Comprendium Translator
System Overview

May 2004

Table of Contents

1. Introduction

Globalization is changing the face of business and with it, come increasing challenges associated with understanding. One of the main problems of globalization is how to overcome the very real language barriers, which divide people around the world. In practice, no one language will ever be universally accepted as the sole vehicle of international communication. The only way forward for global players is to be capable of understanding information and providing content in all the native languages of the countries in which they operate: for multinational customers, work forces, and business partners.

In many applications speed is more essential than polished language: for internal communication within multinational companies and the rapid exploitation of information for business research and monitoring it is often only necessary to understand the content of documents, and having to wait hours or even days for a human translation can lead to a loss of a vital competitive edge.

For company brochures and marketing information top-quality translation is a must - and human translators will always be involved in the production process. But even here, translation tools and processes can help users achieve a quicker time-to- market, reduce costs and relieve the load on translation departments and agencies.

2. What is Machine Translation?

Machine Translation is a technology, belonging to the computational linguistics field that allows computer programs to translate texts from one language to another. Machine translation is an application of computational linguistics whose aim is to get a computer system to translate. This objective is achieved with the integration of a number of natural language processing techniques.

From the human perspective, language is an everyday matter of fact that even very young children can master without any problem. However, the handling of human language in general, and the translation between languages in particular, is one of the most difficult tasks that can be given to a computer program.

Soon after the first machine translation attempts, more than four decades ago, where native word-to-word translation approaches were used, it was clear that the problem required the use of much deeper linguistic devices. Human language is not a mere sequence of individual unconnected words; words group themselves into phrases, and phrases build sentences. Finding out the right internal structure of sentences is crucial if we want to be able to translate it into another language.

3. The Comprendium Machine Translation Technology

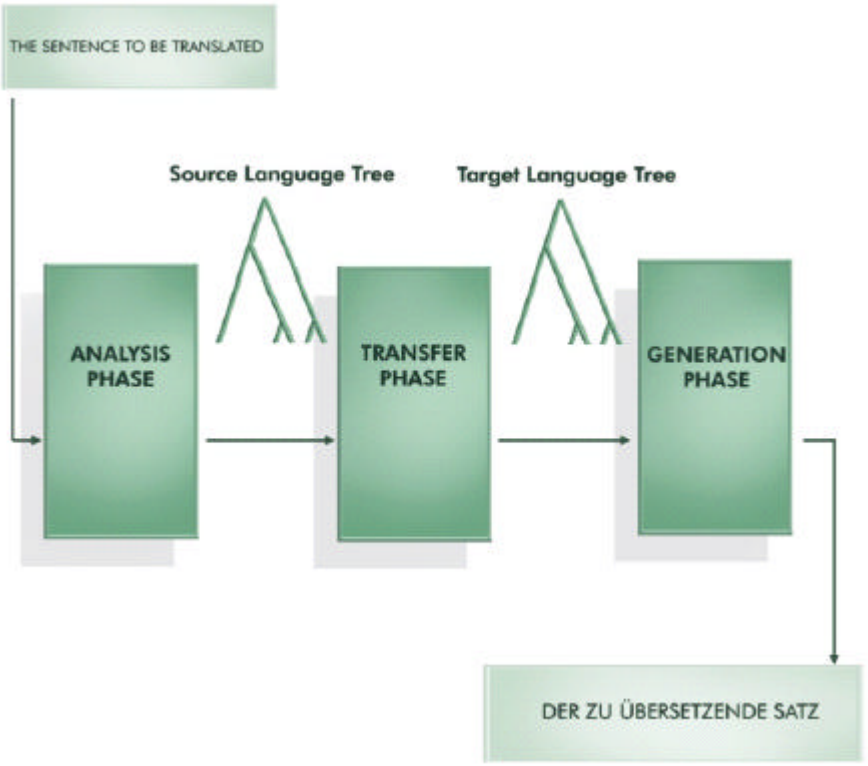
3.1 The Best MT Technology in the Market

The Comprendium Machine Translation system uses one of the most highly developed linguistic approaches currently available. More than 20 years of research and millions of development dollars have gone into the development of this system, which carries out a full linguistic analysis of each sentence before translating them it into the target language. This system has extensive grammars and powerful dictionaries, and is light-years removed from previous word-to-word translation approaches.

The Comprendium MT system yields the best translation quality that can be found today in the market. Translation quality depends on a number of factors (language-pair, type of document, etc.) and is not always easy to measure. One big advantage of the Comprendium MT system is that it can be linguistically tuned to the specific needs of the customer.

3.2 A Transfer Approach

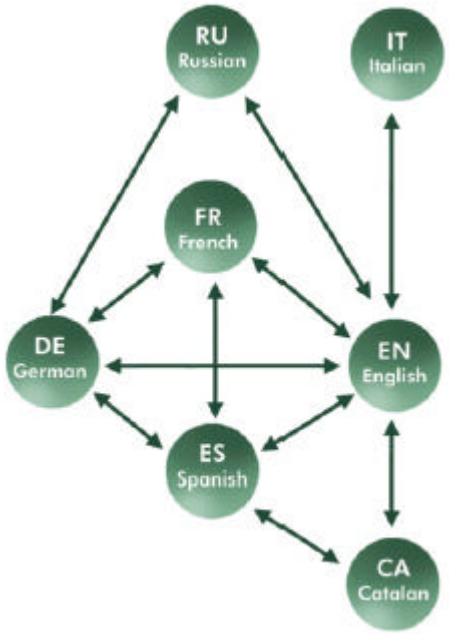
The Comprendium MT system follows the paradigm of transfer-based systems, which have proven to be the only working engineering approach. It uses a tree-to-tree mapping approach. First, an analysis tree is created. This is then transformed into a transfer tree where the basic translation operations are done; finally, the transfer tree is converted into a target language generation tree, which then delivers the output.



Moreover, this approach makes existing analysis and generation components reusable for new language pairs and thus guarantees maximum efficiency and quality standards across different language combinations.

3.3 A Wide Range of Language-Pairs

The Comprendium MT system currently uses a wide range of available language-pairs as is demonstrated below and new ones are under continuous development:



- English <-> German
- English <-> French
- English <-> Spanish
- English <-> Russian
- English <-> Catalan
- English -> Italian
- German <-> French
- German <-> Russian
- German <-> Spanish
- French <-> Spanish
- Spanish <-> Catalan

3.4 Translation Memories

Translation Memory is a technology that leverages the results of previous human translations. It can be used in combination with the MT system. In areas such as software localisation (manuals, process descriptions), many documents are simply modified versions of previously published material that has already been translated into various languages. Translation memory technology stores source documents and their translations on a sentence-by-sentence basis. When a new version of a document is produced, this version is first compared against previously stored versions in order to ascertain what had already been translated. If the identical text is found, this is then re-used for the new version of the document. So-called “fuzzy matching” also makes it possible to find sentences that are very close but not quite identical. Use of Translation Memory means that only the new text in a document needs to be translated and that the retrieved text is of an accepted high quality. This represents considerable savings in time and money.

Translation Memory is best used in combination with Machine Translation: sentences are first looked for in the Translation Memory, the remaining text is sent for re-translation by the MT system.

4. How Does the Comprendium MT System Work?

4.1 The MT System

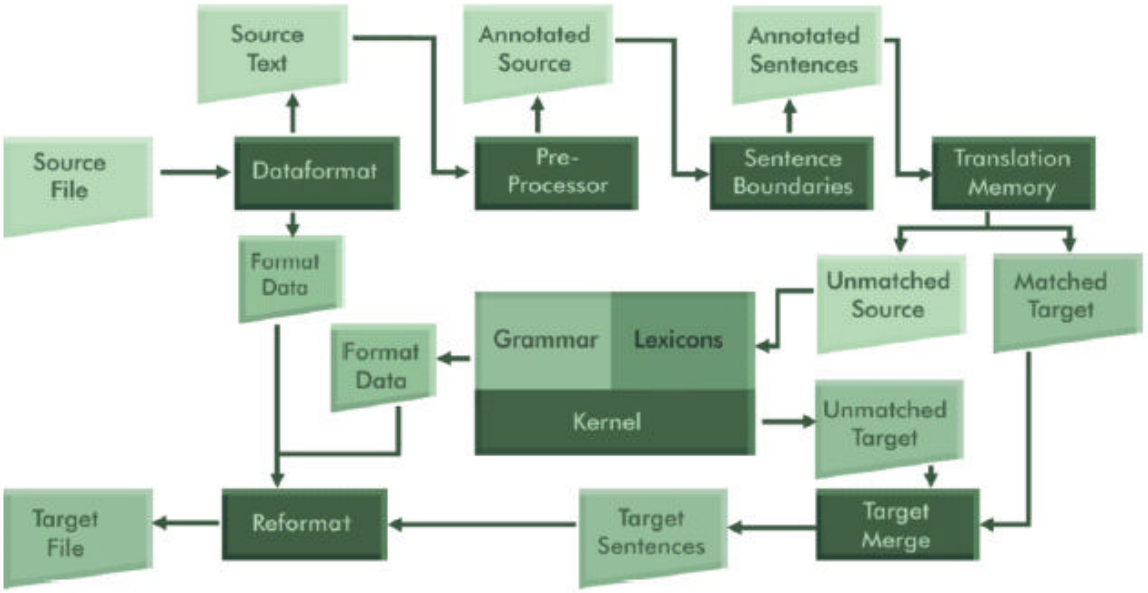
The workflow is as follows:

- Users start translation requests either from their Internet browsers, using a web-based client or from the stand-alone Java application on their PC.
- The translation environment is accessed by the mechanism, which knows which translation engines are on which machine, and activates the respective translation engine.
- The translation engines execute the translation jobs proper, and returns the translated documents. These documents are handed back to the user clients.

For administration, two software components are available:

- There is a component, called LexShop, to administer the lexicon resources of the system. Lexicons change with each application and over time (as new terms are coined, old terms “die” together with the products they describe, like *card punching machines*), so special administration is offered to keep lexicons up-to-date
- From a Comprendium Translator Desktop Power client, the expert user is able to create and administer the translation memory modules.

4.2 Workflow during translation



The translation process starts with a source language file, and then continues in several phases:

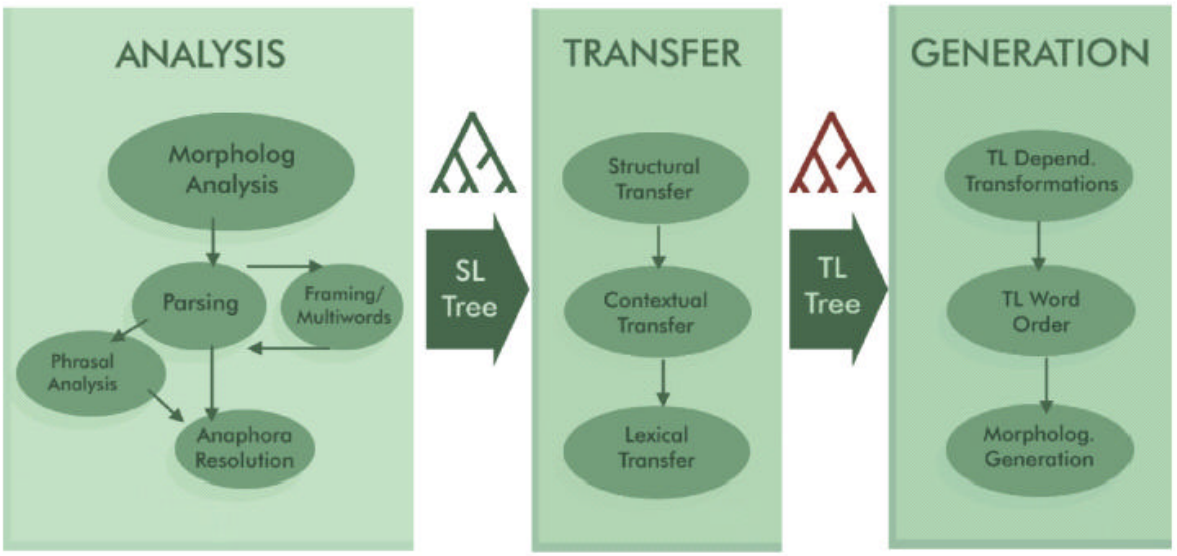
- First, the text is unformatted. The text portions are separated from the layout information and the layout format is stored. This is done because, in the end, the translated text needs to have the same layout as the original document. Creating a new layout would take more time than the translation itself.
- The source text is sent to a pre-processor. This identifies parts that should not be translated (names, email addresses, filenames etc.); these parts are marked
- Then the text is split into sentences because most translation tools work on a sentential basis, and so does the MT core engine
- Next, a translation memory is consulted if available. This can be seen as a large database of sentences where for each source language sentence, its target language equivalent is stored, resulting from previous translation efforts. A memory is a good supportive tool in case there are repetitive texts. If a sentence is found in the memory, it need not be sent to the MT engine.
- If a sentence is not in the memory it is sent to the MT engine, consisting of a software kernel that drives the translation. The translation direction is determined by the linguistic resources (grammars and lexicons) that are loaded into the kernel. The MT engine translates on a sentence basis. The MT engine also updates some elements of the format, e.g. if one boldface word is translated as a phrase in the target language, or is moved to a different position in the target sentence.
- The output of the MT engine (i.e. the target language sentences) is merged with the results of the Translation Memory lookup, so that all document sentences have been translated into the target language.

- The last step is to merge the target language text with the stored layout and to reformat the whole document. This way, the target file is created.

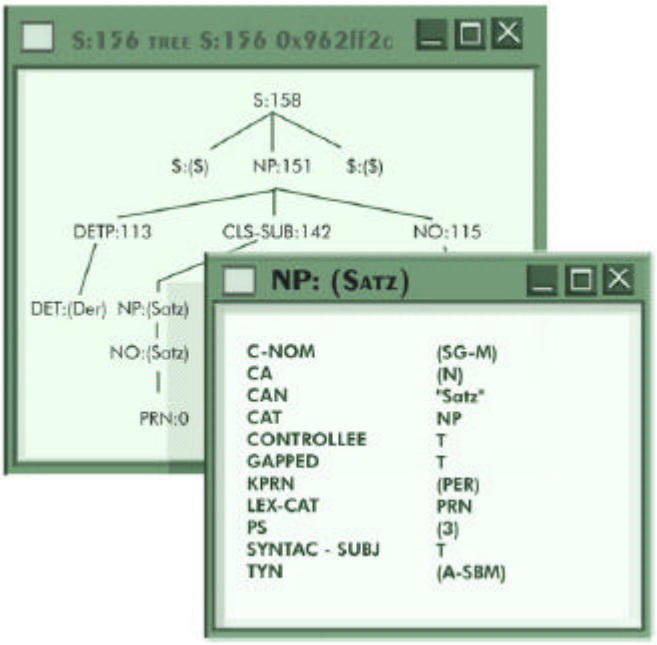
4.3 MT engine internals

4.3.1 The Translation Phases

As we have mentioned above, Comprendium's MT system uses the transfer approach, which means that it uses the analysis transfer generation paradigm.



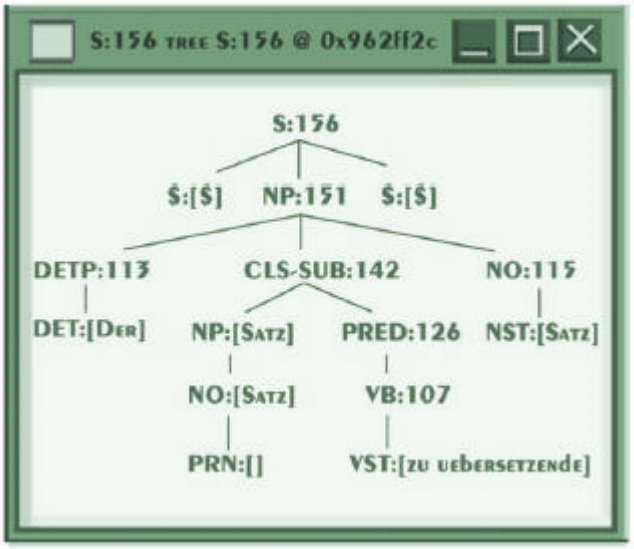
Source and target language trees are representations of the syntactic structure of the sentence being translated. At the same time, trees contain linguistic information in form of feature-value pairs:



4.3.1.1 4.3.1.1 Analysis Phase

The analysis phase consists of five main packages:

- First, morphological analysis is done. Words are looked up in the lexicon, and their specific position in the sentence is interpreted.
- Second, syntactic analysis is done, also called parsing. The system fires a series of grammar rules which create phrases, and combine phrases into sentence parts, and finally into an overall interpretation. The main challenge here is syntactic ambiguity. In sentences like “Paul saw the man with the telescope”, the telescope-phrase can refer to Paul or to the man, but in a similar sentence “Paul saw the man with the children”, the children-phrase can only refer to the man (you can use a telescope to see something, but you cannot use children). Disambiguation is the main task of the syntactic parsing.
- Part of the syntactic parsing is framing and multiword analysis. Framing means to assign certain phrases a grammatical role (like subject or indirect object), multiword analysis interprets groups of words (such as blinder Passagier in German) to be one term, and to be translated in a specific way (into a single word in English: stowaway).
- If the syntactic analysis fails (because the input sentence is too complex or incorrect), a special mechanism (called phrasal analysis) is triggered to try to identify meaningful phrases in the input.
- Finally, there are phenomena that need more than a sentential context, like pronouns referring to some previous sentence. These anaphora need to be resolved. The result is a source language analysis tree per sentence.

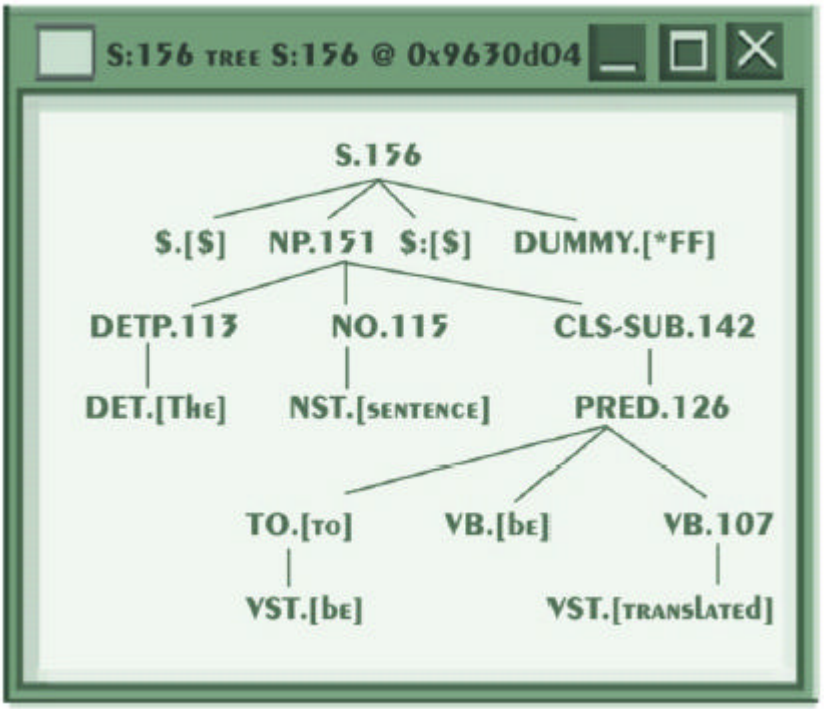


4.3.1.2 4.3.1.2 The Transfer Phase

The transfer phase consists of three steps:

- First structural transfer is performed. This is transfer that is independent of lexical material. Complex pronominal German phrases including present participles will be translated into English subordinate infinitive clauses, independent of the words: der zu uebersetzende Satz -> the sentence to be translated.
- Then, contextual transfer is done, which involved words that influence their context: English like -> German gefallen, but the roles must be changed: he likes her -> sie gefaellt ihm: The English subject becomes the German indirect object.
- Finally, simple lexical transfer is done, replacing a source language by a target language term (German Atomkraftwerk -> English nuclear power plant).

The result is a target language tree:



4.3.1.3 4.3.1.3 The Generation Phase

The generation phase needs to perform the following steps:

- First, some target language specific re-arrangement of the tree are carried out.
- Then, the target language specific word order is created (verb into final position, subject before the verb, etc.)
- Finally, the target words are inflected, and the correct word forms are created (*dog+plural -> dogs*, *go+past-> went*, etc.).

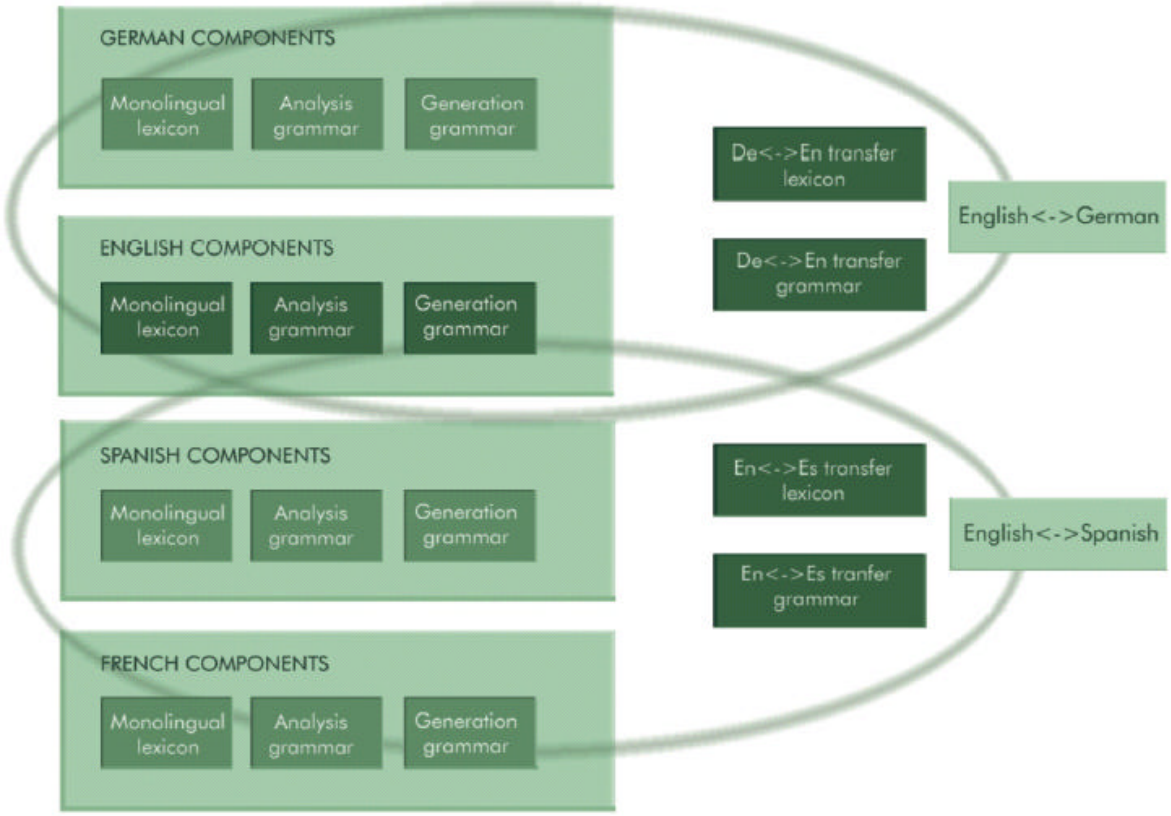
4.3.2 The MT System Building Blocks

The two main linguistic building blocks of the MT System are the lexicons and the grammars.

For any specific language-pair direction (e.g. German-English) the Comprendium MT System uses two monolingual lexicons (one for the source language German - and other for the target language English) and one transfer bilingual lexicon. Monolingual lexicons contain relevant morphological, syntactic and semantic information for every word known to the system. Bilingual lexicons contain all the source-to-target translations for a given word, depending on the syntactic and/or semantic context.

At the same time, the system uses one analysis grammar to carry out the syntactic parsing of source sentences, one transfer grammar to map the source analysis trees into the corresponding target trees and one generation grammar, used to generate the output sentences in the target language.

All these components are re-usable among different language-pairs. Thus, the German-English system and the German-Spanish systems share the same German monolingual lexicon and the same German analysis grammar.



5. Summary

Comprendium Translator offers all the outstanding state-of-the-art machine translation technology described above plus a number of unique characteristics that no other system can deliver.

- Ease of use.
- High-quality output.
- A wide range of language-pairs available.
- Linguistically customizable.
- High performance and scalability.
- A choice of complementary translation technologies.
- Possibility of integration into workflows.

Comprendium Translator is the perfect tool to overcome language barriers within the enterprise.

ABOUT COMPRENDIUM

Comprendium is a leading European provider of enterprise document management & multilingual content integration solutions. With its InfoStore product family, Comprendium not only covers classical application areas such as electronic archiving and document management, but with its multilingual content integration framework, it also enables the combining of documents, content and information regardless of language, company location, or system platform, thereby leveraging an organizations information assets and technology investments.

Comprendium has a single focus: delivering solutions to enable customers' to meet business objectives, with measurable benefits. Whether you are a small or medium sized business with a straightforward requirement to archive and manage your business documents to meet efficiency or compliance requirements, or a large-scale enterprise with a legacy of complex content dependant systems and a growing need to accelerate your response to new business opportunities, Comprendium can support these challenges. Proving the effectiveness of Compendiums solutions are the more than 800 customers across Europe.

Comprendium Deutschland
Balanstrasse 57
81541 München - Germany
Tel. +49 89 24 44 33 0
Fax +49 89 24 44 33 120

Comprendium Austria
Praterstrasse 38
1020 Wien - Austria
Tel. +43 1 596 70 60
Fax +43 1 596 70 60 18

Comprendium Schweiz
Bahnhofstrasse 21
9471 Buchs - Switzerland
Tel. +41 81 755 55 00
Fax +41 81 755 55 01

Comprendium France
3, rue de l'Arrivée, BP 44
75749 Paris Cédex 15 - France
Tel. +33 1 45 38 76 16
Fax +33 1 45 38 58 02

Comprendium UK
40 Portman Square
London W1H 6 LT - United Kingdom
Tel. +44 (0)20 7947 8777
Fax +44 (0)20 7947 8801

Comprendium ESPAÑA
C/ Balmes, 114, 5è
08008 Barcelona - Spain
Tel. +34 93 492 01 60
Fax +34 93 492 01 61