ABBYY®

# Driving Impact From Unstructured Information Analytics

## Managing Deployment and Customization Complexity

WHITE PAPER

Driving Impact From Unstructured Information Analytics

## Contents



# Overcoming the Overwhelming

Over the past 25 years, organizations of all sizes have been implementing various kinds of solutions to action information, automate business processes and improve efficiency.

Many global success stories tell us how market leaders saved billions of dollars on implementing analytic solutions. McKinsey Global Institute estimated[1] several years ago that retailers exploiting data analytics could increase their operating margins by more than 60%; and that the US healthcare sector could reduce costs by 8% through data-analytics efficiency and quality improvements.

Along with Big Data analytics, interest in text (or content) analytics has grown significantly over the last few years. Organizations struggle to use digital content to improve business strategies and eliminate manual document processes.

**The ability to extract insights and intelligence from unstructured text to action information in business critical processes is now crucial for most organizations.**

In a recent AIIM survey[2] over 200 industry leaders provided insights on a series of topics and trends relating to content analytics. While interest in content analytics is high, progress remains mixed.

What are the three biggest drivers for content analytics in your organization?

| | |
|---|---|
| 63% | Improving process productivity by removing manual steps |
| 53% | Providing business insight |
| 45% | Adding value to our legacy content, improving search |
| 41% | Improving the benefits/compliance of our ECM/RM staff are poor at classification |
| 30% | Freeing up process bottlenecks and overloads |
| 25% | Reducing unidentified risk in our “dark data” |
| 24% | Reducing our storage/migration requirements in a defensible way |
| 17% | Detecting fraud, crime, policy infringement, unacceptable use, etc |

Figure 1 - *Content Analytics: automating processes and extracting knowledge*, AIIM Industry Watch, 2015

How would you best describe current progress in your organization towards the use of content analytics?

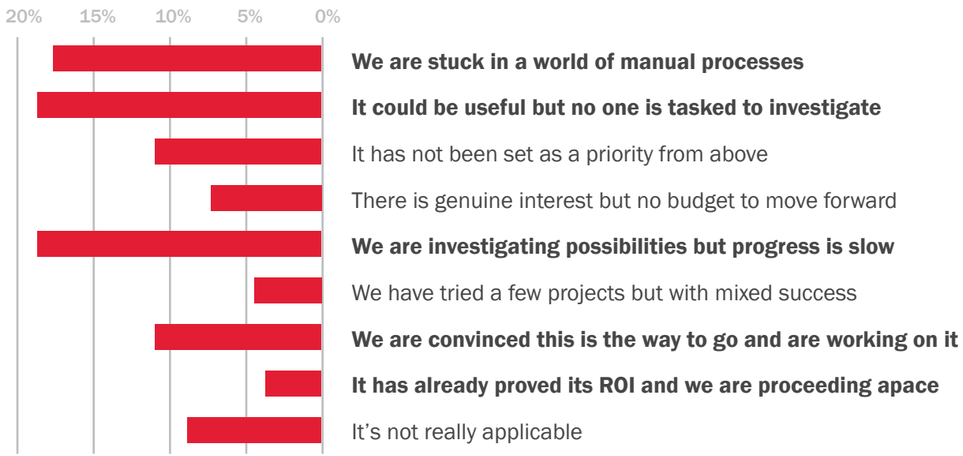



Figure 2 - *Content Analytics: automating processes and extracting knowledge*, AIIM Industry Watch, 2015

## Difficult and time-consuming system tuning, rules-setting and training lead to unpredictable results, costs and implementation timeframes.

**The complexity of setting the analysis rules, unpredictable results and other technological drawbacks often lead to low-quality data analysis.**

In some cases, additional system training or human verification can improve the situation. However, in most cases this wastes time and requires additional investments.

The AIIM survey also documented that companies lack employees with the required skills (lack of expertise tops the list of issues). This complicates both initial deployment and further maintenance.


Source - *Content Analytics: automating processes and extracting knowledge*, AIIM Industry Watch, 2015



**Both customization and maintenance require a specialist who will optimize information extraction. Each approach requires different types of specialists.**

## Efforts spent on customization and maintenance are the single most important factors influencing ROI.

**Hybrid systems** that use statistics (mostly with the use of classification) to train the system to detect entities are usually provided as "do it yourself" toolkits that qualified specialists use during customization and maintenance.

These specialists are usually programmers or linguists with the following skills:

- Excellent knowledge of regular expressions
- Good grasp of the product's "internal" language
- Sound grounding in linguistics
- Deep knowledge of the natural language for which rules are to be created

**Semantic solutions** that use linguistic analysis are usually customized by a solution integrator or the vendor. These solutions are powered by more complex technologies and require an engineer to tune them. But because they require far less customization efforts subsequent to deployment, there is no need for constant maintenance – in contrast to rule-based tools.

### The Customization Process

This process consists of two parts that can be laborious depending on the approach taken ("stages" refer to the table on the following page).

**Preparation work: ≈ 20% of time**

**The work is roughly the same for both approaches.**
*(stages 1-3, 5)*

**Information extraction rules development: ≈ up to 80% of time**

The cost (time and effort) of creating one rule is roughly the same, but the amount of rules required for accurate analysis differs considerably.
*(stage 4)*

### Cost of Maintenance

The cost of maintenance correlates to the number of rules created during customization – the more rules a product uses, the greater the error probability, and the more efforts are needed to support, update and refine these rules.

## Choosing the best approach: Hybrid vs Semantic

* REQUIRED
* DEPENDS ON THE CASE

| | Key Stages | Hybrid Approach | Semantic-based Approach |
|---|---|---|---|
| 1 | **Ontology Creation**<br>An ontology is a model of the data to be extracted. It includes entities, events, their properties, and relationships. | * A linguist or ontology engineer creates an ontology that includes entities, events, their properties, and relationships. | * A linguist or ontology engineer creates an ontology that includes entities, events, their properties, and relationships. |
| 2 | **Dictionary Creation**<br>2.1. The core vocabulary with common lexis.<br>2.2. The specialized vocabulary that contains industry, domain or company-specific keywords | * An engineer creates the dictionary with task-related keywords and synonyms. If predefined dictionaries exist, they have to be adjusted to the use case.<br>* An engineer must create dictionaries of specialized terms and their synonyms. | The core vocabulary is already included in the product, e.g. in the form of a semantic hierarchy.<br>* The vendor may provide industry-specific dictionaries, but company-specific dictionaries are created additionally by the vendor, integrator or customer. |
| 3 | **Syntax Rules Definition**<br>Rules for texts with specific structure are required: e.g., syntax of e-mails differs from agreements' syntax or company-specific keywords | Information extraction rules (see below) usually include syntax rules. | * Typically, a vendor supports specific types of text out of the box. If not, this work is done as a service. |
| 4 | **Defining Information Extraction Rules**<br>Information extraction rules allow the detection of valuable information in text (entities, events, relationships, and properties). For high quality results, rules must resolve natural language phenomena | **80% of customization efforts**<br>* 100*X (hundreds) of rules:<br>~10-20 rules per 1 valuable item (entities, events, etc.)<br>+ Hundreds of rules per item for disambiguation<br>+ 1-5 rules per each other language phenomena | * 100*X (dozens) of rules:<br>~1-5 rules per 1 valuable item (entities, events, etc.)<br>All other rules are unnecessary due to deep linguistic analysis of text. |
| 5 | **Quality Testing**<br>To tune and test the quality of analysis, a set of texts is collected and valuable data is manually marked up. This set is used as a model (master set) for further quality testing and debugging. | * In most cases, an engineer marks up large text corpora, tests and debugging rules manually. | * The vendor marks up texts, tests debugging information and extraction quality by employing his own tools. |



ABBYY

## Language Challenges and Improving Accuracy

Improving the quality of analysis is another crucial task that often is part of system tuning.

Enterprise information is typically text-heavy. Human language is very flexible – we use synonyms to express the same things while in other cases the same words to express different things (homonyms[iv]), we change the order of words in sentences, skip words when the context is clear (ellipsis[ii]), and more.

Overcoming the stumbling blocks inherent in natural language to improve accuracy (the precision and recall of extracted data) requires special rules. The number of these rules and their efficiency differ greatly depending on the approach.

Use Case: Information Extraction "XApp"

* REQUIRED

| Language Phenomenon | Hybrid Approach | Semantic-based Approach |
|---|---|---|
| **Word order alternatives**<br>I need information about features of the XApp product.<br>I need information about XApp product features. | * New case → 1-5 new rules<br>Each case needs particular rules for processing | The order of words does not affect the semantic structure of a sentence. |
| **New words added to the "original" sentence**<br>I need information about features of the XApp product.<br>I urgently need all the information you can provide about the new features of XApp. | * New case → 1-5 new rules<br>In some cases, the distance between words can be specified in the rule. | One rule for all variants will work well, because rules work with the semantic structure, not with words. The structure of a new sentence will be close to the structure of the original one. |
| **Synonyms**<br>Can you tell me about the XApp application pricing?<br>Can you give me some information about the XApp software licensing scheme? | * Each term needs a list of synonyms | Only domain or company-specific terms will need synonyms defined.<br>No synonyms for common lexis. |
| **Ambiguous words**<br>I have a question about the application installation.<br>I have a question about my mortgage application status. | * Hundreds of rules<br>A rule must be created for each specific case of lexical ambiguity (i.e. for a specific word in a specific context). Some products can resolve homonymy[iv] with the help of large text corpora statistics, others use rules. | Words are disambiguated during earlier syntactic and semantic analysis stages. |
| **Anaphora**[i]<br>XApp application. I want to know how to install it.<br>**Ellipsis**[ii]<br>One question is about the features of XApp, and another (question) is about the price.<br>**Co-reference**[iii]<br>Some time ago I installed XApp . I want to know how to enable printing in this application. | * Each case needs 1-5 new rules<br>Information extraction rules usually include syntax rules. | Typically, a vendor supports specific types of text out of the box. If not this work is done as a service. |
| **Semi-structured texts (tables, lists)**<br>Extraction of facts with all related data, e.g. to link Product XApp with the quantity "12":<br>Please send me quotes for the following products:<br>Product: XApp, Quantity: 12<br>Product: XGen, Quantity: 100 | Not applicable<br>Text structure (formatting) is ignored, so structures like these cannot be described by means of rules. | * Extra rules are required<br>Semantic products can extract information from tables, but additional rules must be created. |

## Choosing the best approach: Hybrid vs Semantic-based

Based on the preceding analysis the following recommendations can be used when choosing a technology for implementing unstructured content analysis and information extraction tasks:

### A hybrid-based information extraction product is best-suited when:

- A large text data set is available for testing and debugging the system.
- Input text is patterned in a way that can be described by rules. This works ideally if it is well structured (i.e. conforms to a rigid template) or when the input data is not natural language.
- The customer has a qualified specialist to create and maintain information extraction rules.
- Only a limited number of common entities, events and relationships need to be extracted, (e.g. names and surnames, company names, geographic objects, simple relations like ownership or employment). Complex "multi-event" facts will be unavailable for extraction.
- The data volumes are huge or the task requires real-time processing. Semantic-based products are more CPU-intensive than hybrid-based products.

### A semantic-based information extraction product should be chosen in the following cases:

**Semantic-based Case 1:**
Granularity and accuracy of analysis are more important than speed. In other words, the cost of mistakes is very high. Examples include:

- A company wants to optimize business processes and use information extraction to automate the capture of unstructured or semi-structured documents. Mistakes will require time-consuming and expensive manual verification that prevent the company from achieving cost and time savings objectives.
- When the quality of extraction is critical for a task, e.g. a contract review. The risk of missing any valuable detail can have serious, negative business consequences.

**Semantic-based Case 2:**
Complex facts with all related details (entities, events, relationships, properties) must be extracted. Semantic information extraction recognizes facts, composed of multiple entities, events and relationships. Each property of any of these entities, events and relationships must be related to the correct item. For example, a company needs to analyze an agreement and to identify a seller and a buyer with their individual properties, like name, ID number, address, etc. Semantic approaches identify the seller's properties and buyer's properties separately. Rule-based approaches can't provide this level of relational analysis, thus providing only disjointed information, like relationships between purchase entities "buyer" and "seller", and a set of individual entities, such as "ID number", "address", etc. At best, the nearest entities to a buyer or a seller will be related to them, but close standing entities are far from being related in real text.

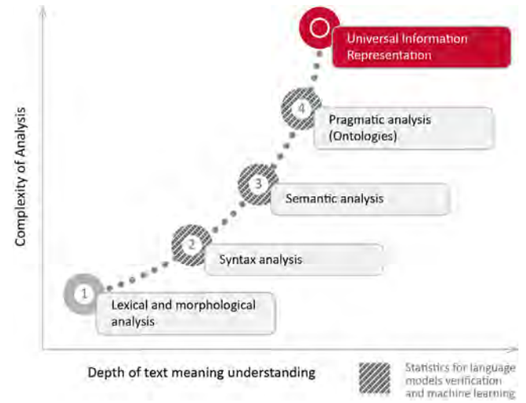


ABBYY

## ABBYY Compreno Solutions

ABBYY Compreno is a natural language processing technology. It is the basis for a new generation of intelligent analytic and discovery solutions to enable business to "understand" and effectively use unstructured information.

Powered by morphological, syntax and semantic analysis, ABBYY Compreno "understands" the meaning of words and reveals relationships between them. Compreno uses this "understanding" to create a semantic representation of text that can be efficiently utilized by a computer for accurate information extraction, classification, intelligent search and other text analytic tasks.

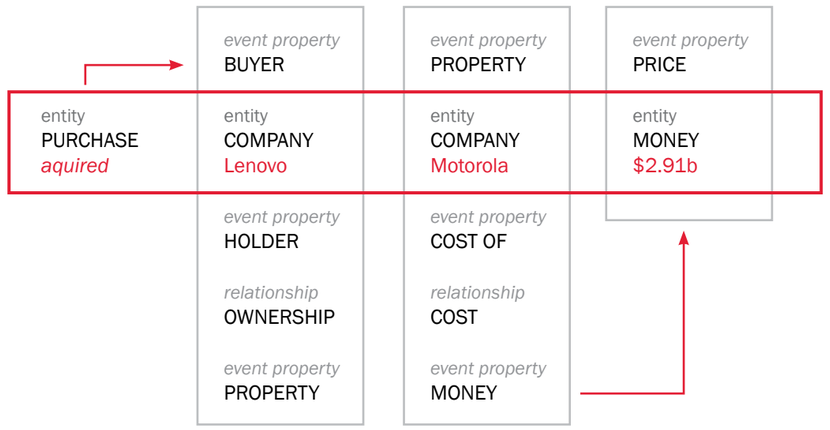



This language-based insight into unstructured information opens up new opportunities to action information and optimize critical business processes, including business process automation, case management, compliance, contract review and other information-intensive tasks that require granular content analytics.

ex

Remembering the entity list and narrative, **"Lenovo acquired Motorola for $2.91 billion."**-

Compreno constructs facts and relationships through the entities and their properties enabling a clearer vision of the story. The above statement can be analyzed as a **Fact of Acquisition**.



## Unique Compreno components

Among morphological analysis and the use of ontologies, which are commonly applied in text analytic solutions, ABBYY Compreno includes three unique components:

### Syntax

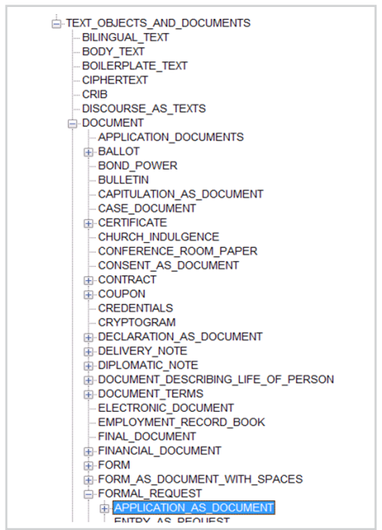
The syntax component detects how words (concepts, to be more precise) are related to one another within one or more sentences. The system analyzes texts and builds a tree of syntactic relations. To make syntactic parsing more accurate, ABBYY Compreno also relies on semantic analysis that uses the hierarchy of concepts described below.

### Semantics

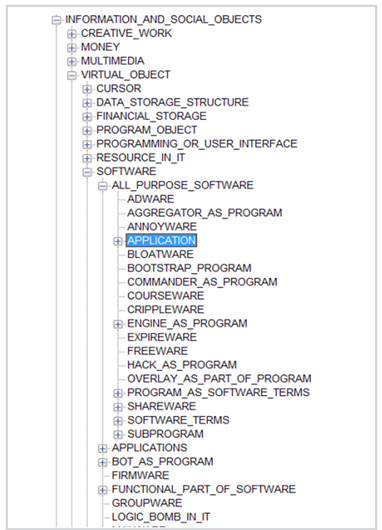
This component is presented in the form of a Universal Semantic Hierarchy (USH) of concepts. Semantic analysis is used to interpret syntactic structures in terms of hierarchy concepts and their semantic relationships and roles.

The USH brings many advantages to Compreno. It helps to resolve ambiguities during analysis. For example, in the illustration below, the notion Application in the meaning of "software" can be easily differentiated from Application in the meaning of a "document".

**USH "Application" for "Document"**




**USH "Application" for "Software"**




**The unique Universal Semantic Hierarchy (USH), developed by ABBYY, provides Compreno with knowledge about the real world.**

**This semantic hierarchy is a tree-like structure of notions organized from general to increasingly specific concepts.**

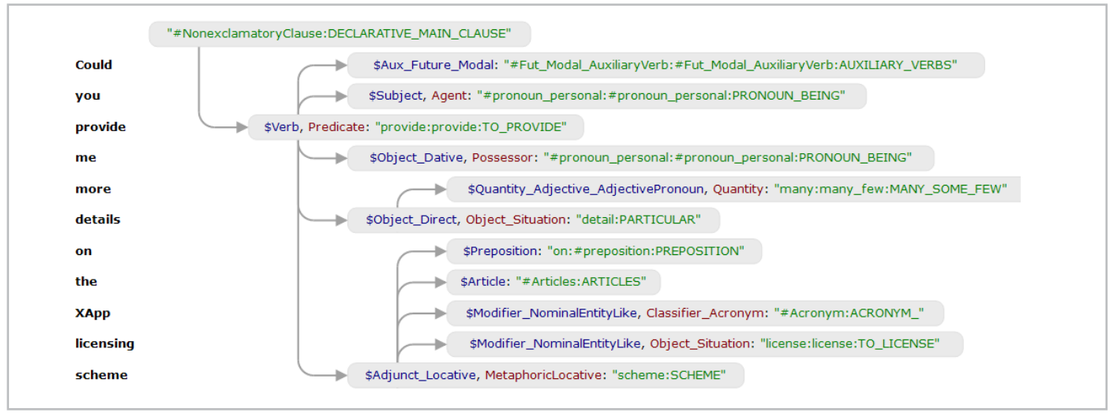
Establishing semantic relationships between meanings within a sentence means, for example, that the system will identify the relations between the verb that means "information request" and its properties — who asks, who is asked, when, where, etc.



Joint use of syntax and semantics enables the system to "understand" the meaning of each word in a sentence.

| I | have | a question | about XApp | application | features |
|---|---|---|---|---|---|
| #pronoun_personal<br>PRONOUN_BEING<br>BEING<br>PHYSICAL_OBJECT<br>ENTITY | have: have<br>EXISTENCE_AND_POSSESSION<br>SITUATION<br>SITUATIONAL_AND_ATTRIBUTIVE_CLASSES<br>LEXICAL_ELEMENTS | question: question<br>QUESTION<br>QUESTION_ANSWER<br>RESULTS_OF_GIVING_INFORMATION_AND_SPEECH_ACTIVITY<br>RESULTS_OF_SPEECH_MENTAL_ACTIVITY | #Acronym: ACRONYM<br>UNKNOWN_SUBSTANTIVE<br>ENTITY_OR_SITUATION_PRONOUN<br>LEXICAL_ELEMENTS | application: application<br>APPLICATION<br>ALL_PURPOSE_SOFTWARE<br>SOFTWARE<br>VIRTUAL_OBJECT | feature: PARTICULAR_FEATURE<br>FEATURES<br>CHARACTERISTIC_GENERAL<br>CHARACTERISTIC_AND_VALUE |

It also detects relationships between meanings and creates a universal semantic structure for a text, which can then be efficiently analyzed by a computer.




### Statistics (machine learning)

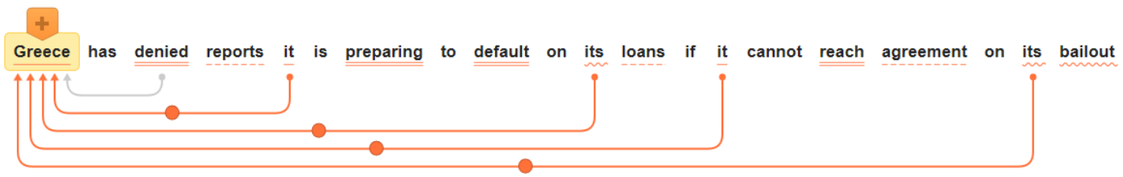
ABBYY Compreno uses statistical data (gleaned from large text corpora) to support syntactic and semantic analysis algorithms - strengthening them while verifying and expanding the formal descriptions available in the system. The statistical component uses texts of different genres and styles to reduce the likelihood of error and misinterpretation, for example, to increase the quality of resolving complex cases of ambiguity and homonymy[iv].

## Technological Advantages

Compreno's built-in capabilities for analyzing texts in natural languages avoid the common challenges of complex customization and poor quality. Deep language-based analysis resolves many complex language phenomena automatically, thus radically reducing the need for customers to describe rules or to train the system. The widespread language phenomena that Compreno resolves includes:

## Relationships detection, including distant and "hidden" relationships

ABBYY Compreno can detect all relationships among words in a sentence and among sentences, including relations between words that stand far apart. The technology also detects words that refer back to an already mentioned object, e.g. by using pronouns it, he, she, etc. (anaphora[i]) or that use other words such as product, company, etc. (co-reference[iii]).

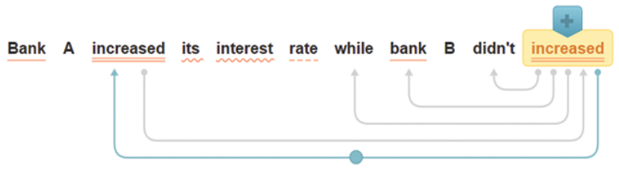



## Word sense disambiguation

ABBYY Compreno detects homonyms[iv] and selects the right meaning by analyzing the surrounding context. It also resolves both grammatical homonyms (same wording, different part of speech: steal – to steal) and lexical homonyms (same wording, same part of speech, different meaning: run = execute, run - move).


## Omitted words restoring (ellipsis[ii])

In natural language, we often skip words that another person might glean from their context. Machines can't read our thoughts, but in some cases it is possible to recover skipped words by implementing deep linguistic analysis. ABBYY Compreno is able to recover words that were omitted, which helps to access information even when it's partially hidden.




Only Compreno's ability to understand relationships, events and facts to grasp the story lines in documents delivers data critical for text-heavy, knowledge dependent activities such as business process automation, case management and information governance.



### References and links

[1] McKinsey Global Institute, Report "Big data: The next frontier for innovation, competition, and productivity", May 2011

[2] AIIM,"Content Analytics: automating processes and extracting knowledge", June 2015

### Definitions

[i] **Anaphora** – the use of an expression the interpretation of which depends upon another expression in context (its antecedent or postcedent). In the sentence "Sally arrived, but nobody saw her", the pronoun "her" is anaphoric, referring back to Sally.

[ii] **Ellipsis** – the omission of words that can be recovered from the context. E.g. "I like swimming, but my brother doesn't". Words "like swimming" after doesn't are omitted.

[iii] **Co-reference** – occurs when two or more expressions in a text refer to the same person or thing; they have the same referent, e.g. "Some time ago I bought XApp. I want to know how to install this application".

[iv] **Homonymy / Homonyms** – the words which are identical phonetically or graphically but which have essential difference in meanings. E.g. application [a text document] and application [a software].

ABBYY®

**ABBYY International Headquarters**
Otradnaya str. 2b/6 127273
Moscow, Russia

Tel: +7 495 783 3700
Fax: +7 495 783 2663

Email: office@abbyy.com

**ABBYY European Headquarters**
Elsenheimerstrasse 49, 80687
Munich, Germany

Tel: +49 89 69 33 33 0
Fax: +49 89 69 33 33 300

Email: sales_eu@abbyy.com

**ABBYY Australia**
Citigroup Building, level 13, 2 Park Street
Sydney, NSW, 2000, Australia

Tel: +61 (02) 9004 7401

Email: sales@abbyy.com.au

**ABBYY North American Headquarters**
880 North McCarthy Blvd., Suite #220
Milpitas, California 95035, USA

Tel: +1 408 457 9777
Fax: +1 510 226 6069

Email: sales@abbyyusa.com

**ABBYY 3A (Asia, Africa, South America)**
Otradnaya str. 2b/6 127273
Moscow, Russia

Tel: +7 495 783 3700
Fax: +7 495 783 2663

Email: sales_3A@abbyy.com