identify the document as an invoice. The next step is to find the data fields. Guided by your past experience or common logic you probably would look for the invoice number, date and vendor address at the top of the first page and for the total amount at the bottom of the last page of the invoice. An invoice may contain several numbers (delivery note number, reference number, customer number, order number, etc.), several dates (invoice date, order date, shipment date, etc.) and various amounts located nearby which should be interpreted correctly. Some key words or elements located near the data field can help you to make a correct decision but there are cases when no key words are available. In such cases you would most likely examine the whole document and take the final decision based on the knowledge obtained about all the elements and their relative positions.

Based on the IPA principles, ABBYY FlexiCapture technology uses the same approach. It does not analyze each object separately but takes into account the relationships between all the objects and characteristics and then determines the best match for the whole set of objects. Such an intelligent and flexible approach to data extraction allows the software to find data anywhere on the document using any information available: content of the field, relation to other objects, the size of the field, lines or gaps nearby, etc. The technology works well even if it has to deal with documents of poor quality that cannot be perfectly recognized by the OCR engine.

The IPA principles-based approach distinguishes ABBYY FlexiCapture technology from the systems that analyze objects in a consecutive order. Such systems examine the relations only between the found and the subsequent element, but do not take into account the relations between all other objects found on the document image. Their main disadvantage is that when the software makes a wrong decision for the first object, it will fail to find other objects related to that first one. The flexible approach of ABBYY FlexiCapture delivers reliable results even while capturing data from documents with very complex structure or extremely variable layouts.

Inside a FlexiLayout®

ABBYY FlexiLayout™ Studio is specially designed to ensure hassle-free use of ABBYY’s cutting-edge technologies. As an integral part of ABBYY FlexiCapture software, this tool is designated for creation, testing and fine-tuning of formalized descriptions of documents’ layouts. A formalized description is called FlexiLayout and is interpreted by ABBYY data capture products for identifying document types and extracting data from the documents which it describes.

A FlexiLayout gives answers to 3 principal questions: how to identify a document, what data to extract and how to find this data. One FlexiLayout describes one document type.

Each FlexiLayout specifies a list of data fields for extraction and a search algorithm for detection of areas that correspond to the data fields. It also provides rules for document identification and detection of the end of a document. The FlexiLayout has a tree-like structure and consists of one block branch and one or more element branches, each branch corresponding to a layout alternative.

Blocks represent data fields which must be captured. A block is characterized by the type of data it may contain and by the area on the image where it is likely to be found (“region” in terms of FlexiLayout Studio). A region can be specified relative to the location of elements, which are detected first, by providing the presumed coordinates on the image or by combining the first two options. A region of a block may coincide with the location of an element meaning that the element serves as data source for the block. One block may contain data of several elements or their combinations (e.g. the block “name” can contain data of several elements “first name”, “second name”, “last name” and so on).

IPA principles

ABBYY recognition technologies 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 mind would analyze them.

Following the principle of Integrity, FlexiCapture treats a document as a single entity 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 instance, a compound element may contain several basic elements.

Following the principle of Purposefulness, FlexiCapture, just like the human mind, purposefully generates hypotheses about objects on a document. It performs this function by interpreting the formalized description of the document’s layout, which comprises all the information known about it.

Built-in Adaptability reveals itself in fuzzy logic applied by ABBYY FlexiCapture. Following this principle, the geometrical and spatial characteristics of an element are described by a range of permitted values and not by discrete quantities. The assessment of quality is expressed in percentage reflecting how well an object corresponds to its description, as opposed to discrete approach with simple ‘true’ or ‘false’ values. On different images the parameter values may differ. As long as they fall within a permitted range ABBYY FlexiCapture will choose the best available. This Adaptability accounts for variations in the document design and location of objects on an image.

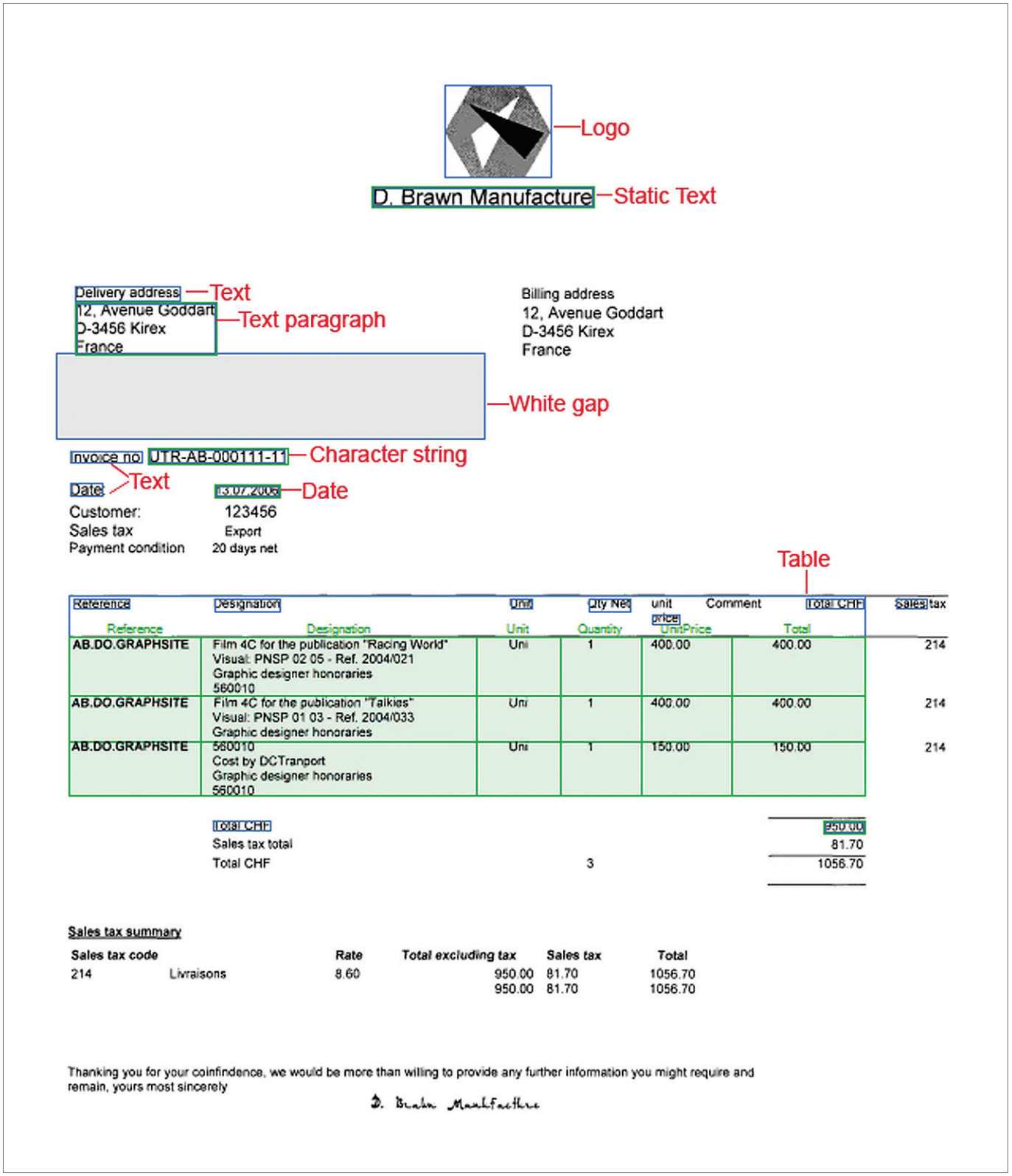

Figure 3. Examples of objects detected on image

An element describes the image object or collection of objects that can be detected in the document. It can be for example a logo, a text paragraph or a gap. Each object is characterized by its type, geometrical features, its likely location, and relationships with other objects. These characteristics or properties of an element allow the software to find the area for the object they describe, which is made of one or more rectangles enclosing the object. After the object in this area is detected, it can be used to locate the data field.

Because the same object on different pages or documents can be different, or even absent on some pages, the element should be generic enough to cover all possible variations of the object it represents.

An image can contain various objects that can serve as reference elements: lines, text strings, pictures, gaps, barcodes, etc. These objects have different characteristics. A barcode, for instance, is characterized by its value and type. A text string is specified by the language and the number of symbols. To simplify the description for all possible objects, ABBYY FlexiLayout Studio provides a sample set of typical elements along with their specific characteristics.

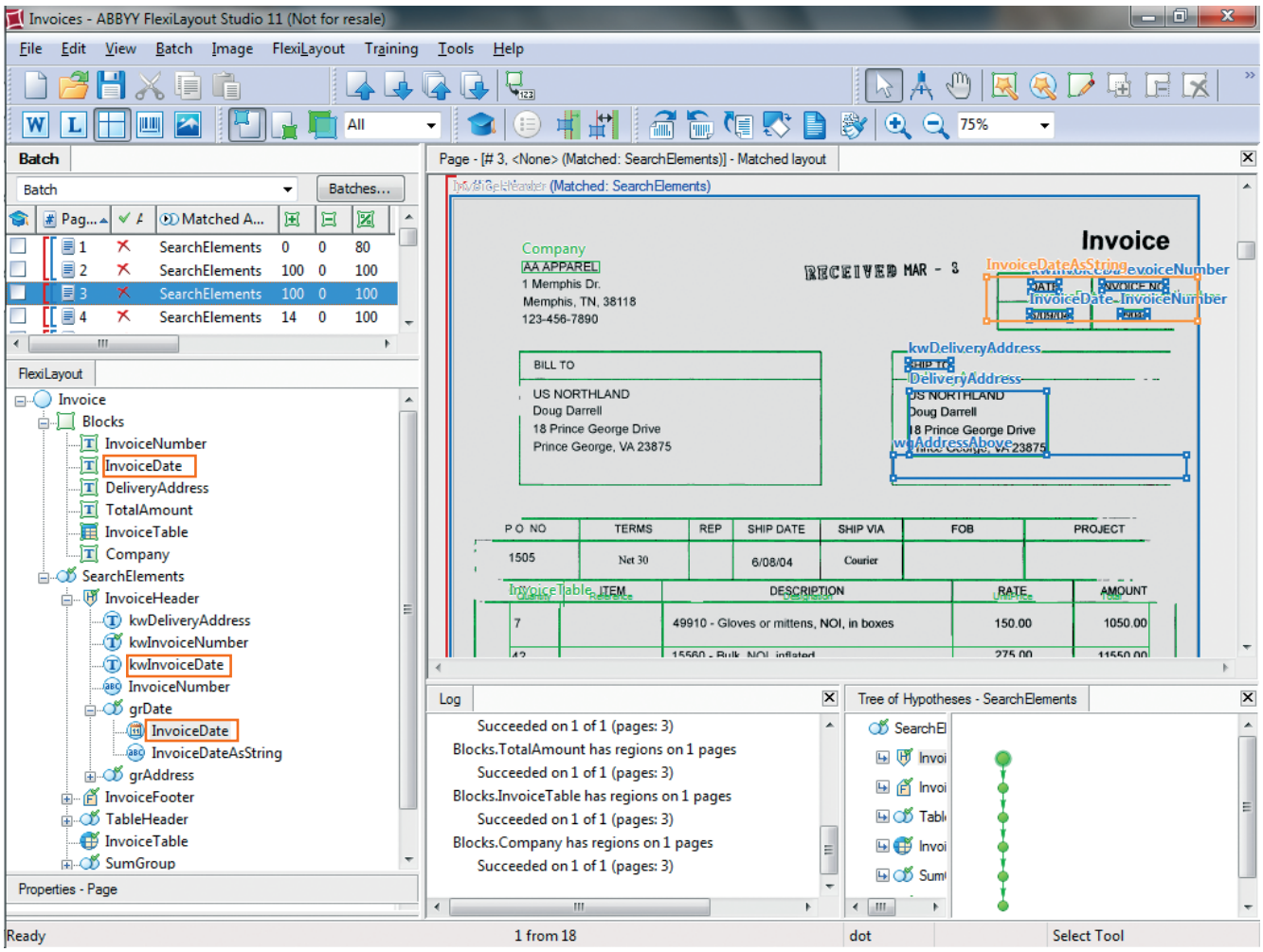

Figure 4. FlexiLayout blocks and elements

Let's consider a practical example. On the invoice above it is required to capture the date. In the branch of blocks, the corresponding block is called "InvoiceDate". To specify the algorithm for its detection, two elements are introduced. Element "kwInvoiceDate" serves as a reference object representing the keyword "date." The search area for this element covers the top section of the page. Element "InvoiceDate" contains the date itself in the numerical form which is likely to be found to the right or below the element "kwInvoiceDate". We should link the block "InvoiceDate" to the element "InvoiceDate". During the processing, the software will find "kwInvoiceDate" first, then detect "InvoiceDate" and finally render the recognized data to the block "InvoiceDate", which in its turn will be exported to a back-end system or database.

Hypotheses tree

When the FlexiLayout is being matched against an image, the system tries different scenarios to establish a correspondence between the elements and the objects on the image. Each possible correspondence is called a hypothesis. The system ranks all hypotheses in terms of quality. If there are 2 hypotheses relating to one element then it will choose the one with the best "quality". It estimates how well a particular object matches the described element – the better the match, the higher the quality of the hypothesis.

The hypotheses generated during FlexiLayout matching are organized in a tree-like hierarchy. FlexiCapture uses each of the hypotheses of the current element as starting points to look for the subsequent elements located below the current element in the tree. Thus, the hypotheses for elements branch out, which results in a tree of hypotheses that contains many more branches than the tree of elements. Ideally, FlexiLayout generates only one chain of hypotheses on every image, which means that all the objects are identified uniquely.

When the software matches a FlexiLayout against an image, it needs to find a complete branch of hypotheses with the highest quality. A branch is complete if it includes all mandatory elements, from the top element to the bottom one.

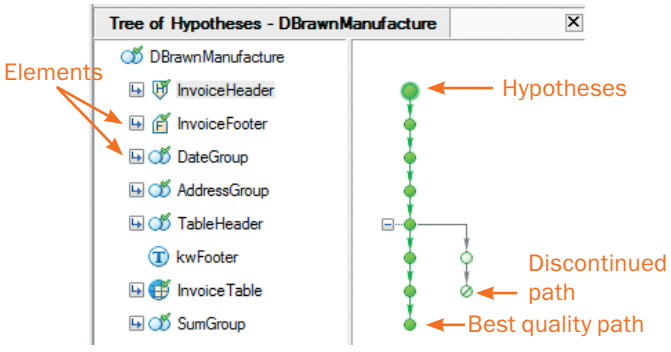

Figure 5. Tree of hypotheses

The user can view the tree of hypotheses. Actually it serves as an effective tool for fine-tuning the FlexiLayout because it clearly demonstrates how many and exactly which objects correspond to search constraints given for each element. Then the user can easily adjust the parameters in order to improve matching.

Headers

Information about blocks and elements in the FlexiLayout also serves to identify document types and differentiate them from each other, i.e. document classification. To speed up this process and ensure correct detection of first pages in multi-page documents, a special element type, which is called "Header", can be used. It is a compound element that can occur in the document only once. Headers of different document types are being matched independently from other parts of FlexiLayout. They don't compete in quality; the first successfully matched header defines the document type. After the document type is established, the corresponding FlexiLayout is applied to capture the data. If no headers were specified, the standard search algorithm will be used and the FlexiLayout with the best matching quality will determine the type of the document.

Headers also help separate documents with variable or unknown number of pages from each other. In this case, "Header" serves as a first page identifier. In order to find the last page of document, a special element called "Footer" can be created. In case the last page has no distinctive features and it is impossible to define the footer, the system looks for another header on the next pages. Once a new header is found, the system defines the previous page as the last page of the previous document. After finding the first and the last pages, FlexiLayout can be matched to this multi-page document.

To speed up classification, it is also possible to set up the order of FlexiLayouts and allow the software to use the first FlexiLayout matched without trying the rest.

Layout alternatives

There are some documents that have identical data fields to capture, but the layout of fields may vary a lot from one document to another. A typical example – invoices from different vendors. To process such document types in the easiest way, a user can design a FlexiLayout that include several layout alternatives, each of them corresponding to a particular group of documents with identically positioned elements. FlexiLayouts with layout alternatives are not only easy to create but also easy to adjust and maintain.

All layout alternatives share one common set of fields but have independent sets of elements and represent separate branches in the tree of Elements.

The order of layout alternatives in the tree determines the order in which they will be applied during FlexiLayout matching. It can be easily changed by dragging and dropping. Once the first match is found, FlexiCapture stops applying other layout alternatives.

Most typically, FlexiLayouts with layout alternatives are created for invoices, purchase orders, explanations of benefits, etc.

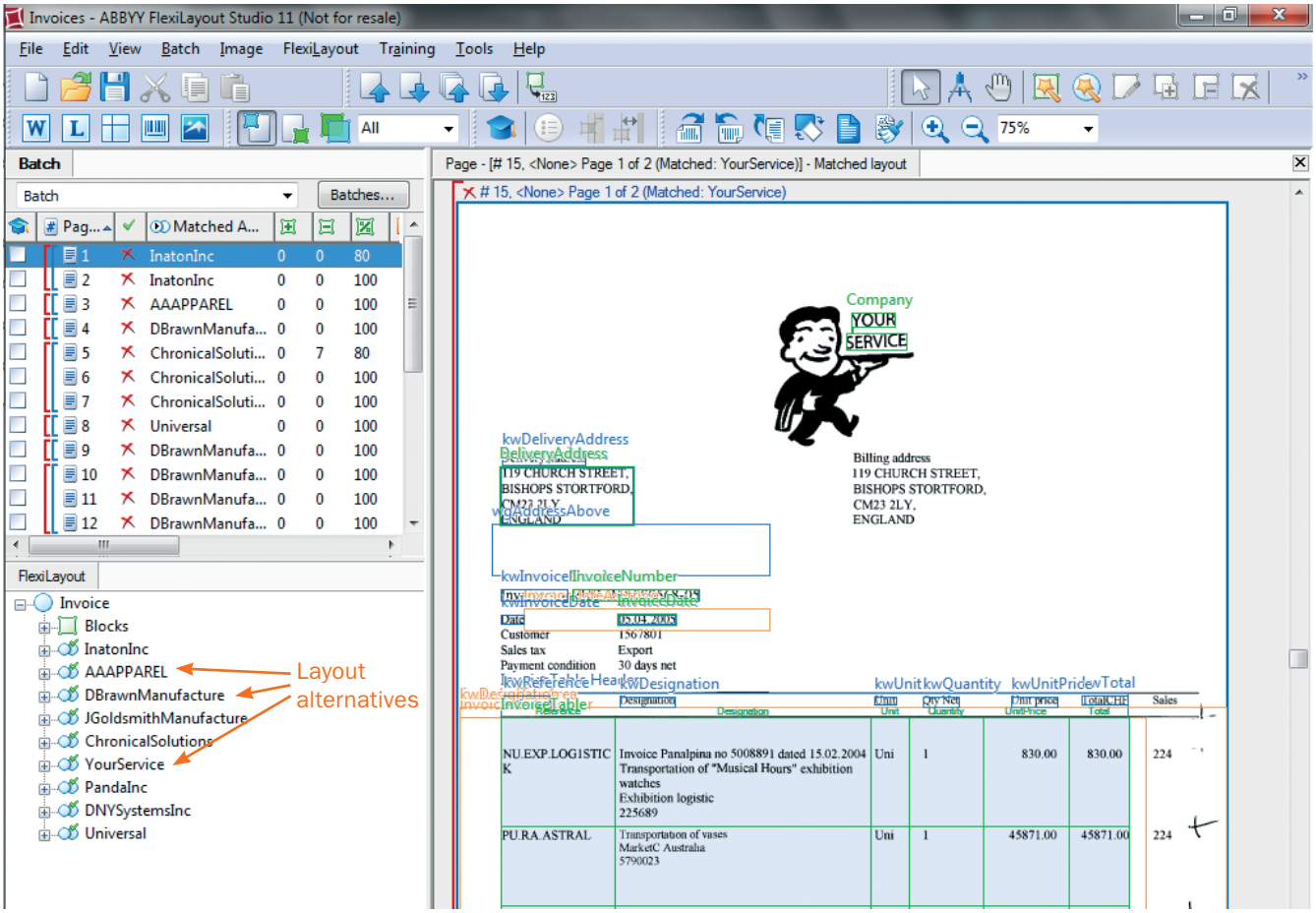

Figure 6. FlexiLayout with layout alternatives

Using the technology

ABBYY FlexiLayout Studio has a user-friendly interface and features a set of intuitive tools and instruments that do not require extensive training. It is notable for its universality and flexibility, providing inexperienced users with an immediate start while enabling the experts to use ABBYY's highly accurate classification and data capture algorithms to the full.

The process of creation of FlexiLayouts is pretty straightforward and includes three steps:

1. Creating a new project and loading images
2. Definition of data fields, search elements and specification of their properties
3. Fine-tuning

Creating new project

As a first step, it is necessary to create an empty project and add sample images of documents intended for further processing. The list of image files as well as all other relevant information about a FlexiLayout is stored in the project. Generally, up to 10 images for each document type is enough. It is advisable to select images that reproduce layout variations of the same document type.

Defining the data fields, search elements and specifying their properties

At this stage, ABBYY FlexiLayout Studio offers the user two alternatives: to manually configure a layout using one's creativity and logical skills or to leverage automated FlexiLayout generation and featured auto-learning capabilities. Let's have a look at both options.

Automated FlexiLayout generation

A FlexiLayout can be automatically generated as a result of training on a set of images. ABBYY FlexiLayout Studio enables the user to mark out necessary data fields and reference elements on the image in the training mode. While training, the software learns the location of fields and elements on each image and automatically determines properties and relations between them. When fields and elements are marked on each sample image, a FlexiLayout can be automatically generated by clicking on the relevant button.

All automatically generated FlexiLayouts are available for review and correction. ABBYY FlexiLayout Studio gives the user the full control of the layout design. It doesn't use any hidden

logic, and this distinguishes ABBYY software from other document capture systems featuring auto-learning capabilities.

Automated FlexiLayout generation ensures excellent results for one-page documents with relatively simple structure containing text, black lines or barcodes that can be used as reference elements to establish the location of the data fields. Correspondence is a good example. When dealing with multi-paged documents or documents with sophisticated layouts, automatically generated FlexiLayouts can be used as a basis instead of starting from scratch. Such a combination of the auto-learning capabilities of the software and expertise of the user results in unsurpassed accuracy in data location and capture while minimizing time required for setup of the system.

Manual FlexiLayout creation

The process of manual FlexiLayout creation, actually, can be rather enjoyable. The software offers an impressive set of easy-to-use visual tools to specify the required relations and characteristics and perform fine-tuning of the FlexiLayout.

Following the search algorithm of the software, the user must specify the list of data fields to capture, determine reference objects that would help to detect the fields, and describe the relations between them.

There are many alternative ways for selecting objects to detect, defining their corresponding elements, for combining elements into compound elements, and specifying relations between them. ABBYY FlexiLayout Studio's interface enables users to effortlessly evaluate various possible scenarios.

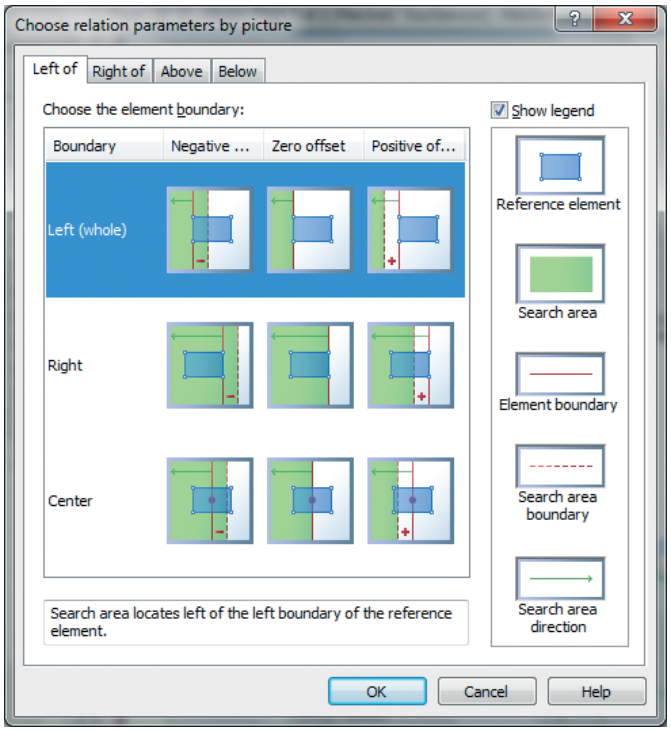

Figure 7. Visualized adjustment of relations between elements

In complicated cases requiring more detailed customization and assistance, FlexiLayout Studio provides direct access to its internal language for greater flexibility and control. Each properties dialog box contains an "Advanced" tab where the user may specify any additional relations or properties using FlexiLayout structural language.

Manual FlexiLayout creation obviously takes more time as compared to automated generation. However, this method gives an advanced user the full scope of possibilities available to them from using ABBYY's intelligent document recognition and data capture technologies.

Fine-tuning

After a FlexiLayout is generated by ABBYY FlexiLayout Studio, it can be tested on a batch of images and fine-tuned in case of mismatches. Sometimes it turns out that the set of element properties and relations that worked perfectly for one group of images, does not work for other images that differ from that first group. Such problems can be fixed by introducing new elements or by modifying the properties of the existing elements and relations between them. For that purpose, ABBYY FlexiLayout Studio was enriched with auto-learning tools enabling automatic adjustment of elements. To correct the detection of an element on a particular image, one can simply draw the right position and initiate training by selecting the appropriate item from the element context menu. The software calculates a new set of parameters based on the objects found around the new position and displays both the previous and new values to the user for comparison and approval.

When making modifications to a FlexiLayout, it is important to make sure that you actually improve the results on the whole test batch of images and do not degrade them. For the purpose of fine-tuning the FlexiLayout, it is possible to save a reference layout for each page and document that shows the desired location of fields on the image. Reference layouts allow you to compare the results of matching your FlexiLayout with the desired results. Reference layouts can be created manually or based on the results of FlexiLayout matching.

Classification

As it was mentioned before, once created a FlexiLayout is used by ABBYY FlexiCapture for document identification, classification, separation and data extraction. If the task is to process a large variety of documents in one single stream, the maximum productivity can be achieved with the use of a classifier. In this case, classification stage somewhat precedes data extraction. However, in terms of the software, document recognition is a single comprehensive process.

A classifier represents a small FlexiLayout that describes the unique features of the first page of each document type facilitating its identification. It doesn't contain any data fields. Thanks to its simple structure, the process of classifier matching takes very little time.

ABBYY FlexiCapture offers three classification modes: auto-learning, rules-based and combined. In the auto-learning mode, the software learns to identify different document types automatically in the course of training on a set of sample images. The rules-based classifier detects document classes using the instructions specified by the user. In the combined mode, both classifiers can be consecutively applied: the automatic classifier runs first and in case of unconfidently classified images they are additionally checked against the rules-based classifier.

When processing documents, ABBYY FlexiCapture refers to the available classifier first. The document type of images successfully passed through classification is added to their properties. Images that could not be confidently identified by the classifier are attributed to an "unknown document" class. On the next stage, the software applies FlexiLayouts based on the classification results instead of blind matching. For images in the "unknown document" class, it tries different FlexiLayouts just like it would if no classifier is used. Thus two different FlexiLayouts are applied to one image: the first one to identify the document type and the second one to find the data.

If the number of possible document types intended for processing is not high, there is no need to focus on the classification stage that much and separate it from the data capture process. FlexiLayout matching can be accelerated through their prioritization and the introduction of header elements.

Choose ABBYY FlexiCapture. Why?

Used by thousands of organizations worldwide since 1997, ABBYY's document identification and data capture technologies have been proven in thousands of successful projects. Skillfully packed into a single comprehensive software solution, these technologies deliver unmatched accuracy, streamlining your document-driven processes.

ABBYY FlexiCapture wraps the technology in a user-friendly interface, equipping the user with an impressive set of various tools that can be flexibly applied. A clever use of the software's capabilities fulfills a wide spectrum of document processing tasks like

- Automated processing of documents of different kinds in a single stream
- Data extraction from poor-quality fixed forms (e.g. faxes) or with significant distortions that cannot be compensated by traditional template-matching technologies
- Recognition of documents with complex structure (with continuous tables or repeatable document sections) and multi-paged documents
- Intelligent identification of document types and assembly of pages into documents
- Intelligent classification enabling documents to be arranged according to user preferences
- Highly accurate document classification based on auto-learning techniques

To guarantee easy and efficient adoption of its powerful technologies, ABBYY provides excellent technical support and professional services.

Transforming lists and tables

Excellent recognition of table-like data is what we are justifiably proud of, and it is perhaps one of the most intricate tasks in document processing. When transforming lists and tables, the software has to omit the appealing formatting of a document (letterhead, logo, subtotals, etc.), and produce only pure data in the output.

Whatever the requirements to the output data may be, whatever the layouts of the original documents may be, ABBYY FlexiCapture carefully extracts information and transforms it into ready-to-use data.

Here are just several examples of the software's potential:

- Conversion of brickwork-like tables
- Accurate recognition of tables containing subtotals
- Preservation of the continuous numbering of line items when transforming multi-page tables
- User-defined order of columns in the output, allowing variations and arrangement according to rules
- Selective data capture from user-defined columns of the table

About ABBYY

ABBYY is a leading provider of document recognition, data capture, and linguistic technologies and services. Its key products include the ABBYY FineReader line of optical character recognition (OCR) applications, ABBYY FlexiCapture line of data capture solutions, ABBYY Lingvo dictionary software, and development tools, supporting a variety of platforms and computing environments. ABBYY Language Services provides comprehensive linguistic solutions to corporate customers. Paper-intensive organizations from all over the world use ABBYY solutions to automate time- and labor-consuming tasks and to streamline business processes. ABBYY products are used in large-scale government projects such as those of Australian Taxation Office, Lithuanian Tax Inspectorate, Ministry of Education of Russia, Ministry of Education of Ukraine, Montgomery County Government of the USA, and Government of Canada. Companies that license ABBYY technologies include BancTec, Canon, EMC/Captiva, Hewlett-Packard, KnowledgeLake, Microsoft, NewSoft, Notable Solutions, Samsung Electronics and more. ABBYY OCR applications are shipped with equipment from the world's top manufacturers such as Epson, Fujitsu, Fuji Xerox, Microtek, Panasonic, Plustek, Ricoh, Toshiba, and Xerox. ABBYY group consists ***of 14 global offices*** located in Russia (5 offices), the USA, Germany, the UK, Japan, Ukraine, Australia, Canada, Cyprus and Taiwan. Most of the research and development projects are conducted in Moscow headquarters.

"...It's been fantastic. ABBYY accommodates nearly all of the insurance carriers' different forms, which is no small feat since there are hundreds of types. The hours we no longer have to dedicate to data entry makes ABBYY FlexiCapture an outstanding value, we're so much more productive."

Chelsea Sprague
Billing Specialist and Client Liaison On-Site Anesthesia Services, Inc., USA

"At first, it was actually a little unexpected to see such high speed and yet top accuracy in processing sophisticated and complicated forms like HCFA. Thanks to ABBYY, we are able to submit projects to our customer on time and with a high quality."

Manny S. Miranda
Project Manager, Accurance-AABP
The Philippines

"ABBYY FlexiCapture simply makes possible tasks that we would not have been able to undertake before and has streamlined our survey process so that our staff has less work to do on administration and can focus their time on working with young people."

Steve Maddison
Manager for Connexions, United Kingdom

We would like to thank you for taking the time to learn more about our technology.

Please, take the opportunity to try ABBYY FlexiCapture and enquire about training and certifications for the product by contacting our sales offices.

ABBYY USA
880 North McCarthy Ranch Blvd., Suite #220
Milpitas, California 95035, USA
Phone: +1 408 457 9777
Fax: +1 408 457 9778
E-mail: sales@abbyyusa.com

International Headquarters
Otradnaya str. 2b/6, 127273, Moscow, Russia
Phone: +7 495 783 3700
Fax: +7 495 783 2663
E-mail: office@abbyy.com

ABBYY Europe GmbH
Elsenheimerstrasse 49
80687 Munich, Germany
Phone: +49 89 69 33 33 0
Fax: +49 89 69 33 33 300
E-mail: sales_eu@abbyy.com

ABBYY 3A (Asia, Africa, South America)
Otradnaya str. 2b/6, 127273, Moscow, Russia
Phone: +7 495 783 3700
Fax: +7 495 783 2663
E-mail: sales_3A@abbyy.com

ABBYY Ukraine
Moscovsky av. 13-B, 04073 Kyiv, Ukraine
Phone: +380 44 490 9999
Fax: +380 44 490 9461
E-mail: sales@abbyy.ua

ABBYY Russia
Otradnaya str. 2b/6, 127273,
Moscow, Russia
Phone: +7 495 783 3700
Fax: +7 495 783 2663
E-mail: sales@abbyy.ru

ABBYY®