

Computer Aided Translation in an Integrated Document Production Process

Tools and Applications

Gerhard P. Freibott

Krupp Industrietechnik GmbH, Zentrale Informationsdienste
Franz-Schubert-Str. 1-3, D-4100 Duisburg 14

Abstract

The internationalisation of markets, the ever shortening life cycles of products as well as the increasing importance of information technology all demand a change in technical equipment, the software used on it and the organisational structures and processes in our working environment.

Translation as a whole, but in particular as an integral part of the document production process, has to cope with these changes and with new and additional requirements.

This paper describes the organisational and technical solutions developed and implemented in an industrial company for a number of computer aided translation applications integrated in the document production process to meet these requirements and to ensure high-quality mono and multilingual documentation on restricted budgetary grounds.

TABLE OF CONTENTS

- 1. INTRODUCTION**
- 2. GENERAL SITUATION**
- 3. ANALYSIS OF THE EXISTING ENVIRONMENT**
- 4. THE ROLE OF TRANSLATION IN DOCUMENT PRODUCTION**
- 5. THE SYSTEM OF TOOLS AND APPLICATIONS**
- 6. ECONOMICAL AND SOCIAL ASPECTS**
- 7. CONCLUSION**

INTRODUCTION

The precise nature of changes is one of the most difficult things to identify in the business world, particularly when one's own organisation is itself subject to major changes as has been and still is the case at Krupp Industrietechnik. The factors of change are both general and specific: global economic and political forces, mergers and acquisitions, new developments in technology, new market conditions and legal regulation.

Office communication and automation is currently subject to constant and rapid change with all its implications for the translation business. Apart from the technological changes and transformations, we see ourselves confronted with developments that are taking place parallel to the evolution of world-wide commercial change and the changing patterns of managerial work and expectations in our business.

The changes and transactions of EDP applications that took place in so many fields of a company's working environment, with significant increases in effectivity and productivity must also be introduced in language processing as part of the overall activities of a company's internal services.

Indeed, it is undeniable that computer based applications in office communication and specifically in language processing haven't had the anticipated economic effect. To put it in the words of Lester C. Tretow, the Director of the MIT School of Management: "We do not have a general productivity problem but a productivity problem in the administrative sector. Our factories function well, the office does not"

In my opinion, there are two major reasons for this:

☐ Lack of appropriate NLP-tools (Natural Language Processing Tools).

In office communication, apart from the organisation of the technical operations of text and document production and of information management, we are dealing with a phenomenon which cannot be treated with traditional methods of electronic data processing. Here we are dealing with language, natural language, and the electronic processing of natural language obviously needs special treatment and follows different rules.

☐ Lack of standards and exchange facilities.

These facilities are needed to allow the gathering, processing and distribution of information on different hardware levels and in different software environments.

GENERAL SITUATION

Translation within the document production environment today has to cope with new and additional requirements.

Due to a change in the judicial basis caused by new product liability laws in the EC countries, the EC-guideline "Machines", the EC Conformity Certificate for the import and export of goods into EC countries and now other European countries as well, it is necessary to deliver the documentation accompanying the product in the official language/languages of the individual country of destination and use.

Documentation is expected to be more user-friendly. The user-oriented approach understands documentation and in particular multilingual documentation as part of a company's public relations, image improving and acquisition activities.

The requirements to be met in the production of mono and multilingual documentation as a high-quality end product have increased immensely. It has become normal practice to combine the judgement of the quality of the product with the judgement of the quality of the accompanying documentation and in particular with the judgement of the quality of the translation.

Translators certainly welcome this development but must try to live up to these requirements.

The success of an industrial product today is only conceivable in connection with qualitatively high value services. The main product in many branches is no longer the hardware itself, but high quality service in the form of know-how transfer, transfer of knowledge about construction, production, operation and use of the product. This knowledge is presented in different forms which to a very high extent assist language arrangements. Documentation is first to be mentioned here. Documents in paper form, in electronic form, as graphics or moving pictures, as multi-media application etc.

The problem that we are faced with, is that the requirements for documentation have quantitatively and qualitatively risen and that this increase in quantity and quality is to be realised without significantly higher costs, without overstressing existing budgets.

To cope with this situation, we need new effective tools and methods, new organisational structures to fully exploit the possibilities of technology and to realise the undoubtedly existing synergy effects.

General Situation	Requirements	Applications
<ul style="list-style-type: none"> ☐ Rapid and constant growth in volume 	<ul style="list-style-type: none"> ☐ Higher output without significantly higher cost 	<ul style="list-style-type: none"> ☐ Use of information and language technology
	<ul style="list-style-type: none"> • Better quality regarding both content and form 	<ul style="list-style-type: none"> • Direct and immediate access to and retrieval of:
	<ul style="list-style-type: none"> • Shorter production times 	<ul style="list-style-type: none"> - standardised modules (pre-processed, pre-formatted) via descriptors of automatic matching,
	<ul style="list-style-type: none"> - adaptability to market requirements 	<ul style="list-style-type: none"> - interactive and automatic matching of words, compounds, collocations, sentences, parts of speech
	<ul style="list-style-type: none"> - shorter life-time cycles 	<ul style="list-style-type: none"> - linguistic, conceptual, terminological, pictorial information
	<ul style="list-style-type: none"> • Language independent processing 	<ul style="list-style-type: none"> • Direct storage of reusable parts of translated text and documents
		<ul style="list-style-type: none"> • Definitions and explanations of terms and hints regarding possible homographs, synonymity etc.

Although technology plays a major role in all these applications, document production and in particular foreign language document production cannot be only a question of technology. It is also a question of establishing and optimising organisational and functional chains of individual operations, a question of qualified project management, computerised document management and information management

Organisational planning has a key function in the successful use of a document management system. Workflow patterns that begin even before a single word has been translated and committed to paper must be analysed and streamlined.

ANALYSIS OF THE EXISTING ENVIRONMENT

It cannot be disputed that a tremendous change in the working environment has taken place during the last few years due to the introduction of EDP-applications in the language business and that there are quite a number of tools available on the market to increase productivity and quality. The information technology world market and with it the language technology market are currently enjoying a great amount of dynamism. We are witnessing an enormous growth in computer-linguistic tools and applications. But what we also expect are economically successful products that will help to assure that the growing complexity of knowledge can be appropriately refined and presented.

As became evident from a number of market studies and the analysis of the existing working environment at Krupp Industrietechnik, the majority of translators prefer to use very basic and traditional translation aids which are obviously sufficient to meet their demand for information in the translation process:

- terminological database in glossary form integrated into the word processing environment
- "cut and paste" functions
- very basic data exchange utilities
- spell checkers and hyphenation features based on simple processing algorithms.

This all comes to the surprise and in contrast to the imaginative power of a highly ambitious information technology community.

For translators, automation of the translation process seldom seems to go beyond these features. There is one exception, however: access to and extraction of information from parts of previously translated text, i.e. information extracted from whole parts of text.

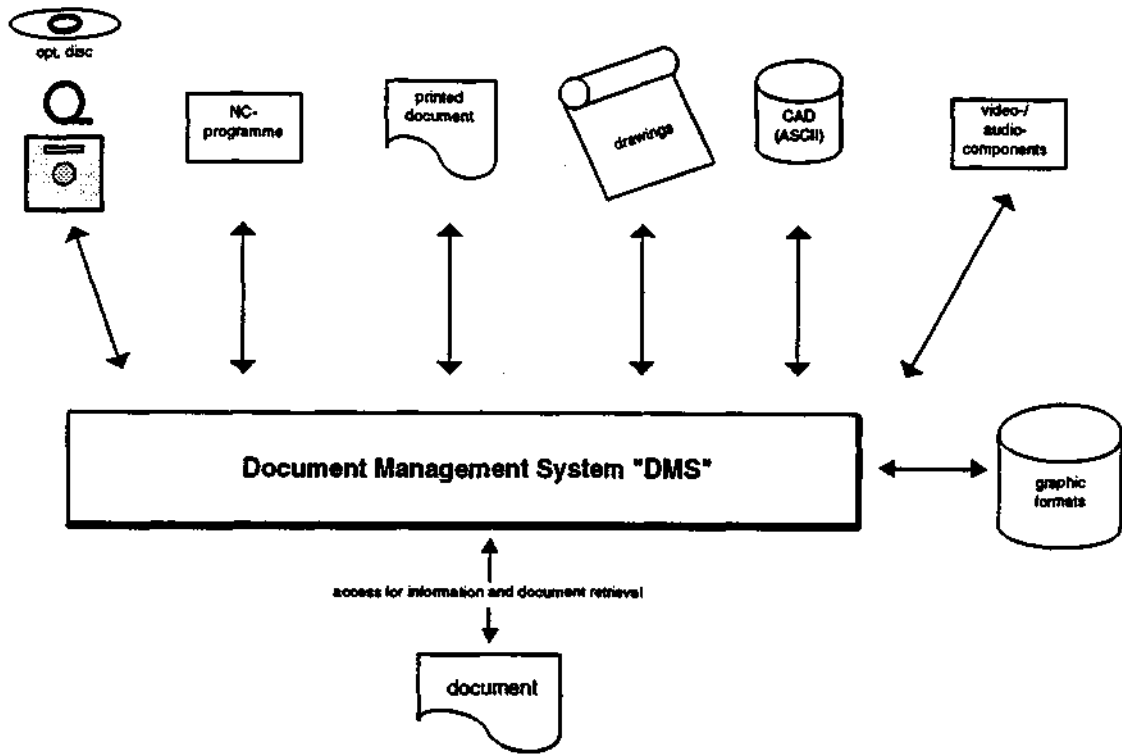
Having realised this, an attempt was made to find out what could be improved in the strategic planning of the development of the office communication and automation environment. The analysis of the existing situation showed the following deficits in the factual organisation:

- lack of effective search and retrieval functions,
- operational and organisational fractures due to heterogeneous hardware platforms and software environments,
- lack of integration in the text generation, text processing, translation and document production processes.

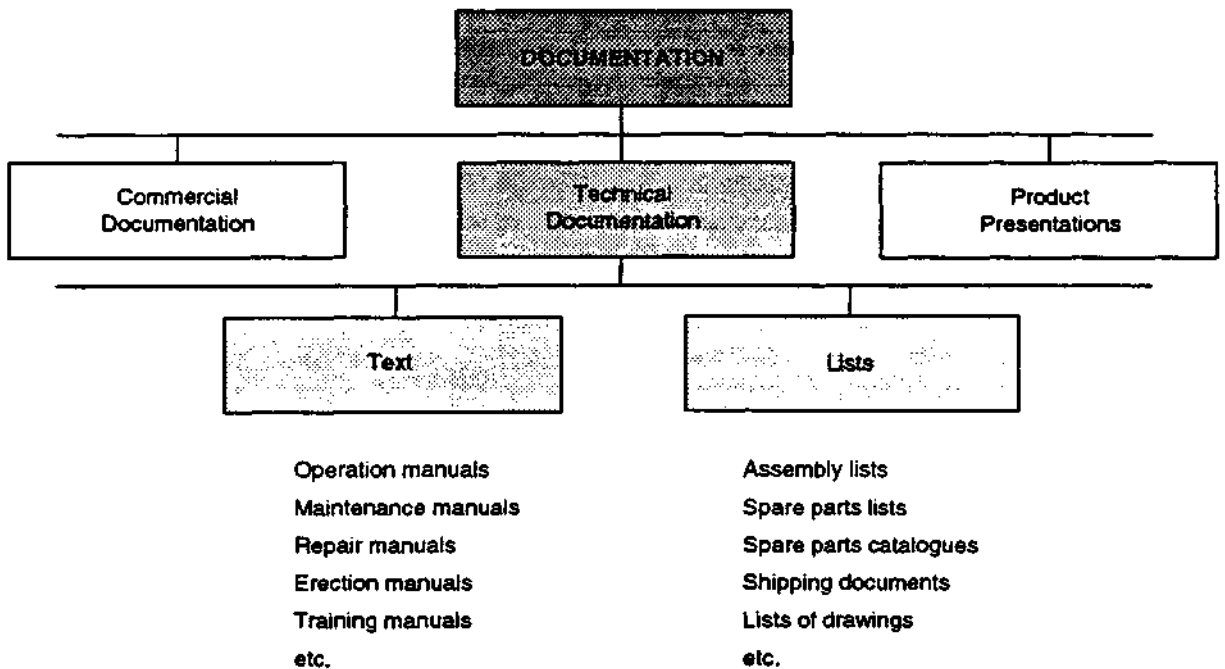
These deficits became more and more evident. It was, however, found to be impossible to implement one commercially available and applicable system that could meet all the expectations.

To understand the complexity, the individual processes, services and products in question were analysed and defined.

What is a document?

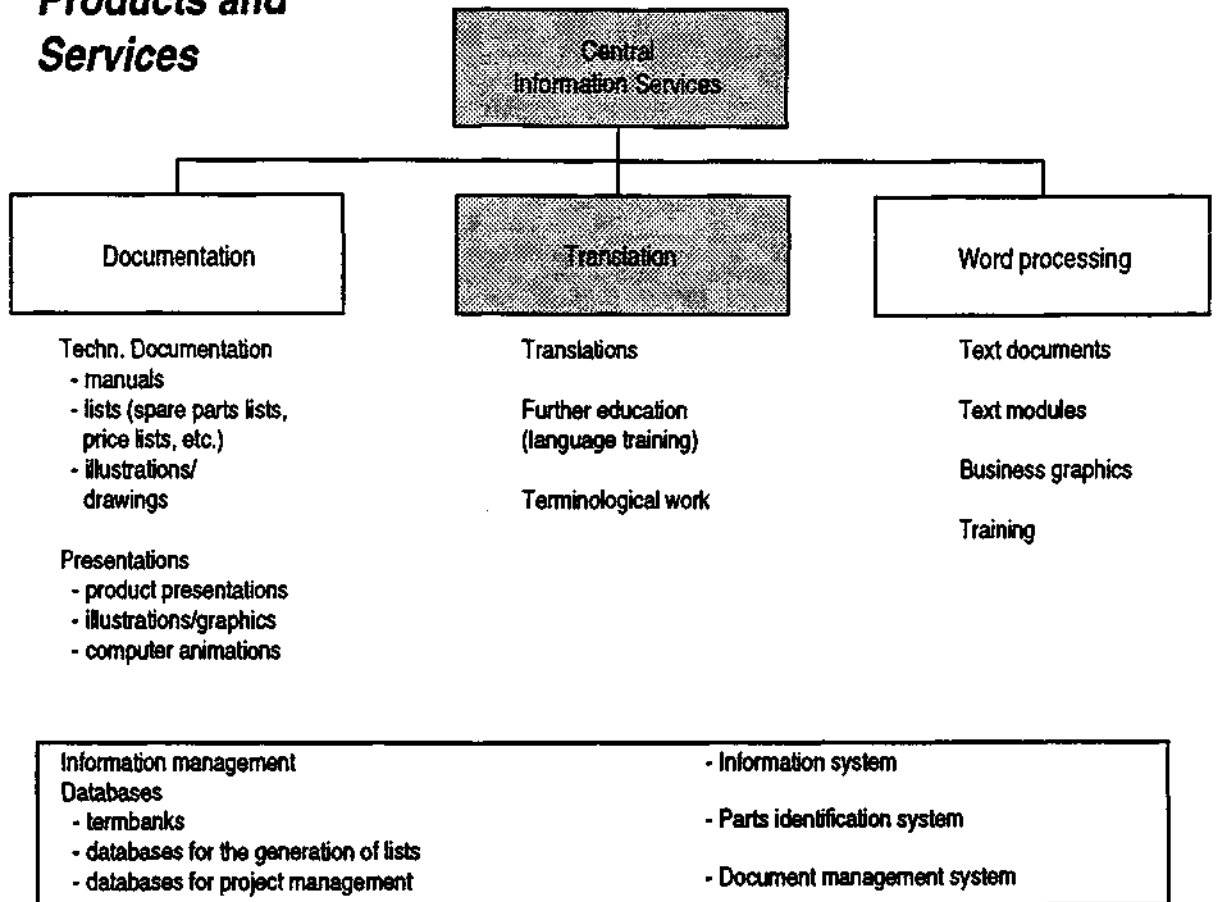


The range of "documentation"



(For the purpose of the analysis, the range of documentation was restricted to technical documentation such as manuals and lists. In principle, however, it was found, that the results of the analysis more or less applied to all kinds of texts forming part of "documentation".)

Products and Services

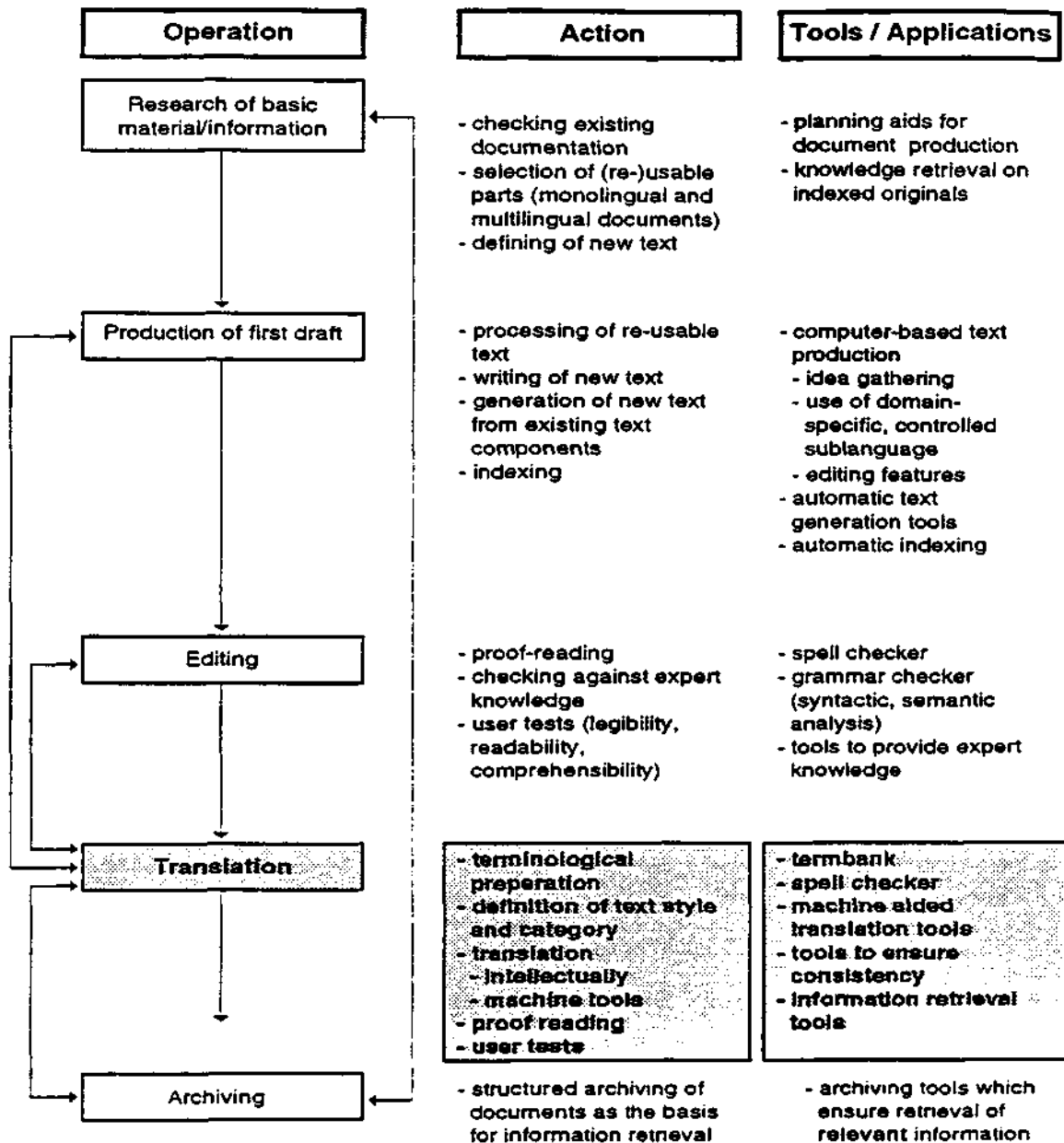


THE ROLE OF TRANSLATION IN DOCUMENT PRODUCTION

The integration of all phases of the document production process: text creation, translation, editing, reviewing and publishing into one, easy-to-use application is a prerequisite for both cost effectiveness and quality assurance, as well as for the protection of the investments in software, hardware and training.

In the overall document production chain, the role of translation is of central importance. Accuracy, consistency of terminology and content, speed and quality are necessary criteria both for source language and target language documents.

Technical Documentation - Text Production Process



The multilingual element of text production is an integral part of all individual stages of multilingual document production, beginning with the research into basic information through to the structured archiving of the finished document.

Whereas during the initial stages the process of gathering information is concentrated on terminological and linguistic data, information about the product and processes gets more and more important during the later stages of processing, in particular when the contents of all the lists and manuals of the documentation are compiled.

Terminological and linguistic data

Terminology work is carried out both independently of targets and also specifically for a special purpose, a project, a product, a domain or a process. The development of denomination and classification systems with unmistakable, unambiguous concepts is a key function for text and document processing applications. The result of the terminological work is

- a) the basis of the source language text generation and processing,
- b) the basis for all automated translation operations and
- c) the basis of all retrieval operations.

Not only must the individual terminology of the documents be correct, unambiguous and consistent, but the descriptors for search functions too. Concepts, their expressions in the source and target language, synonyms and target language equivalents with their respective classification are part of the overall information required.

Terminological work in the sense in which it is used here also include data and information about products, processes etc., i.e. subject information. It is also a prerequisite for :

Consistency of terminology and content

The demand for terminological consistency results from criteria such as product liability, user security and comprehensibility. Terminological inconsistency arises in particular when large amounts of documentation have to be written, rewritten, edited, compiled or translated by a number of authors or translators simultaneously. To ensure consistency, it is necessary to define concepts and their denominations clearly, eliminate ambiguous terms, avoid synonyms and categorise the terms to be used.

Document consistency on the other hand must be assured by the document management system. It means that the same content in the individual parts of a documentation must be the same in various parts of the documentation and over periods of time. E.g.: safety instructions in the operation manual, maintenance manual, repair manual etc.

Another vital part of translation which has to be fulfilled in this context is

Target group orientation

For the translated version of a document, too, an analysis of the target group is necessary. The results of this analysis define text style, text category and terminology. Sometimes it may be necessary to change both the style and the terminology of the source text document because the users of the target language version are different from the users of the source language document. This can mean that the translator may have to change the document, in order to explain the complex processes and operations that are more or less clear to the reader of the source language document in more detail.

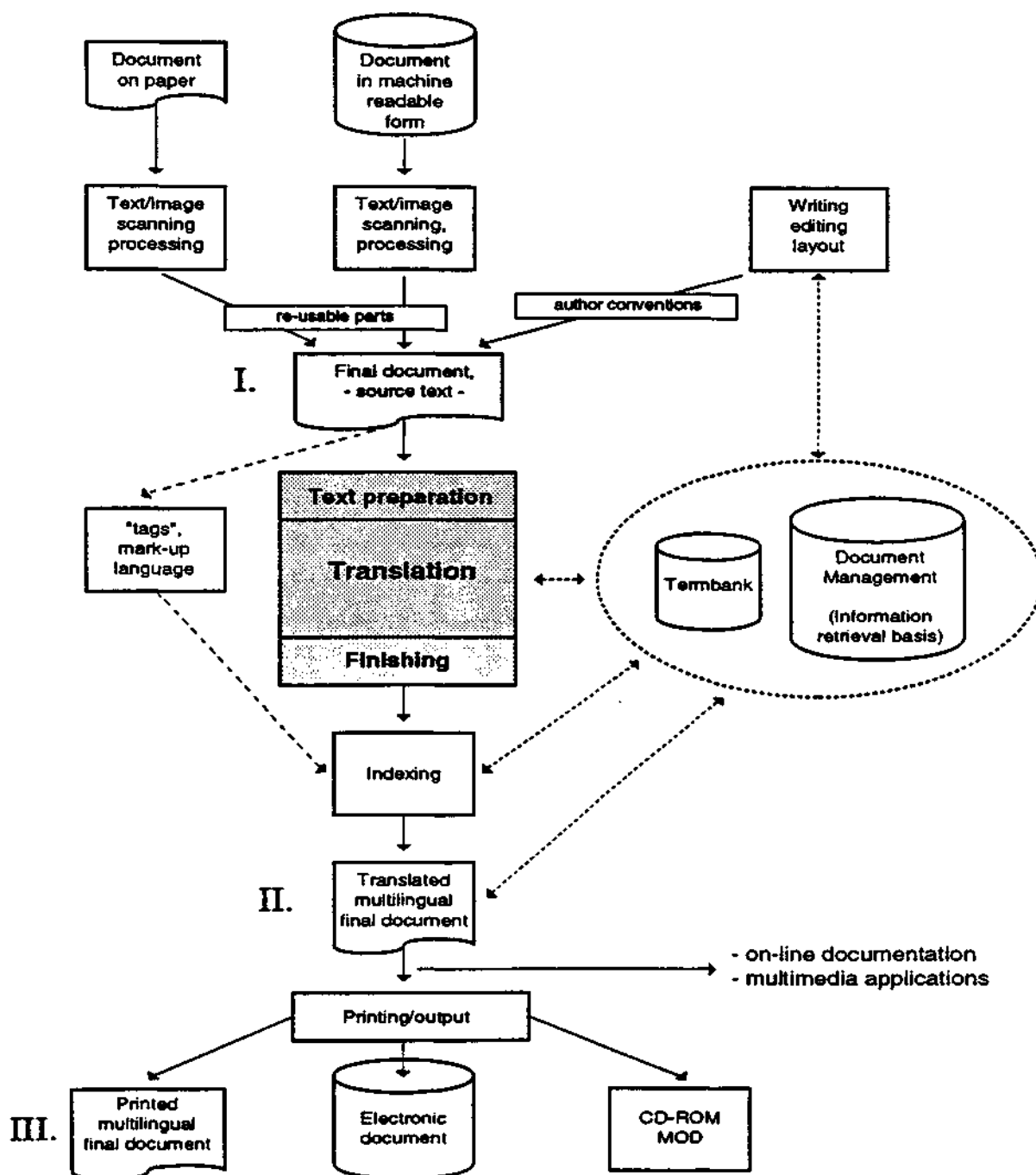
For this intellectual work, the translator/author must have tools at hand that give him the necessary and often very specific information at any time during the translation/writing process. This means that immediate and unconditioned change from one text style to another and from one terminology specification to another (general language, sublanguage, restricted language) must be assured.

THE SYSTEM OF TOOLS AND APPLICATIONS

The fulfilment of the described tasks requires intelligent but not too sophisticated electronic aids. Undoubtedly, such aids were and are available on the market. Their insufficient performance, however, and the inability to allow optimisable and programmable integration into existing or planned operational processes and environments led to the decision to build an application of own design.

How are the translation tasks organised and integrated into the electronic document production scheme?

Translation in the Document Production Scheme



Based on the fact that translators naturally have a keener interest in language and the traditional form of processing than in technology and the use of sophisticated tools and applications, the solution was to be as close to traditional processing methods as possible. The system of translation aids was to be created according to the following principles:

- ☐ Simplicity, comfortability.
As simple and at the same time as comfortable as possible for the user.
- ☐ Effectiveness, productivity.
As effective as possible with the highest possible increase in productivity.
- ☐ Optimal technical and organisational integration into the existing hardware and software environment
- ☐ Use of commercial software whenever possible
- ☐ Network solution.
- ☐ Solution of interface problems using standards, filters and conversion programmes.

According to the scheme shown above the information needed in the translation process concerns language problems that require knowledge extracted from a base of linguistic knowledge and expert knowledge extracted from a base of subject knowledge.

Tools

Generally speaking, the translator must have access to the same knowledge base as the author/writer of the source language text in addition to the extension of foreign language equivalents. Consequently, the following tools were developed or - when commercially available - integrated:

1. For text and corpus analysis:

a commercial programme, which makes use of available text by extracting the terminological content and entering it into a database structure for further processing.

2. For linguistic and subject related information:

products of own design. "TermBase", a terminological and lexical database, and "KISS" (Krupp Information Support System).

"TermBase"

The information retrieval process within "TermBase" is carried out by traditional data access mechanisms, however, with some additional features in comparison to commercial products. The data kernel of "TermBase" can be adapted to various user- and application-oriented menu structures and front ends such as:

- a menu system, adapted to the MS-Word 5.0 user surface,
- an application, adapted to the user surface structure of Windows (Winword 2.0),
- an Apple Macintosh front end, programmed in hypercard (prototype).

In the menu systems and the front-end, the user has the same surface as in the normal word processing environment. This considerably reduces training time, improves the acceptability of the application and ensures maximum retrieval results.

The following information features can be retrieved from "TermBase":

From the main menu:

From the sub-menus:

☐ **Comment** to the lemma entry,
giving prequalifying hints for immediate
disambiguation

☐ **Source**

☐ **Country code**

☐ **Subject area**

☐ **Morphosyntactic features**
noun, verb, transitive verb, adjective,
adverb, preposition, conjunction,
pronoun, determinant, affix

☐ **Abbreviation/acronym**
with a cross-reference to the long form
record and access to all its information

☐ **Graphics**
Pictorial information such as photographs,
technical drawings, illustrations.
The flow of information starts with the
lemma entry and is pursued via the
"graphics", display with the possibility of
identifying the object and defining its
place in the hierarchically ordered system
of broader, narrower and related terms
(tree structure of products and processes)
leading back to denominations which are
lemma entries on a different level.

☐ **Documentation**

- definition - citation form
- context - broader term
- explanation - narrower term
- remarks - related term
- usage - identification number

☐ **Collocations**

- noun : verb - verb : adverb
- noun : adjective - adjective : adverb

☐ **Synonyms**
Information retrieval in a thesaurus-like
application of synonymous relationships
with prequalifying features such as
SYNONYM, QUASI-SYNONYM,
PARAPHRASE. Further information is
given with the indication of the
individual
term's position in classification systems,
the subject area allocation and textual
information in the documentation field
(definition, context, explanation etc.)

Number**Foreign language equivalents**

To be used for translation and language teaching purposes with prequalifying features such as EQUIVALENT, QUASI-EQUIVALENT, PARAPHRASE. Further information is given through the indication of the individual term's position in classification systems, the subject area allocation and textual information in the documentation field (definition, context, explanation etc).

Gender

A number of features may not seem necessary for the immediate translation process, but are regarded as indispensable for further development in particular for planned NLP-applications.

On the output side, the content of "TermBase" can be represented in various forms:

- as an ASCII coded text file
- as a DTP file in postscript format ready for output on a laser printer or a typesetting machine. The printing/typesetting commands ("tags") are programmed into the output interface and can be changed according to user or application requirements.

"KISS"

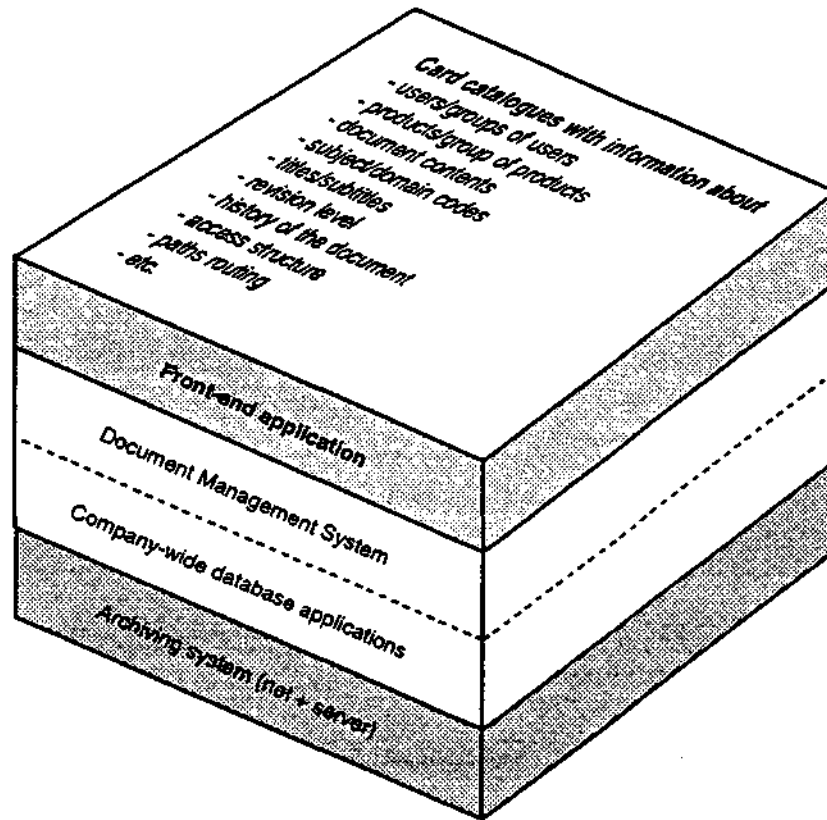
The need to include hypertext functions in the retrieval processes meant the integration of hypertext capabilities in a comprehensive system of applications; capabilities which not only extend the use of a text or document for retrieval purposes but also the use of the retrieval function within databases.

"KISS" is a front-end prototype development designed for use at any working place in the company that needs access to information contained in documents and about documents (linguistic and textual information, factual information, pictorial information etc.).

The basis is represented by a mainframe application built on traditional database technology. The intermediary device between the archiving system and the mainframe application on the one side and the front end on the other side is the document management system "DMS".

The "front end" is organised like card catalogues containing entries for each individual object, such as information about users, groups of users, products, groups of products, document contents, the subject or domain codes, titles, subtitles, revision level, history of the document, access structure, paths routing etc.

Krupp Information Support System (KISS)



To give an idea of how this is organised, here's an example. Through a series of hierarchically organised menus, the translator locates the particular document or part of the document he/she is looking for and gives the commands to the system to deliver it. In accordance with the user modelling structure the system checks the appropriate access to the file and then places the correct version of the document or the individual part of the document on screen, so that the research into the document can begin and the relevant information can be extracted.

To define and manage the document workflow, objects can automatically be routed through the production stages (creation, translation, editing/layout and completion), and the distribution channels (for example translator, reviewer, editor, layouter). When the translator finishes a document and checks it in as complete, the document management system recognises that the translation step is finished, changes the status of the document, gives access for reviews and notifies the reviewer that the document is ready for review.

Access to components within a document or to specific parts of a document is facilitated by marking the repetitive parts. A single object, such as the description and translation of a specific product or process, an illustration or a table can be used in many different documents. This feature is important for consistency and for retrieval as well.

In the hypertext document, the author sets links between related pieces of information. Whenever the document is re-used for document generation or for translation, an active area (a word, phrase or picture containing a hypertext command) is clicked on, and the programme then carries out the command such as bringing the user to a related graphic, description or definition etc. E.g.: an author inserts a series of hypertext links in a technical manual. Any user can now easily navigate through the manual via the embedded hypertext links.

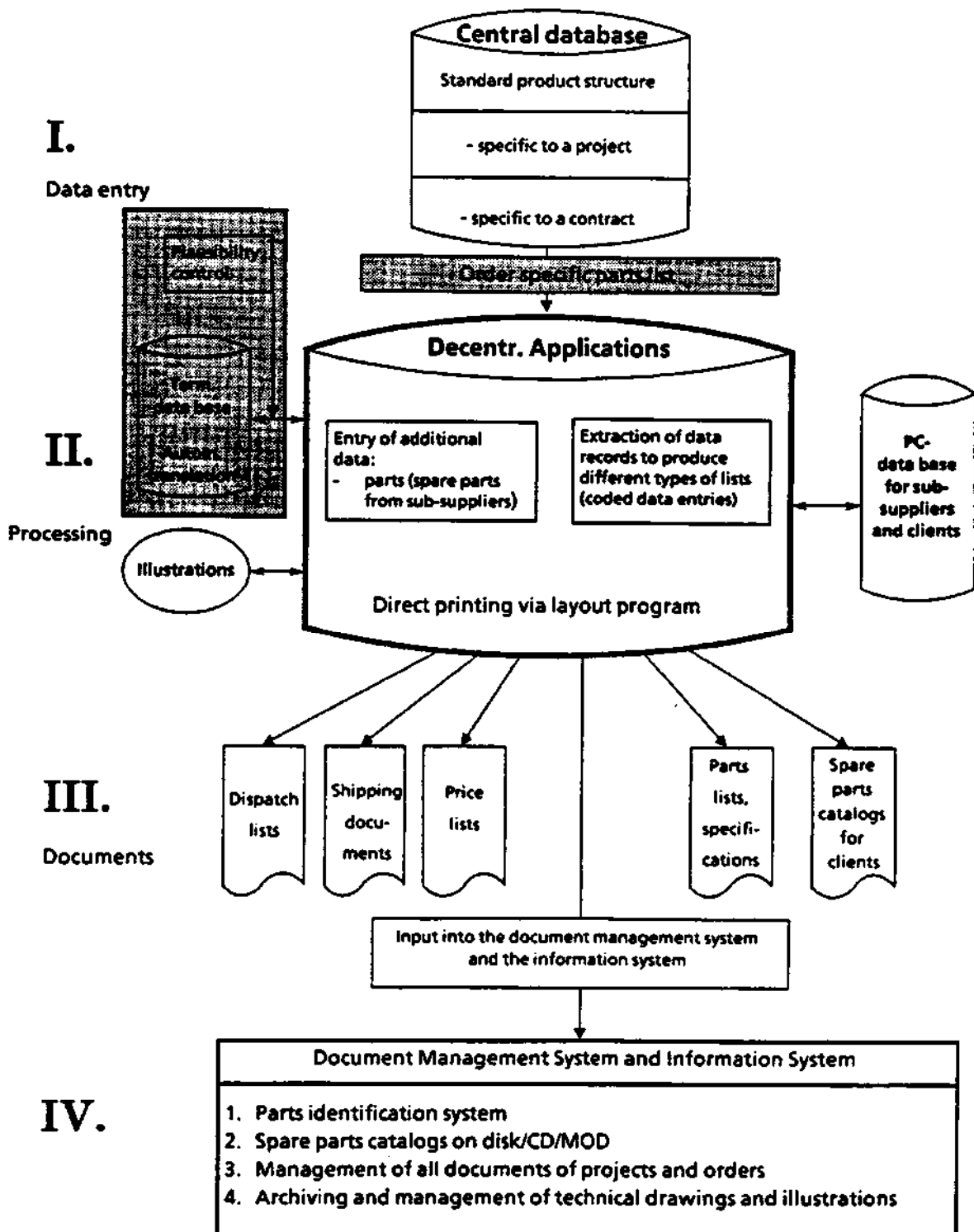
Applications

Based on these tools, the following applications were developed and are in use:

- Dictionary Lookups with "cut and paste" functions
- Macroprogrammed routines at word processor level
- Programmes for the automatic matching of terms (pattern matching)
- Automatic and interactive matching of text modules (sentence parts, sentences, sections of text)
- Applications for tag recognition and automated term matching for text files which can be extracted from a layout programme (commercial product). This application protects the codes (the "tags") through inversion, colour marking or simple copy marks protection.

These applications are used, e.g., for the translation of lists:

Integrated List Production



There are two ways in which machine-aided translation for lists:

- A programmed routine in batch mode. It matches the text parts of a list with the corresponding entries in a database. The source text terms are replaced by foreign language equivalents or additionally pasted into the list. If necessary, this process is preceded by a plausibility control routine between term and identification number to ensure unambiguity. In this batch process modification, updating, elimination of mismatches etc. have to be carried out by post-editing of the already translated document. The output can be an ASCII text file or an DTP-file including all printing commands programmed into the database output interface.
- A programmed macro (set of macros) within a word processing environment in the following stages:
 - Window 1: Loading of source language text file (list) into the word processor.
 - Window 2: Extraction of corresponding terminology (sublanguage terminology) from the database and loading as text file into the word processor.
 - Matching process similar to the routines described above.
 - Window 3: Output of mismatches into "Window 3" for later revision.

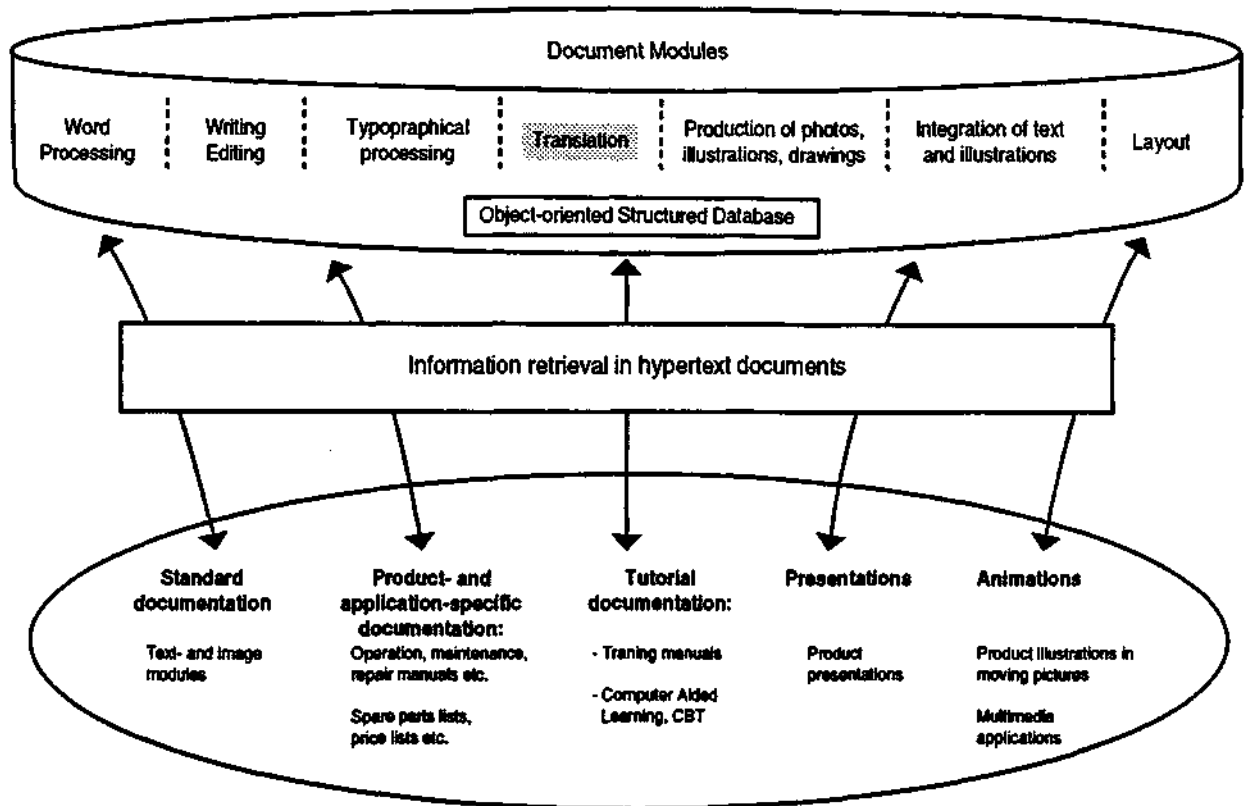
This process is interactive. The macro can be stopped (interrupted) and restarted. Depending on the predefined layout of the lists, the foreign language equivalent replaces the source language term or additionally places it into the text. Principally, both applications enable the fully automatic production of lists plus translation.

The translation of running text

There are also two ways in which machine-aided translation of running text can be carried out:

- *Interactive mode:* Direct database access from the translator's working environment in an interactive mode.
 - a) DOS-version
A memory-resident (TSR) version of the database "TermBase". The translator switches between the word processor or any other DOS programme and the database (the programmed interface is activated by a hotkey).
 - b) Windows version
The user switches from the Windows environment to the database integrated into Windows. This version additionally allows input and output of large parts of text and documents and is thus the basis of a "translation memory"-like application.
- *Automatic matching:*
The terminological content of a subset (selected, domain-specific terminology) of the database is converted into an ASCII coded text file with predefined syntax. During translation, terms, compounds, parts of speech etc. are read out of the screen display and are automatically matched against the text file. (Commercial programme). The successful matches are highlighted indicating that there is one or there are several translation proposals. By pressing a hot key the foreign language equivalent can be pasted in. This applies also to large parts of text. The interactive "cut-and-paste" function can be replaced by automated batch mode processing.

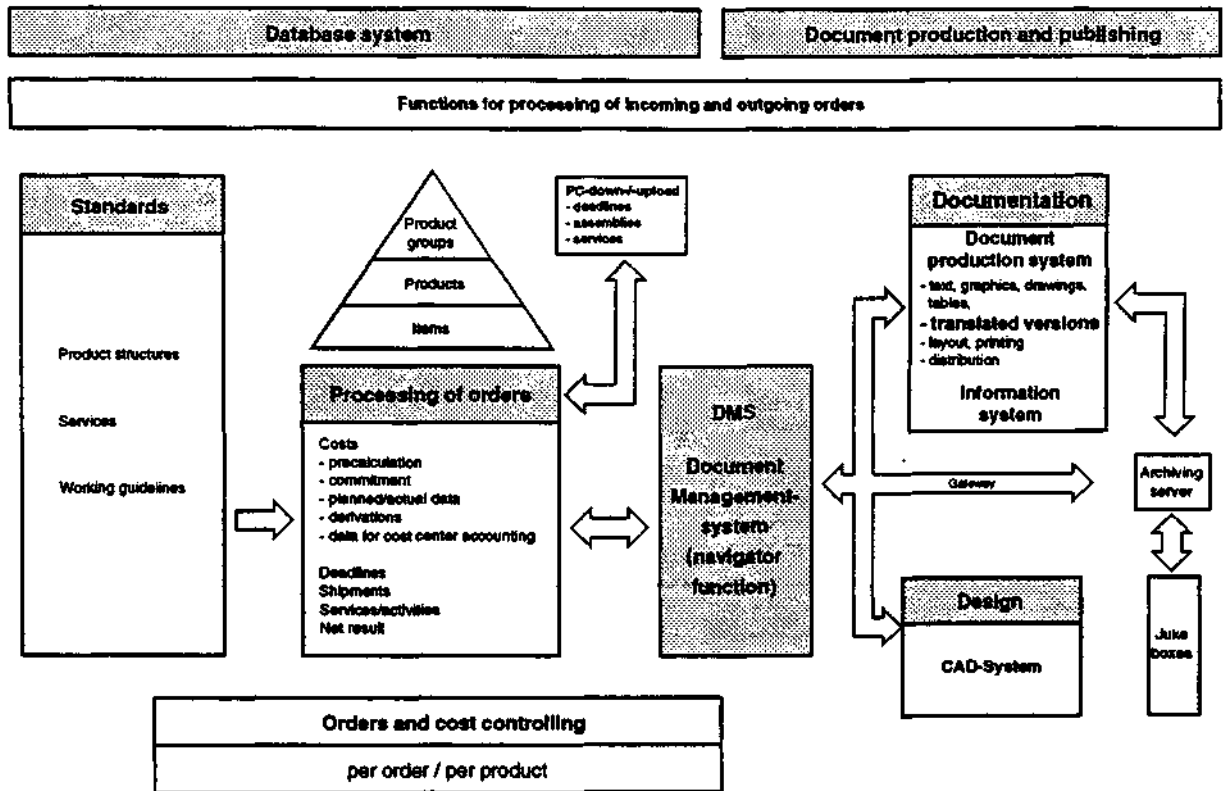
Information Retrieval into a Common Data Pool



The text/document modules are the basis for the "Translation memory"

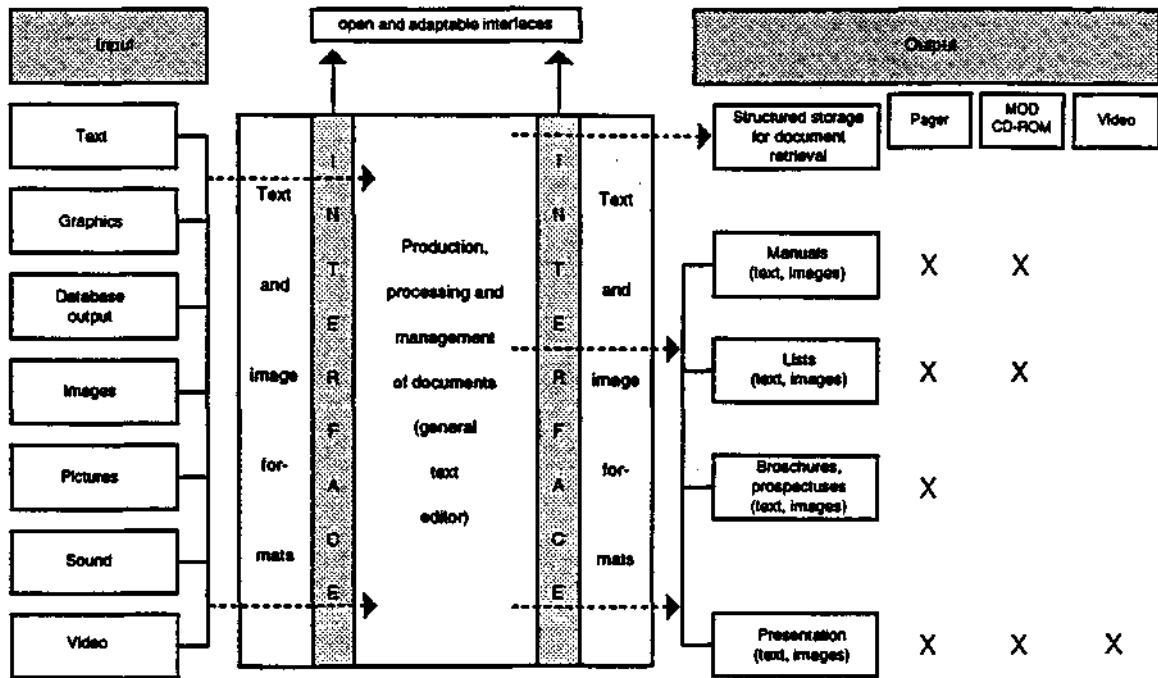
All the described tools and applications for the translator are integrated in document production process which itself is integrated into the whole scheme of integrated company-wide system of EDP-applications:

Technical and Organisational Integration



This integration, of course, is only possible when as many standards, filters, or conversion programmes as possible are applied and integrated.

Input/output Interface Manager for Document Production and Management



Although all documents internally are produced in a common text editor with hypertext function and mark-up language ("tags"), there is a large part of documentation delivered by suppliers/sub-suppliers with different formats. This requires adaptable interfaces on the input side. The same applies to the output side, as most clients want to have their documents in a special format for their internal use and processing, which requires interfaces for the export into the document formats required.

These standardisations and interfaces do also have a highly productive effect on the use of the database output. A standardised data and system structure is a prerequisite for the internal and external proliferation of these data for sub-suppliers and clients.

Translators are very much interested in the exchange of terminological data and text files. This makes it necessary to have at hand standards for the macrostructure of the different databases in use (modelling of the lexicon) and the microstructure (modelling of the terminological entry) via a standardised data definition language.

For "TermBase", an output was programmed in an easily convertible mark-up language based on "tags" that can be transformed into ASCII codes and thus converted by anyone who wants to use the data.

ECONOMICAL AND SOCIAL ASPECTS

Organised as a profit centre within a large company, the organisational unit "Central Information Services" is very much interested in automating its production processes both from an economical and qualitative point of view. The cost-saving possibilities are evident.

Direct access to relevant information saves time and money, as e.g. on multiple terminological research, elimination of ambiguous synonyms, equivalents etc.

The cost-saving potential - especially in combination with the word processing and technical writing sectors and when DTP/CAP-tools are being used - is enormous in comparison with the traditional working methods. It, however, necessitates:

- use of "controlled language" and introduction of author conventions,
- consequent re-use of repetitive document modules,
- optimal organisation of the production line,
- preparatory planning,
- use of electronic aids.

Although a lot of preparatory work had to be done for the realisation of the described system of tools and applications, it is now very successful in economical terms. The organisational integration of writing, editing, word processing, computer-aided publishing and user programming in one organisational unit has proved to be very efficient

The social aspects of the streamlining and rationalisation processes have both positive and negative sides. It requires, on the one hand, highly motivated and qualified employees that are prepared and willing to work in project management teams, and better qualification as a rule means better reputation. It assures both a higher status in the social environment of an organisation and a working place guarantee for qualified and skilled labour. On the other side, the significant rationalisation effects have their implications on the employment of less qualified staff.

An analysis of all translated texts carried out before the introduction of the described applications showed that approximately 40% of the translation work amounted to lists and technical specification with restricted uniform syntax and semantics, demanding very strict adherence to consistent and unambiguous use of terms and phrases. Another 20 - 25% were covered by highly repetitive text.

Thus, approximately 60% of the texts were very apt for the use of machine aids in computerised applications. None of it certainly very exiting and challenging for the translator but, nevertheless, work that assured a constant work load. Now, with approximately half of the work being processed by the computer, this means at least 50% less employment.

Fortunately, we have until now been able to counterbalance redundancies by re-training and transferring personnel to other sectors within the company or through early retirement.

CONCLUSION

The purpose of building a system of own design was to increase productivity and efficiency of the translation and document production process, and to reduce workload and costs.

The problem with which we are faced as an internal profit centre is that we have to deal with the cost structure of a large company with all the inflexibility and, in particular, the large overheads on the one side and the need to be competitive in terms of market conditions on the other side.

This can only be done successfully when the products and services are offered as packages on a long-term basis, so that synergy effects, technology advantages, and the know-how of insiders can be fully exploited. External resources are used, however, with complete integration into the internal process and adaptation to existing structures.

Translation is a service that has precipitated an enormous demand and rapid growth during the last few years. The costs of producing multilingual documentation as a whole have risen dramatically because of the rapid increase in the amount of documentation to be produced and translated.

The main question was and is: How can one cope? How can one cope without incurring significantly higher costs?

Principally, what is needed for the future is a low-cost form of on-line document and data distribution which ensures that all those involved in the production process as well as the end user have access to the most up-to-date information, and new intelligent applications which not only provide help for the technical side of the production process, but also for the automation of processes which are still being carried out manually and intellectually today.

Such tools are in the pipeline. Research and development is taking place at a great pace. Hopefully, this part of the translator's work will experience the same keen development in the future which we have enjoyed ever since electronic data processing was first introduced in the office communication environment.