

ABBYY FlexiCapture White Paper

ABBYY FlexiCapture: What It Is

Automated data input technologies have a relatively long history – dating back to the days when the first optical reading systems were developed to recognize stylized symbols drawn according to templates. Since that time, they have evolved to support a vast industry, utilizing a large set of very different technologies.

The traditional machine-readable form processing technologies of today are well-established. A large choice of systems capable of processing many types of machine-readable forms is now available. Today's advanced systems can accurately capture machined printed and hand-written characters and process thousands of documents per day. ABBYY FormReader is one of the leading products in the field, capable of handling both printed and hand-printed forms (see <http://www.formreader.com> or contact ABBYY for a whitepaper and additional information on ABBYY form processing technology).

Yet while today's form processing systems are very advanced, they are still limited in functionality. For example, the task of processing semi-structured documents, or forms and documents on which the sizes and locations of fields of key pieces of data varies from document to document, still remains the most challenging task in data capture. While the demand for solutions to address this area is extremely high, form processing programs have not been flexible and intelligent enough to process these types of documents without extensive customization and system training. Access to an easy-to-deploy, cost-effective solution for processing such documents as invoices, order forms, legacy forms, and template-based contracts has been, until now, virtually inaccessible by a large audience.

For these types of cases, even when full-text documents are being handled, the ultimate aim is to extract a particular set of fields, or key pieces of information, from a given page. We will refer to such documents as **flexible forms**.

Flexible Forms

Forms normally contain data that is required or requested by the organization using the form. This variable data can include such things as names, addresses, and monetary amounts. On traditional forms, key pieces of data can be found in exactly the same fields, located on exactly the same position on the page, in the same sized field, from document to document. Processing this type of document is relatively simple if the form and the form template are designed carefully. The system simply needs to match the scanned form with the template to know what information to extract and how to extract it.

Flexible forms are documents in which the placement of the data, and the fields holding the data, varies from document to document. In most cases, only a few key pieces of information, or certain fields, are actually truly important. So the challenge is not only to find the important information, but also to understand which information is not important and then should be ignored. An example of a flexible form is a standard contract in which the name of the person, which is important for indexing purposes, is contained in the body of the contract text. In the majority of cases, the placement of the specific field cannot be defined or predicted exactly by a template, sometimes it cannot be specified even roughly. So the technology must find a way to locate these fields. The lack of distinctness makes the development of flexible forms capture technology a real challenge.

After many years working with companies processing a variety of flexible forms, ABBYY has developed FlexiCapture™: a technology that enables the capture of data from semi-structured documents. FlexiCapture is based on award-winning ABBYY FineReader OCR technology, and delivers unprecedented power in handling different kinds of complex documents.

ABBYY FlexiCapture: Why You Need It

Data on paper can be very difficult to capture. Even in the best cases in which the original documents were designed with data capture in mind, one single mistake or omission may turn a seemingly easy data capture project into very complicated task. In very many situations, the tools that are available in the traditional data capture environments are not flexible enough to solve a particular task, or they require solutions that are complicated, expensive, or difficult to use and maintain.

ABBYY FlexiCapture provides a solution for a wide variety of data capture problems by giving the system increased intelligence and flexibility. Through the FlexiLayout™, a logical definition of layout of data – FlexiCapture frees you from the restrictions of template-based form matching (such as reliance on the exact placement of fields on the page). The system can find fields elsewhere, using any information available: relation to other objects on the page, contents of the field, its size, lines drawn around, etc.

- Have to deal with faxed and copied forms, where there is a distortion that cannot be compensated by traditional form template matching technologies?
- Need a tool to process custom forms collected from different sources?
- Have to extract some specific information from text documents, such as contracts or business letters?
- Have to develop a customized invoice processing application?

FlexiCapture is exactly what you need.

ABBYY FlexiCapture: The Technology

From the description above it looks like ABBYY FlexiCapture is magic that solves all data capturing problems. Of course it is not magic. FlexiCapture is specialized technology based on ABBYY's more than 10 year experience in recognition and document analysis technologies. Initially designed as an in-house technology, it has served as a platform of many successful projects for more than 4 years. In fact since 1997, a customized data capture application based on the FlexiCapture technology has been used by several large banks across Eastern Europe for the specific task of processing paper-based payment orders.

ABBYY Technology for Bank Payment Orders

ABBYY FlexiCapture used in a specialized and customized document capture solution created by ABBYY for financial institutions in countries where bank money transfers are made based on so-called "payment orders." A typical payment order contains about 20 fields, with the most important fields being the name of the "payer," the account number and bank information of the "payee" or "beneficiary," the purpose of payment, and the payment amount. As these documents are used for each money transfer in all the banks, the total circulation of such documents is normally very large. That's why there are typically some regulations in these countries about how the payment order has to be laid out. However, these regulations are normally very loose in the way they define the exact layout of information, and are more focused on the content of specific fields. As a result, the layout of the actual documents can vary significantly. For each different form, the payment amount, for example, could be found on a different part of a page, at the top, at the bottom, to the right, to the left, etc. The key for the banks is to be able to process large amounts of these forms, no matter what the format.

In answer to this need, ABBYY developed a product based on FlexiCapture technology. This product is used by more than 500 banks worldwide, and more than 300,000 pages of payment orders are processed every day by ABBYY technologies.

Over the years, ABBYY has received numerous requests to transfer this custom-built technology into a tool for integrators. That is the reason why ABBYY developed FlexiCapture Studio.

The FlexiLayout

ABBYY FlexiCapture Studio is a tool that allows VARs, system integrators or IT department personnel to create, test and fine-tune flexible form definitions, referred as FlexiLayouts. The FlexiLayout is then interpreted by the FlexiCapture technology incorporated in several ABBYY products, including ABBYY FormReader and ABBYY FineReader Engine: it serves as a set of rules for the form matching process. The FlexiLayout enables the recognition system to easily find necessary fields on the flexible form. Once located, the data in the fields can be extracted using the OCR/ICR/OMR and barcode-reading functionality provided by ABBYY FineReader Engine.

To implement the FlexiCapture technology you may choose to either use ABBYY FormReader to create a solution for the customer, or integrate FlexiCapture into your application using ABBYY FineReader Engine SDK.

(For more details about these products refer to their appropriate whitepapers, available from ABBYY).

FlexiCapture is built on powerful and time-tested ABBYY technologies based on IPA principles that imitate the way humans recognize objects. The same principles lay in the basis of the proven and award-winning ABBYY FineReader OCR technology and ABBYY's ICR technologies. With FlexiCapture Studio, ABBYY provides intimate access to its internal technologies – at a level that has never been available before.

ABBYY FlexiCapture: Where To Use It

One thing has to be made clear: ABBYY FlexiCapture is a technology and 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 FlexiCapture Studio to create solutions for capturing a wide range of document types. The result is the creation of a FlexiLayout. The end solution that is delivered to the customer, is the the FlexiLayout and a product that has the FlexiLayout interpreter incorporated in it (one of the ABBYY recognition products, see below for an outline of products supporting FlexiCapture).

FlexiCapture can be used to capture data from the following major classes of documents:

- printed forms, including invoices, purchase orders, reports, etc.
- documents having some regular fields (e.g. standard contracts, business letters, etc)

It is quite possible that there are projects that involve a need to process yet another type of semi-structured documents. FlexiCapture is flexible and powerful enough that it could be applied to a number of unspecified tasks without limit. Benefits of using the FlexiCapture are clear enough. Normally, the kind of problems that are solved by FlexiCapture technology require use of some recognition engine and development of sophisticated algorithms to search for specific information on the page. That approach is expensive, risky, time consuming, and complicated. This approach also does not allow you to create simple working prototypes easily.

Using FlexiCapture Studio does not require programming skills and prototyping a solution is a breeze.

In most of the cases, it is possible to create first working prototype that reliably captures few most important fields just in an hour or two.

ABBYY FlexiCapture: How It Works

In theory, capturing data from documents seems to be a relatively easy task. Even when one has to deal with semi-structured forms, there is no major problem from the first glance. The solution seems quite simple: find

anchor objects for the fields for which you are looking, detect those objects based on their contents, and then easily locate the fields nearby.

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 instance, if there is a field with a person's address, then normally one should expect the word "Address" before the field (for instance, below the field). These guiding objects on the form, normally designed to simplify its perception by a human, are very useful when searching the form for fields. Such elements are referred as "anchor objects." The traditional approach is to use some OCR technology to read the contents of the page, then program a procedure that looks for specific fields by using anchor objects. One obvious drawback of such approach is the possibility that wrong anchor objects are identified, because the form is not considered in the whole. For instance, 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, the incorrect data will be captured. FlexiCapture utilizes a much more powerful technique, based on the IPA technology, which allows the technology to analyze the form as a whole and consider all the possibilities of data placement to determine the best choice.

Managing Variations on Flexible Forms

In practice, however, the situation can actually 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 extract only a part of the text, or the text with some mistakes. Also, the same word, or even few words, found in the anchor text could also be printed somewhere else. There is also a chance that on some forms the field is on one single line, while on others it is several lines, and it is not obvious how to distinguish lines belonging to the field from the 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 bad handling of the paper.

So in the real world, one has to expect vast variation from the ideal form model. The real working technology must be able to account for all variations and deliver reliable results.

The utmost strength of FlexiCapture is that its document model works even though such variations exist. Because it is based on ABBYY's intelligent technology and the principles of IPA technology (see inset on IPA), FlexiCapture technology never relies on any fixed presumptions: you may specify any object or its properties to be tentative. Using the principles of IPA, FlexiCapture 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 of the objects and the characteristics of each of the objects. Only then does it determine the best match for the whole set of objects.

Another system that analyze individual objects one by one – without accounting for the whole and without examining the relationships between objects – the system would fail. If a wrong decision was made for the first object the system would fail to find all other objects that are related to that first object.

IPA Technology

ABBYY recognition technologies are built using 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**, FlexiCapture 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 their 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 brain, purposefully generates hypotheses about objects on a document. It performs this function by interpreting the FlexiLayout, which is*

the essence of the human programmer's knowledge about the specific flexible forms. The **Adaptability** built in into the technology, allows FlexiCapture to more precisely generate hypotheses about specific objects based on the information collected from other parts of the image. With further technology improvements, the adaptability will go even further, making automatic adjustments to improve the FlexiLayout based on the analysis of the real documents being processed. In other words, the system learns and trains itself over time.

The reason that IPA works for flexible form processing is 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. When such an approach is used, there is no way to cover large variations of the form design unless one were to develop a separate new template for each and every possible variation. Such an approach, is obviously expensive, difficult to maintain, and limited to only specific cases when it is still possible to use fixed templates.

How the FlexiLayout Works

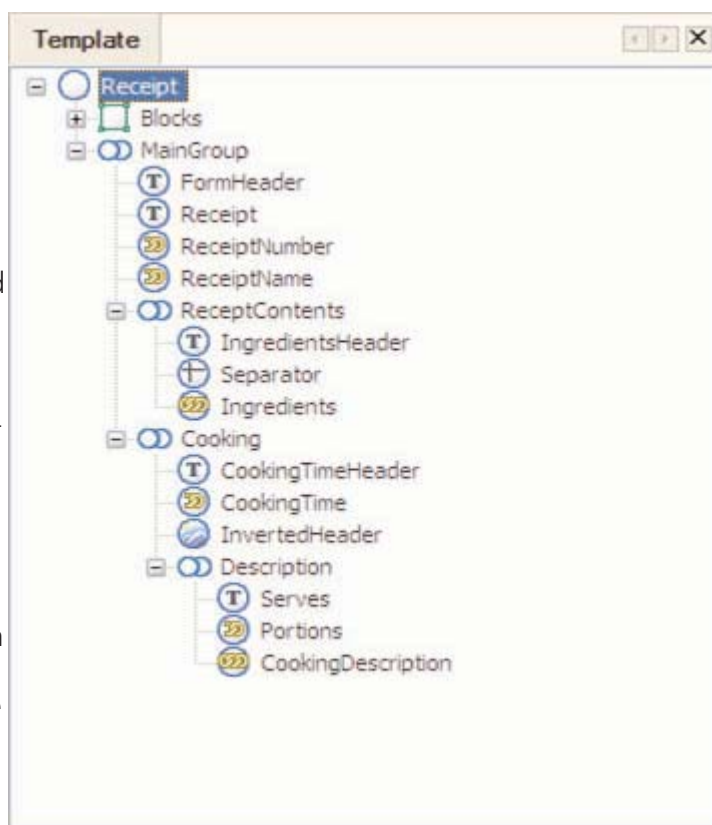
The FlexiLayout is a set of elements organized in a tree.

An element is a set of distinguishing characteristics of a specific object on the page. When the FlexiLayout 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.

FlexiCapture provides powerful tools to allow for such a generic description. You may specify ranges for its placement, the contents in terms of regular expressions, substrings or set of variants, relations to other objects on the page. It is possible that the element cannot be found on the page. It is also possible to describe the single object using several elements of different types and then pick up the element that was found during layout analysis. This is very useful when on different forms the same logical element has different contents. If the FlexiLayout is well-developed, there are always several ways to find necessary information and when one fails, the other works.

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 separate the whole form into logical groups of elements instead of dealing with a large set of separate elements. Any compound element could in its turn be combined into another compound element. There is no limitation on nesting (see the below picture for the example of the element tree).

When the FlexiLayout is matched against the page, the system tries different scenarios for setting correspondence between the elements and the objects on the page. Each specific possible correspondence is called a hypothesis. All the hypotheses generated during the FlexiLayout matching are also organized in a tree. This is because for any ele-



ment, there may be several hypotheses, and as hypotheses for the elements are interrelated depending on the hypothesis of one element, different hypotheses for the related elements are generated. By navigating and examining the hypotheses tree, the developer can find out the reasons why something occurs during the template matching, identify problems, and modify elements to improve layout matching.

Below we will see how the developer may effectively control the way FlexiCapture detects objects on the page. But first we will learn to make first steps with the FlexiCapture Studio.

ABBYY FlexiCapture: Hands-On

The process for using ABBYY FlexiCapture technology is simple to learn because it is very similar to the process for fixed forms. The major difference is simply that one must use ABBYY FlexiCapture Studio to develop the FlexiLayout, instead of developing a fixed form template using a form template editor.

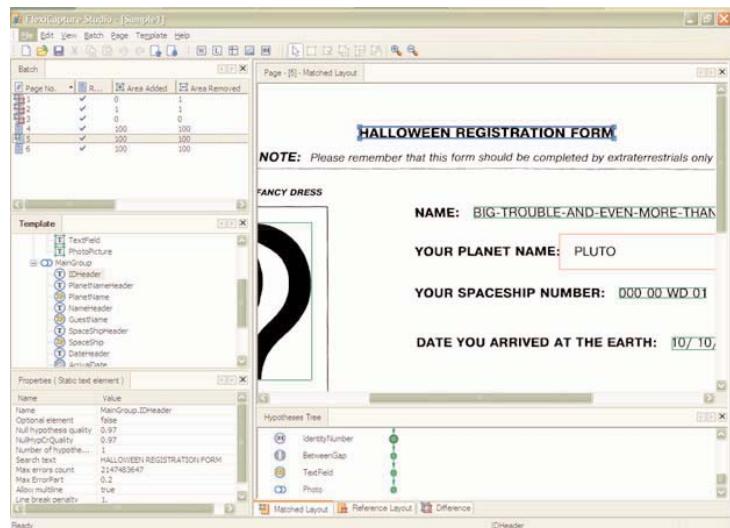
Creating a FlexiLayout

The steps for creating the FlexiLayout are as follows:

Creating the project

As a first step, an empty **project** must be created. The project stores all the relevant information, including the elements, the list of image files, the FlexiLayout data, etc.

Next, it is necessary to add images of sample documents to be processed. The set of images should be varied enough to represent the range of forms to be processed. Very similar documents do not really help to improve the quality of FlexiLayout, but any sufficient variations of the document layout should be represented. The more variations of the document layout are covered the better. The main window of the FlexiCapture Studio is shown on the picture.



Defining the fields

The next step is to name the fields we are going to capture. It is very important at this step to make it clear what should be included in each field and what should be skipped. That is because in many cases not only the layout of the form is "flexible," but the rules outlining how the form should be filled out could be vague. For instance, in one case you may have the first and second name put into one field, while in other case they could be separated. The list of fields could be easily modified later on. Each field should be later related to one or several elements. There is a way to describe any complex relations between an element and the fields.

The best way to get acquainted with ABBYY FlexiCapture Studio is to try it. It is accompanied by a tutorial containing step-by-step instructions how to create FlexiLayouts for two example forms. The material covered by tutorial is enough to start successfully using the product. Also ABBYY provides trainings and certification for the product, see <http://www.abbyy.com> for details.

Defining elements

As soon as all the fields are defined and are properly described by means of FlexiCapture Studio, you may proceed with creating the elements. This step requires some experience, because there are several possibilities for selecting objects to detect, defining corresponding elements, combining elements into compound elements, and selecting relations between elements. The user interface of FlexiCapture is organized in the way makes it possible to investigate different possible scenarios without too much effort.

What happens if at some point you find out that there was a mistake, for instance in the way you grouped elements in the tree? The solution is simple: just drag the element from one place in the tree to the other. If such action results in invalid relations, the elements that have to be reviewed are marked in the tree. In that case you must modify their settings by double-clicking on these elements in the tree to open the properties dialog box.

The more experience you have with the system, the easier it will be to figure out the right strategy. And will make fewer mistakes in the way you organize elements in FlexiLayout.

Defining the element

The element is created by defining its properties. There are several types of the elements, such as "Static Text," "Separator," "White Gap," etc (see side frame for details). These are basic objects that FlexiCapture is able to sense on the image. For instance, static text is detected by passing the image through an OCR procedure, while some other elements (such as separators) could be detected directly on the image.

Types of elements sensed by ABBYY FlexiCapture technology

Static text – any text on the page. The static text element is defined by the possible contents of the text. It is possible to specify different variants of the text to be found.

Separator – horizontal or vertical black line. It is possible that there are many such lines on the page. It is important to specify enough information to allow the system to distinguish the specific separator from others. Relations to other objects are normally very helpful, and specifying some general constrains on its length also works well.

White gap – horizontal or vertical area containing no objects. White gaps are very helpful for distinguishing between several specific layouts, as they are easy, fast and reliable to detect.

Barcode – any one-dimensional or two-dimensional barcode. ABBYY FlexiCapture detects barcodes types supported by ABBYY FineReader Engine.

Character chain – text that varies and may contain several lines, so that it cannot be reliably detected by the Static text element. For the character chain, it is possible to specify a regular expression for its contents, as well as specify a set of characters that can be found in it.

Object collection – any object, such as text, a punctuation mark, a picture or a checkbox. Can be useful for locating very generic contents on some area of the page.

Date – a text representing date. It is possible to specify which date format is possible to encounter in the specific case.

Here you can see an example of the "Static Text" properties dialog box.

Please note that the element description allows for some percentage of OCR errors, as well as some omitted words.

The "Search Constraints" page is used to define different relations between elements on the page.

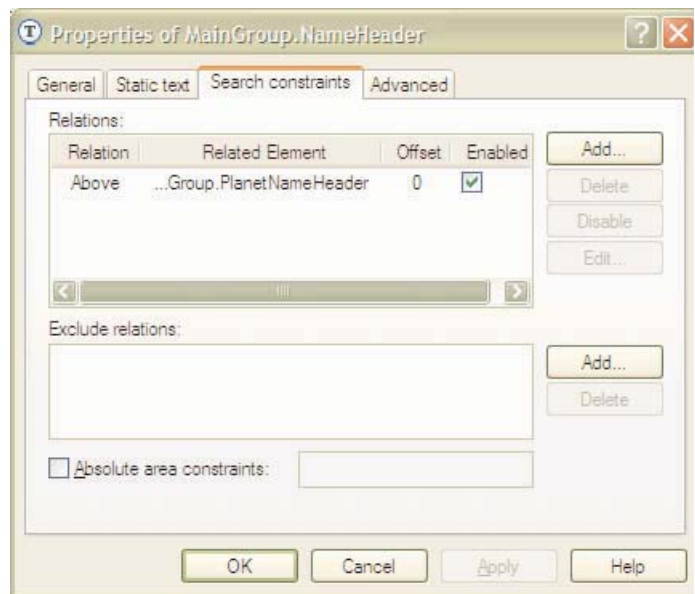
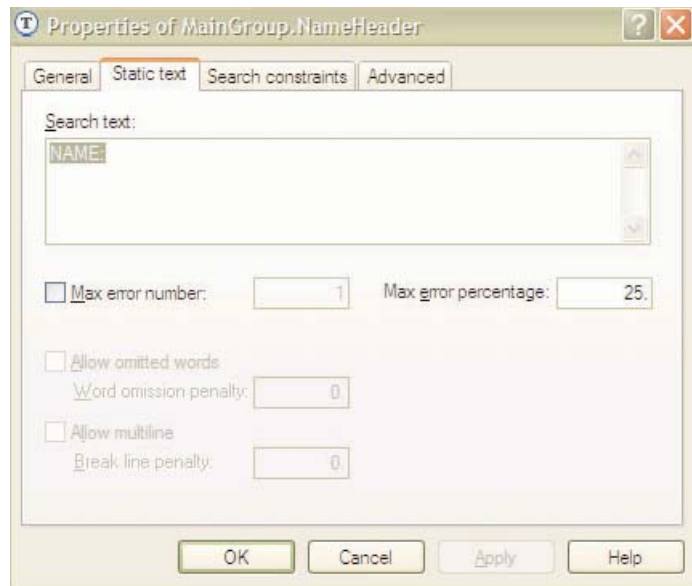
As you can see, the semantics of the relations is pretty straightforward. For instance, in this example, the relation states that the object that corresponds to the element `MainGroup.NameHeader` should be above the object corresponding to the element `MainGroup.PlanetNameHeader`.

Any such relation could be enabled or disabled, which makes for very convenient debugging.

The element can be defined elsewhere in the hierarchy. Elements can be grouped into compound elements, and one may easily move or copy elements from one place in the hierarchy to other using simple "drag&drop" technique.

When the first element is created, it is important to figure out how it is detected by FlexiCapture.

It is possible that additional relations have to be defined with other elements to improve the quality of detection. However one has to make sure that right properties are set so that even without relations to other elements the system generates reasonable hypotheses about the location of the corresponding object across different images.



Fine-tuning the FlexiLayout

During the process of fine-tuning the FlexiLayout, you must analyze the layout generated by the FlexiCapture Studio and if necessary, analyze the hypotheses tree to find out the reasons of possible mistakes. At the first stage most of the mistakes come from the variations of different documents in the batch. It turns out that the set of element properties and relations that worked perfectly for some subset of images do not take care of other images that have some differences from that subset. These problems can usually be fixed by reorganizing the element tree, modifying properties of elements and relations between elements.

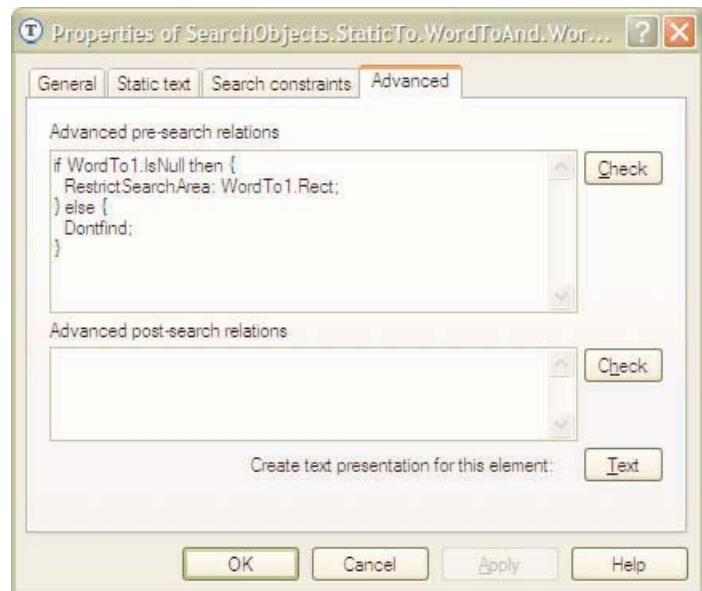
Of course when making modifications it is important to make sure that you do not spoil your current work by degrading the existing results. To preserve the integrity of existing results, you must save a "reference layout." For each page you may save the layout of the objects with which you are satisfied, and after making some modifications to the elements, you may trace the differences from the reference layout that appeared as a result of your modifications.

The **hypotheses tree** is another good tool used during the fine-tuning process. It visualizes the decision making process of the system. When a specific hypothesis in the tree is selected, a window shows the properties assigned to the hypothesis. Quality is one of the hypothesis' properties. The quality defines how well the object that is found on the image corresponds to the properties of the element. Qualities of the hypotheses in the chain are multiplied. The best chain in the hypothesis tree is the one having the best resulting quality. Modifications to element properties result in changes to the quality of hypothesis, thus defining which hypothesis will be finally selected by the system.

There is a good rule: the smaller the hypothesis tree, the better the FlexiLayout. A balanced approach to assigning quality property settings and a carefully designed FlexiLayout tree, with the idea in mind that the tree shall be as short as practically possible, not only reduces the effort required to fine-tune the FlexiLayout, but it also makes the whole template matching process much faster, as lesser variants have to be checked before the final decision is made. The ideal case is when only one chain of hypothesis is generated on every image, meaning that there is actually no choice for any element: all the objects are identified uniquely.

Advanced features

The FlexiCapture Studio user interface is designed to simplify FlexiLayout creation by directing the developer through a set of dialog boxes. In complicated cases requiring more detailed customisation and assistance, FlexiCapture Studio provides direct access to its internal structural language for greater flexibility and more control. Each properties dialog box contains an "Advanced" tab, where you may directly specify any additional relations or properties using the FlexiCapture structural language. For instance, in the example below the current element will be considered as not found in case the element WordTo1 exists.



Using the flexible document definition

ABBYY FlexiCapture Studio is available as companion product to the following ABBYY recognition products:

- ABBYY FormReader form processing solution
- ABBYY FineReader Engine Software Development Kit.

Both systems when purchased with a "FlexiCapture" enabled option, are able to recognize flexible forms using the FlexiLayout.

At least one runtime license of either product is required to process forms using a FlexiLayout. Existing customers should contact their local supplier about upgrading their current systems with a "FlexiCapture-Enabled" option and pricing information for FlexiCapture Studio.

ABBYY Software House
(Headquarters)
P.O. Box #54
Moscow, Russia, 129301
Tel.: +7 (095) 783 3700
Fax.: +7(095) 783 2663
office@abby.com

ABBYY USA
47221 Fremont Blvd.
Fremont, CA 94538, USA
Tel.: +1 510 226 6717
Fax: +1 510 226 6069
sales@abbyusa.com

ABBYY Europe GmbH
Anglerstrasse 6
Munich, Germany, 80339
Tel.: +49 (0) 89 5111 59 0
Fax: +49 (0) 89 51 11 59 59
info@abbyeu.com

ABBYY Ukraine
P.O. Box #23
Kyiv, Ukraine, 02002
Tel.: +380 44 490 9999
Fax: +380 44 495 2080
sales@abby.ua

ABBYY Europe GmbH, UK Office
3 South Mill Trading Centre
South Mill Road, Bishops
Stortford
CM23 3DY, England
Tel.: +44 (0)1279 323766
Fax: +44 (0)1279 323767
mail@abby.co.uk



© 2004 ABBYY Software. ABBYY, FineReader, FormReader, FlexiCapture and FlexiLayout are registered trade marks of ABBYY Software Ltd. All other trade marks are property of their respective owners.
ABBYY Software: P.O. Box 54, Moscow, 129301, Russia.
Tel.: +7 (095) 783 3700, Fax: +7 (095) 783 2663, office@abby.com; www.abby.com