InfoCodex
Semantic Technologies

Added Value from Information

by Efficient Knowledge Management

General Situation

Knowledge has indisputably become an essential production factor in today's economy. Vast quantities of information are available on the Internet, in e-mailboxes and on file systems and databases across large enterprise networks. Although this information could, in principle, be accessed, the question arises, however, whether it is really exploited to it's full potential. For example:

- A customer contacts a company with a request. How does the company's representative find previous similar requests and the corresponding answers?
- An engineer would like to develop a new machine component. How does he find out if another department has already undertaken similar development work? How does he find the relevant technical specifications and the country-specific safety regulations to be accounted for in the product development?
- How can a document be found that one recalls having seen some time ago as an e-mail attachment?
- How can the employees' knowledge of one department be made available to other employees of the company in a rational and profitable way?

Knowledge Discovery by Means of InfoCodex

A large proportion of today's knowledge is only available in the form of unstructured, individually formatted documents comprising pictures, graphs, tables and text. Typically, these documents are spread across a diverse variety of repositories and are written in different languages.

A software system capable of effectively supporting large companies and organisations in the discovery and management of knowledge must be able to gather information from various platforms and heterogeneous text sources. In order to exploit fully this information, the system must be able to recognize and represent the content of a document and must realize that the English and German version of the same document contain the same information. In addition, an automatic indexing for efficient searches across languages is indispensable.

This is the world of InfoCodex. Its advanced and unique information gathering and analysis technology (linguistics + taxonomy + statistics + self-organizing neural networks) is merged with a large multilingual, linguistic database whose entries are linked to a universal taxonomy for content recognition. This empowers InfoCodex for the following tasks:

- Automatic content recognition and logical organization in an information landscape (→ *"arranged into a well organized bookshelf"*) without human intervention, i.e. without a cumbersome case-specific training of the categorization by experts.
- Automatic high quality key word generation.
- Efficient full-text search including synonyms and cross-language retrieval.
- Similarity search, i.e. retrieval and ranking of documents which are similar in content to a given query text.
- Automatic assignment of new documents to existing document categories (→ response management, e.g. automatic e-mail distribution).

Application Examples

The examples below illustrate how InfoCodex is currently being used by private companies and government organizations.

- *Market Observation, Competition Monitoring*

  Periodic download of information from the Internet or from specialist databases (can run automatically in the background); owing to its unique technology, InfoCodex can also find and analyze the latest Web pages that are not yet listed on major search engines.

  InfoCodex provides a thematic overview of the collected information organized in a meaningful way which enables efficient location of documents of particular interest.

  Benefit*:* Substantial time savings in the routine work of information research on the Internet and specialized databases; well-founded market knowledge; solid basis for decision-making.

- *Patent Research*

  Efficient search for existing patents describing inventions similar to the technical description of a potential product or patent submission.

  Benefit*:* Substantial improvements in the economy of the time-consuming process of patent research; avoidance of costly re-inventions of existing products.

- *Enterprise Search-Engine and Coordination of Knowledge Resources*

  Knowledge retrieval can be cumbersome even if documents are stored in a sophisticated DMS system. Since manual key-word setting by qualified personnel is often too expensive and the viewpoints can change in the course of time, the key words are often ambiguous and hence an effective retrieval of specific information can be difficult and cumbersome.

  InfoCodex creates a virtual order according to the current status of the documents, independently of the underlying archiving system and without human intervention.

  Benefit*:* No manual work required for the organization of the archive and the generation of key words; nevertheless, a high-quality thematic overview of the entire collection of information and powerful search mechanisms are available at any time.

Information Research on the Intra- and Internet

The development of a new product requires substantial effort being put into retrieving national, multilingual safety regulations and technical standards, sometimes even more than for the technical research and development work. With InfoCodex the access to the relevant regulations becomes substantially more efficient.

<u>Benefit:</u> Elimination of cumbersome routine retrieval; improvement of the information basis.

- *Key Word Setting for Documents*

  When archiving documents it is often necessary to tag them with key words so that they can be found again at a later time. Without even addressing the cost problem, it is unlikely that an individual can be found to do this job for all topics covered in company documents (how should an engineer correctly tag a biology document). Owing to its large knowledge database InfoCodex can automatically generate key words across most subjects.

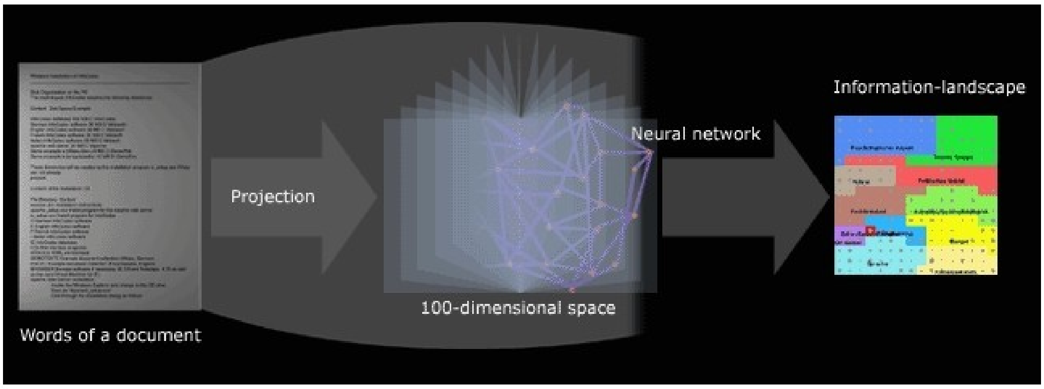
  <u>Benefit:</u> Combination of time savings with an increased quality of key words in the archiving process.

Technology

InfoCodex' spider agents collect the documents from multiple sources (Internet, Intranet, local clients, mailboxes, databases). In addition to the thematic content, a set of metadata is extracted from the documents and stored to assist in subsequent retrievals: title, author, date, language, format (Word, PDF, HTML, XML, ppt etc.), document origin.

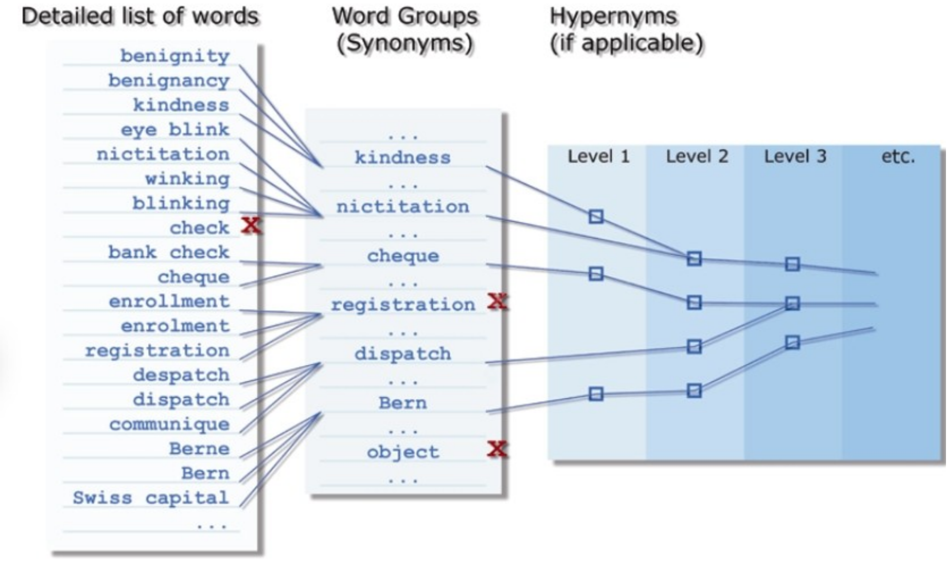
Words and expressions (= collocated words like "human rights commission", "magnetic resonance imaging") are checked against the entries of a multilingual linguistic database, condensed into synonym groups and assigned to the nodes of a taxonomy tree. The hit distribution on the taxonomy nodes provides a clue to the thematic character of a document (→ **content recognition**).

On the basis of the effective content of the entire document and information collection and its internal universal knowledge repository, InfoCodex constructs a 100-dimensional semantic space whose axes are chosen to characterize the content of the documents in an optimized fashion. The system takes into account both the relative significance of the words according to its linguistic database and collection-dependent quantities derived from information-theory such as entropies and relevance indices. After a projection onto the 100-dimensional space, the documents are logically clustered (→ **arranged into a well organized bookshelf**) by self-organizing neural networks. Additionally, the documents are indexed and provided with lexical descriptors for an efficient search. All this happens without any human intervention and hence no training is required.



The encapsulated linguistic database contains, at present, more than 2.4 million entries in English, German, French and Italian. Words/expressions with the same meaning and their morphological variations are aggregated into synonym groups. For each item the word type (noun, verb, adjective, adverb etc.), the language, the significance (on a 0-4 range) and the assignment to a taxonomy node are stored. In addition, a hash-value for the accelerated identification of composed expressions (e.g. "European Court of Justice", "President of the United States") and a flag for the treatment of ambiguous words are stored.

This linguistic database covers the whole range of human knowledge. It is based on well-established works such as WordNet of Princeton University, Eurovoc of the European Union, Jurivoc of the Swiss Federal Court of Justice, the taxonomies of several universities and professional organizations in the private domain and the public administration. This base structure can easily be complemented with a customer-specific linguistic database which is then superimposed to take precedence over the InfoCodex database. This is particularly important when company-specific word meanings and codes have to be taken into account.



The InfoCodex technology is patent-protected in Europe (including Switzerland) and the USA (patent pending).

Benefits of the InfoCodex Technology

- Logical clustering/categorization of documents without human intervention (i.e. no training required)
  → overview of a whole document collection in well-organized bookshelf.

  The cumbersome categorization training by means of prepared document sets and the time-consuming preparation of content-specific taxonomies are not necessary (as opposed to the needs of competing technologies).

- In addition to the traditional full text and index searches, InfoCodex also offers similarity searches (thematic searches).

  Documents are retrieved on the basis that they are similar in content to a query text submitted in colloquial terms.

- Establishment of a well-founded measure for the content and thematic similarity of documents: this measure is used for both the clustering in the information landscape and the similarity retrieval.

- True multilingual functionalities: cross-language retrieval, even in similarity or thematic searches.

- Visual representation of the content of a document collection in an information landscape and heat-map representation of search results.

- Automatic key word generation of high quality.
  A company has evaluated and compared the quality of key words generated manually by five well trained professionals and key words generated fully automatic by InfoCodex. The generated key words were assessed by experts in the relevant fields which revealed substanital advantages for InfoCodex:

InfoCodex	90% acceptable to good	10% bad
Manual	50% acceptable to good	50% bad

Installation Requirements

- *Platforms*

  Windows NT, Windows 2000, Windows XP,
  Linux or Unix (IBM AIX, Sun Solaris, HP UX ...)

  Minimum of 512 MB Ram, 10 GB disk space (the InfoCodex software including the linguistic database require about 1 GB but some additional temporary disk space is required as swap space and for the index database)

- *User Interface*

  Standard browser (Internet Explorer, Netscape 7 or Mozilla)

- *Software Requirements*

  Only the operating system and the Web Server (Apache or IIS) are required in addition to the InfoCodex software package. The Apache Web Server can be delivered with InfoCodex if required.

- *Installation*

  In standard cases neither special preparations nor interface programming are required.

  Documents are only read, not copied or stored. The distribution of documents has no influence on the virtual order established by InfoCodex.

  As a result, InfoCodex can be fully installed and commissioned within approximately 2 hours.