

User Manual

Part 1: Usage

Table of Contents

1. What is InfoCodex	2
1.1 Possibilities and Resources of InfoCodex	2
1.2 Typical Application Examples	2
2. Starting InfoCodex	3
3. Searching and Finding	4
3.1 The Information Landscape Map	4
3.2 Retrieval with Search Text	5
3.3 Searching using Document Characteristics (Metadata)	6
3.4 Display of Search Results	7
3.5 Document Families	9
3.6 Alternative Display Possibilities	10
4. Creating a New Collection	12
4.1 Collections and Domains	12
4.2 Setting up a Collection	12
4.3 Setting the Characteristics of a New Collection	12
4.4 Advanced Options for Collection Set-up	13
5. Adding Documents	15
5.1 Selection of Data Sources	15
5.2 Import execution (Content Analysis / Indexing)	17
6. Analysing Content	19
6.1 Categorising Documents	19
6.2 Add Documents	20
6.3 Delete Document Entries	20
6.4 Create a Sub-Collection (see “Analyses”)	20
6.5 Compare New Entries with Basic Set / Trend Recognition	20
7. Synonyms / Taxonomy	22
7.1 Synonym Groups	22
7.2 Hierarchy of Meanings	23
8. System Management (Summary of Part 2 of the Manual)	24
8.1 Data Protection / IC Domain Concept	24
8.2 Collection Administration	25
8.3 Influencing the Categorisation	25
9. Selected Examples	27
9.1 Information Research	27
9.2 Finding Available Know-how	27

1. What is InfoCodex?

The "*InfoCodex Semantic Engine*" is a software instrument for knowledge and document management. It filters the textual content, by means of its spider agents, from a diverse range of document types over various different platforms (intra- and internet, mailboxes, etc); analyses their content irrespective of language; groups the documents thematically and presents the entire document collection in a clearly arranged information map – and all without human intervention.

Tapping Knowledge from Different Sources

An instrument, which should really ease the burden of sifting through masses of information, must be able to collect the documents from multiple platforms and document formats and correspondingly group them as clearly as possible according to their thematic content. It should also recognize that the English translation of a German document has the same thematic content as the original. The automatic indexation for efficient cross-language searching is indispensable.

1.1 Possibilities and Resources of InfoCodex

<i>Document formats</i>	Spider agents for PDF files, Word documents, Excel spreadsheets, PowerPoint files, PostScript and EPS, RTF and plain text files, HTML, XML, diverse Mail formats incl. attachments, PST files (Outlook-archives), JPG, GIF, TIF, BMP, Lotus Notes, ZIP-, GZIP- and GZ files
<i>Content recognition</i>	Based on a multi-lingual linguistic database, whose entries are linked to a universal taxonomy . The database, with meanwhile more than 4 million entries, supports itself on renowned works such as WordNet from Princeton University, EuroVoc, AgroVoc, JuriVoc, CIS and many other technical vocabularies. It encapsulates English, German, French, Italian and Spanish. Russian, Chinese and other languages are supported by external translation resources.
<i>Categorisation</i>	Semantic clustering (by means of the linguistic database), information theory analyses and neural networks (Kohonen map)
<i>Search indices</i>	Words and collocations (phrases) for classical full-text search, synonyms for advanced and cross-language search, similarity measures for similarity search (based on the neural network), e.g. queries using free blocks of text

1.2 Typical Application Examples

- Search machine for company internal network ("Enterprise Search Engine")
- Information research: Combination of the results of different internet search engines with one's own documents, intranet documents, emails etc.
- Market observation, competition monitoring, patent research
- Knowledge coordination (within project groups and external posts)
- Automatic keyword generation and topic categorisation of documents
- Profile matching (comparison of CVs with employment position descriptions, classification etc.)

2. Starting InfoCodex

InfoCodex runs via a standard browser (Internet Explorer, Firefox, Chrome).

Start address: [servername/infocodex.html](#) or [servername/infocodex5.html](#)

Login and choice of interface language (English/German/French/Italian/Spanish; for Spanish the GUI is in English) are entered on the start page.

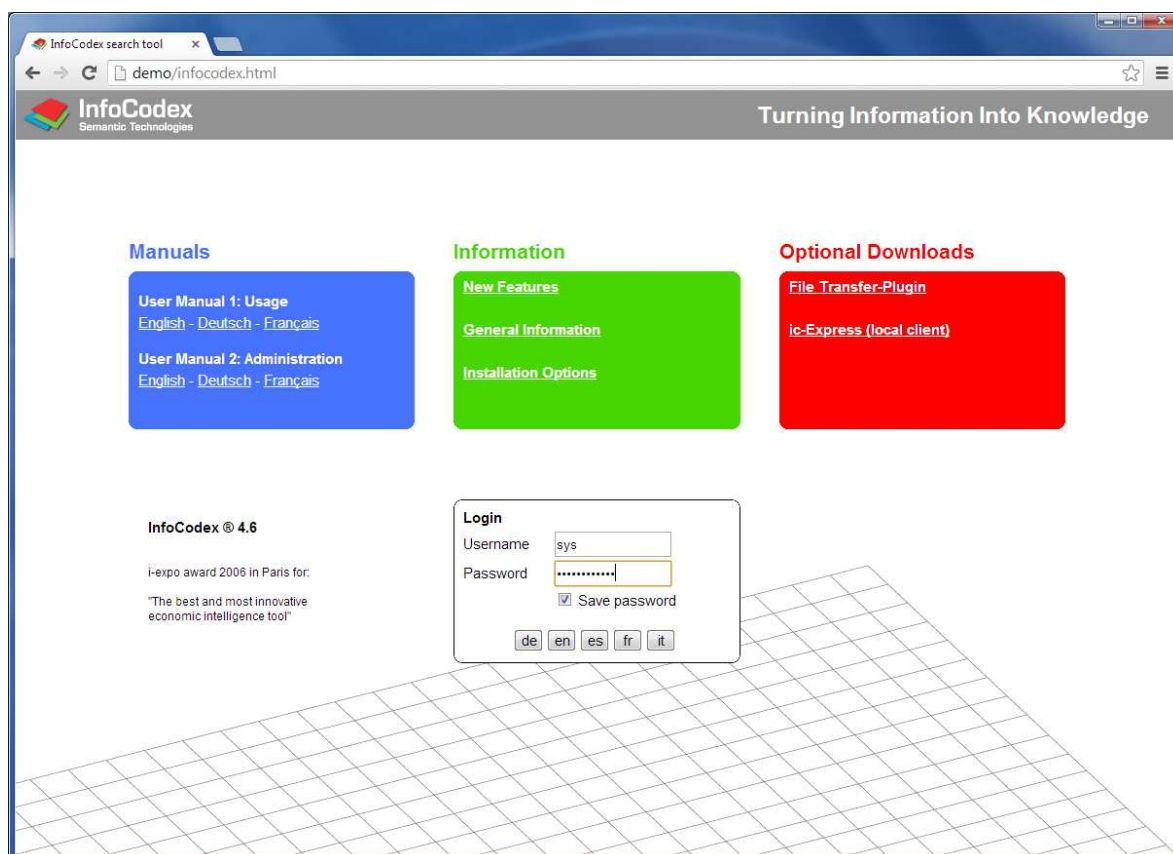


Fig. 1: Start page

The kind of login is installation dependent:

- Username / password corresponding to those of the underlying network
This is the case when the user administration of InfoCodex is coupled with LDAP ("single point of administration", e.g. ADS from Windows). When the access rights of the operating system are to be observed (File System Security), this organisational form is mandatory .
- Username / password are InfoCodex-specific
In this case the user administration and allocation of access rights are governed exclusively from within InfoCodex. This organisational form allows for a "public" user without a password.
- Automatic login (no password prompt)
No username/password is required with this "single sign-on" set-up.

The standard search mask appears after login, where the user can select document collections and specify searches.

3. Searching and Finding

In InfoCodex searching occurs within a previously constructed document collection. A single document collection can be constructed from up to 10,000 different freely selectable sources, e.g.

- documents on the company-internal network,
- pages from the website of a particular organisation,
- results from internet researches using different search engines,
- documents from a CD of the lectures of a professional conference,
- emails and their attachments.

All collections are independent of each other and individually organised (i.e. shelved in a virtual book cabinet) and indexed by InfoCodex. Searches occur within a single collection.

The creation of a new collection is outlined in Section 4. Let us assume, for the moment, that several collections already exist, and outline how documents are searched and found .

First, we select a collection from the list top/left of screen, after which the corresponding information landscape map is displayed on the right hand side showing the thematic layout of the collection.

3.1 The Information Landscape Map

The information map provides a graphical overview over the content of the entire document collection, comparable to a library bookshelf where the documents are ordered by the character of their logical content.

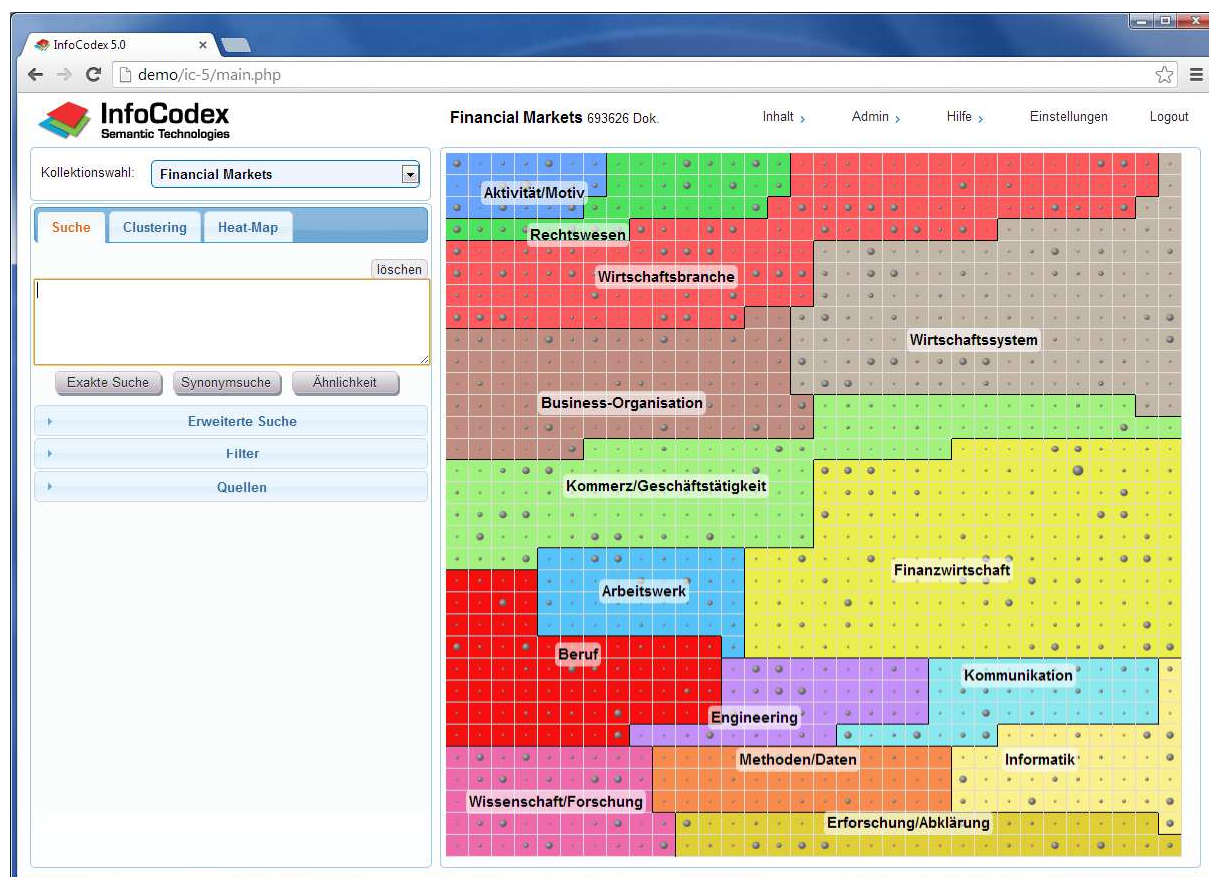


Fig. 2: Search mask (left) with information landscape map (right)

Each coloured field in the information map corresponds to a principal topic. Each topic consists of one or more fields ("neurons") which in turn contains documents of similar thematic content. The size of the dots in the centre of the neurons is an indication of how many documents are contained in that neuron.

By hovering the mouse over a particular neuron, further information is displayed e.g. keywords to the main topic and the number of documents in the neuron. On clicking the neuron, the actual documents contained within can be displayed.

3.2 Retrieval with Search Text

Enter the search text / terms in the text field in any of the languages: English, German, French, Italian or Spanish. Start the search with one of these three choicess:

Exact Search

The search terms must occur in the exact form they are entered

by Synonym

Cross-language search including synonyms: a search for "pushbike" also finds "Fahrrad", "Drahtesel", "bicycle", "bicicletta" etc.

by Similarity

Documents with similar thematic content are found; a matching of search terms is not necessary

Exact Search

The exact search corresponds to the classical full text search familiar in the established search engines like Google, Yahoo, Bing etc. An exact match of the search terms and the words in the documents is required, albeit case-insensitive. The word "bike" will neither match with "bikes" nor "bicycle" nor "pedal cycle". Numbers and insignificant words like "the", "it", "that" etc. will be taken into account.

Additionally, the use of the asterisk "*" as a wildcard is possible: a search term like "neuro*" will match with anything that begins with "neuro", e.g. like "neuroleptic", "neurological", "neurologist".

If a search should return very few documents, InfoCodex offers the chance to include documents that do not contain some of search terms. Such documents are weighted less relevant than those containing all search terms.

Synonym Search

The synonym search is a true cross-language, semantic search.. As opposed to the exact search, this facility interprets the search terms intelligently:

- Synonyms are taken into account:
A search for "bicycle" will also match "bike", "pushbike", "racing cycle", "bicycles", "pedal cycle" etc.
- A cross-language search takes place:
A search for "bicycle" will also find "Fahrrad", "bicyclette", "bicicletta", "velocipede", "Drahtesel" and such.
- Collocations (compound terms of more than one word) are also recognised:
"Europäische Union", "common sense", "pomme de terre"

Similarity Search

With the similarity search, none of the terms of the search text need to match the terms in the documents. Documents are compared in terms of their thematic content and ranked in terms of similarity. The search text in these cases is typically a longer passage of text in prose form, like a piece of text copy/pasted from some existing document. The thematical comparison occurs using InfoCodex' similarity measurement methodology.

Typical examples for using the similarity search:

- Patent research
- Panning of research work: Comparison of new research intentions with existing works in a document collection
- Comparison of foreign publications with one's own work
- Processing of client queries: Finding similar past cases
- Finding norms and legal regulations to a given problem

A pure similarity search with few search terms and without any further restrictions is generally fairly useless because the content similarity of the search is very insignificant. In such cases InfoCodex will recommend the use of a synonym search instead.

AND/OR Combinations

All words separated by a space or a comma must be present (AND operation like Google, for example). OR operations can be stipulated with ";" (semicolon).

Search terms	Meaning
traffic car	Both "traffic" and "car" must be present
traffic; car	Either "traffic" or "car" must be present
traffic car; transport	Either both "traffic" and "car" must be present, or the term "transport" alone.

Such combinations make no sense in a similarity search, and are ignored.

Boolean Search

With the boolean search, arbitrarily complex search texts can be specified and combined with a similarity search. The rules can be viewed under the menu point:

[Help > ? Search instructions](#)

3.3 Searching using Document Characteristics (Metadata)

As well as searching for document content, it is also possible to filter results using some document characteristics. The so-called metadata.

Advanced Search

The advanced search permits a search for metadata. These include characteristics like author, title, source and date of the document. The document's language and the date of import are also stored as metadata.

Documents' file names can also be used as a filter constraint with asterisk "*" wildcard possibilities like "proj*measur".

If the import has provided further characteristics such as client number or branch, these

The screenshot shows a web-based 'Advanced search' interface. It features several input fields for filtering documents: 'Comment', 'File name (parts)', 'Title (parts)', 'Author', 'Doc. date' (with 'from' and 'until' sub-fields), and 'Import date' (also with 'from' and 'until' sub-fields). Below these are checkboxes for 'File type' (Word, Excel, PDF, HTML, PPT) and 'Language' (en, de, fr, it, es). At the bottom, there are buttons labeled 'Filters' and 'Sources'.

can also be searched for. In the default case, such extra metadata is entered in the "Comment" field (except when custom-made additional fields have been introduced).

Filters

Search results can be further constrained using filters. InfoCodex creates filters automatically to dominant persons, topics, organisations and locations in the analysed documents. In this way, one could restrict a search of documents, for example, concerning John F. Kennedy and Berlin.

The location filters can optionally be hierarchically structured; in the above example "Berlin" is a subordinate element to "Germany". In this case, a restriction to "Germany" automatically leads to "Berlin" too. The linguistic database is the basis for the establishment of this hierarchy.

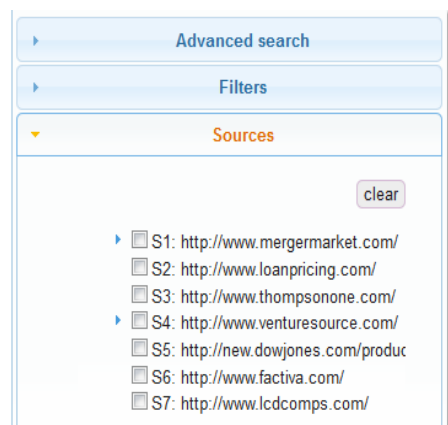
Depending on the collection further user-defined filters can be available.



Sources

With this option the search results can be restricted to particular sources – the physical origin of the documents.

A collection typically contains documents from diverse sources: like local files from different network storage areas, web pages, search results from search engine queries, etc. All documents are taken from the given data sources in the collection that InfoCodex can analyse. The documents' source can be arbitrary; it has no influence on InfoCodex' categorisation.



3.4 Display of Search Results

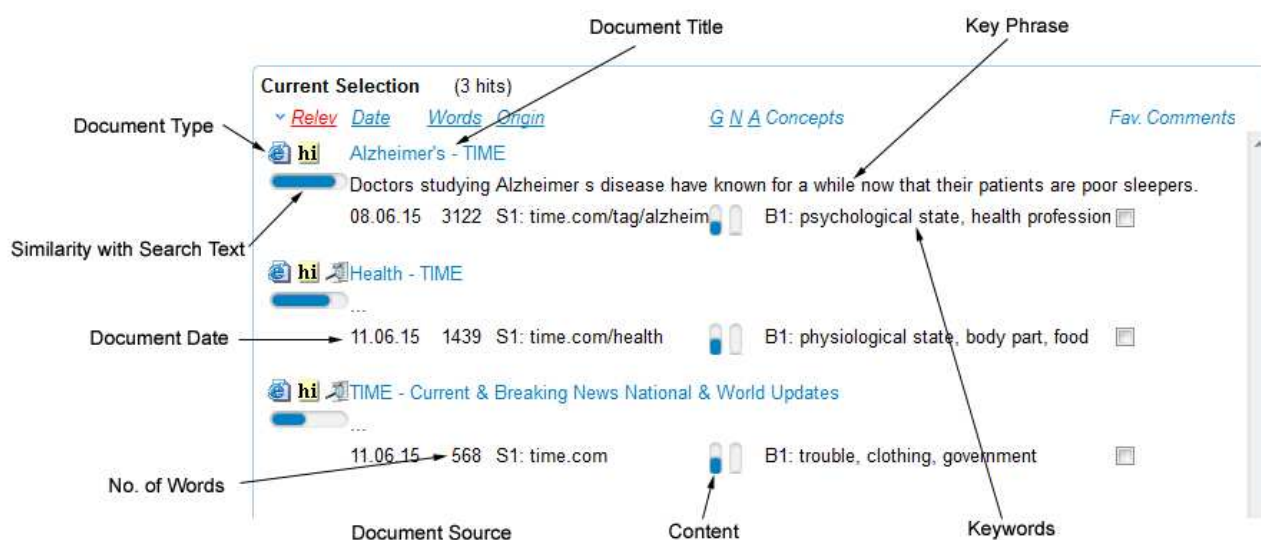


Fig. 3: Detailed View of the Hit List

The following information can be displayed from the search result hit list:

Detailed Document Information

Clicking on the document title displays the following extended information:

- The *Abstract* is an automatically generated summary of those sentences in the document which best reflect the core statements of the document.
- The *Descriptors* consist of up to 18 keywords with which the document can be categorised (displayed in the chosen interface language)
- The *Comment* field can be used to associate any arbitrary text to the document.

The screenshot shows a document entry with the following details:



- Title:** BMW: A Company on the Edge - TIME
- Thumbnail:** A blue bar with a white 'hi' icon.
- Description:** In this video, TIME looks at how the top-selling premium manufacturer BMW is exploring new technology ranging from ...
- Date:** 17.06.15
- Count:** 623
- Source:** S1: time.c...mpany-on-the-edge
- Category:** A2: passenger car, vehicle, enterprise
- Abstract (english):** As part of a strategy, partly overseen by its 49-year-old CEO, Harald Krueger, BMW has been aiming to make 30% more vehicles with the same number of workers while trying to reduce production costs per vehicle by raising economies of scale in components, drive systems and modules. ... In this video, TIME looks at how the top-selling premium manufacturer BMW is exploring new technology ranging from self-driving vehicles to virtual reality in an effort to keep pace with the competition. ... Brands ranging from Toyota to Hyundai are also trying to sell more premium vehicles. ... As part of a strategy, partly overseen by the 49-year-old executive since late-2007, BMW has been aiming to make 30% more vehicles with the same number of workers while trying to reduce production costs per vehicle by raising economies of scale in components, drive systems and modules.
- Descriptors:** A list of 18 keywords: BMW, car manufacturer, premium, vehicle, Mercedes, research and development, technology, consumer, car, fuel efficient, new technology, car rental, Automotive industry, Toyota, Volkswagen, drive system engineering, economies of scale, production cost.
- Author:** ---
- Origin:** <http://time.com/topic/bmw-a-company-on-the-edge/>
- Comments:** A text input field.

Fig. 4: Document Information with Abstract und Descriptors

Original Document

Clicking on the file symbol to the left of the title will, if available, open the original document. If that is not possible, e.g. because it was temporary information or the document does not exist as a single data file (like an email with attachments and threads), an automatically generated substitute file containing the textual content is displayed.

Text Version with Highlighting and Visualisation of the Similarity

Clicking on the symbol  evokes a pure text version of the original document, in which the located search terms are highlighted. By clicking on the blue similarity bar  underneath

the **hi** symbol, the similarity of the document with the search query is highlighted. This representation offers a quick overview over the relevant sections of the document. In addition, one can observe how InfoCodex has evaluated this document in regard to the query.

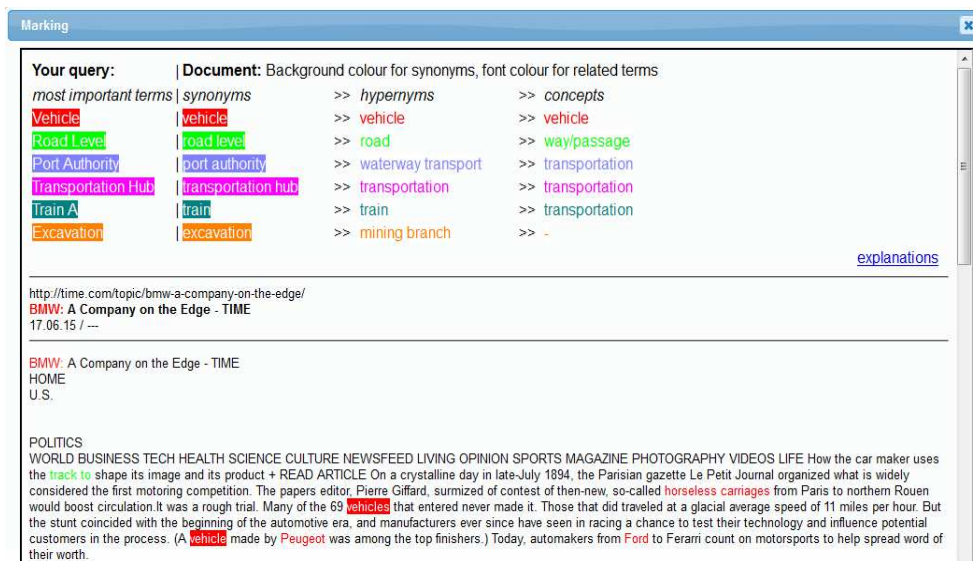


Fig.. 5: Visualisation of the Similarity with the Search Query

3.5 Document Families

InfoCodex can identify documents of very similar content, e.g.

- slightly altered or updated versions of the same document,
- the same document but in a different format (Word, PDF, ...), or
- the same document in a different language.

Only one document per family appears in the result hit list. A special symbol is displayed when a family of documents is present.



Document families have a double function:

- The hit lists are shorter and more overseeable.
- For clean-up operations duplicates can be easily identified and located.

3.6 Alternative Display Possibilities

If a search query returns a very large quantity of results it can become confusingly overwhelming. In these cases there are two further simple ways available to thematically reduce the result list:



Fig 6: Further Display Possibilities

Clustering

With clustering, all located documents are sorted by topic and listed in a concept hierarchy. In this way the result hit list can be quickly and simply further thematically constrained..

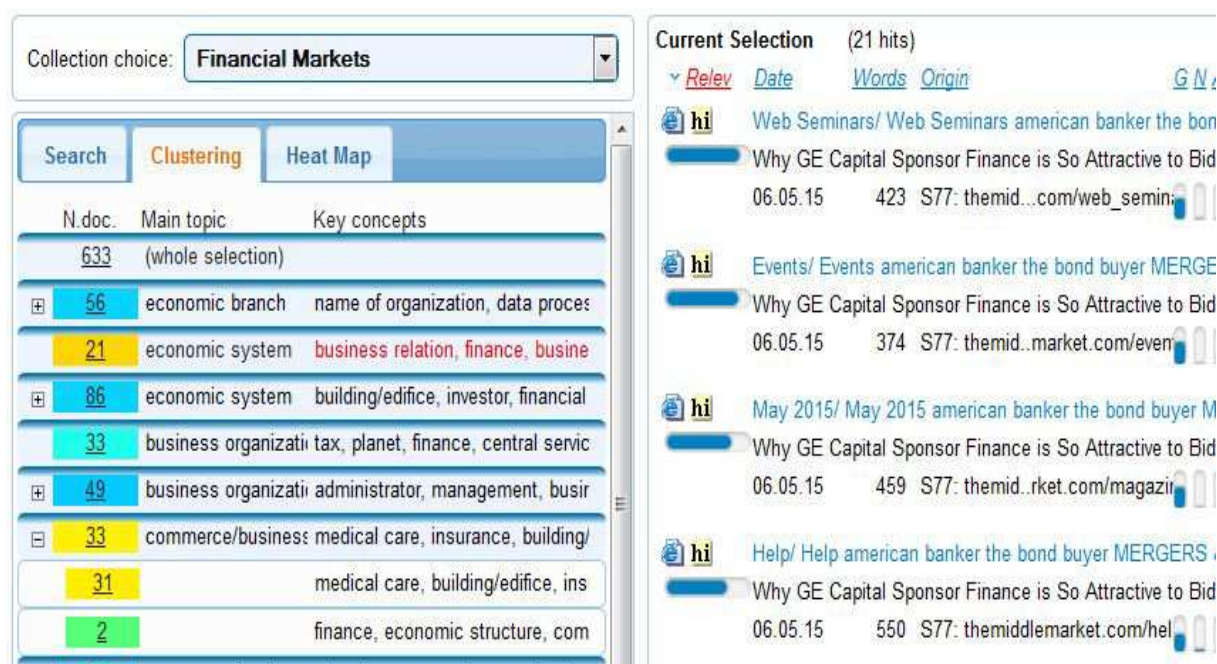


Fig. 7: By clustering (left) the located documents are subdivided into topic areas

Heat Map

The heat map corresponds to the information landscape map (see chapter 3.1), where the individual fields (neurons) are coloured to reflect their relevance to the search query. The heat map provides a quick overview over very large numbers of search results. Importantly, it also shows which topic areas in the collection contain how many relevant documents.

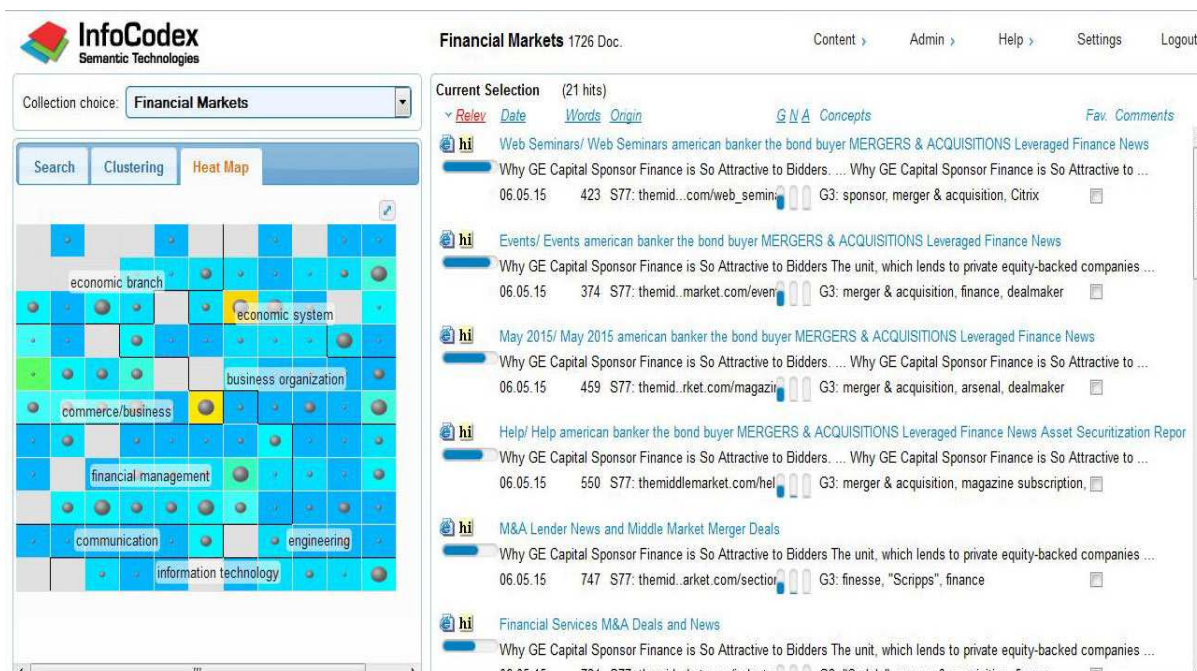


Fig. 8: The heat map (left) offers a quick overview over the main thematic topics of the located documents

4. Creating a New Collection

4.1 Collections and Domains

InfoCodex organises the indexed documents into so-called **collections** (see chapter 3 and 8.1). When documents are included into a collection, they are neither changed nor moved, but solely read and categorised. Only the link (URL) to the original document is stored by InfoCodex for the purpose of viewing.

InfoCodex collections can be structured into distinct **domains**. A domain is a repository of collections and serves mainly the separation of user groups and the allocation of user access rights to confidential or sensitive data.

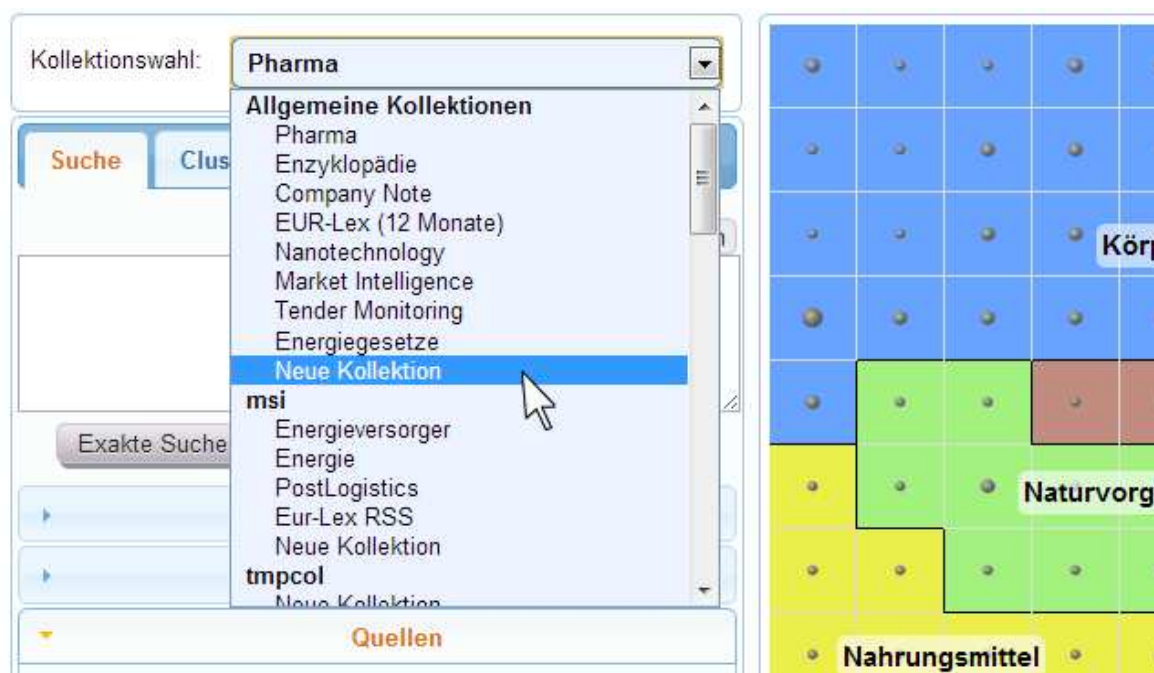


Fig. 9: Collection selection with the domains "General" (main domain), „msi“ and „tmpcol“. The accessible collections and the menu point "New Collection" are listed under each domain

4.2 Setting up a Collection

For users with sufficient access privileges there is a selection option at the bottom of the collection list of each domain for "New Collection". On selecting this item a new window is opened to allow a new collection to be opened.

4.3 Setting the Characteristics of a New Collection

On setting up a new collection, a few options are to be provided:

- Import Mode:
In "update" mode changes in a document will be added to the entry. That is the default case. By unchecking the checkbox, InfoCodex will make a new entry for every change to the document.
- Generation of abstracts, Build document families:
see. chapters 3.4 und 3.5.

- Proximity search, Highlighting:

The proximity search enables the searching of concepts that occur near each other in a document. Also, a simplified text version of all imported documents is stored that enables the colour-highlighted display of the search terms..

Create New Document Collection

Collection name	<input type="text"/>	
Import mode	<input checked="" type="checkbox"/>	Update, i.e. for future incremental loads, documents with the same filename as existing entries will be updated (otherwise the documents would be appended to the collection)
Generation of abstracts	<input checked="" type="checkbox"/>	yes (but it takes some time)
Build document families	<input checked="" type="checkbox"/>	yes (identify similar documents)
Proximity search/highlighting	<input checked="" type="checkbox"/>	yes (enable)
<input type="button" value="Advanced options"/>		
<input type="button" value="OK add documents"/> <input type="button" value="OK save empty collection"/> <input type="button" value="Cancel"/>		

Fig. 10: Creating New Collection

4.4 Advanced Options for Collection Set-up

InfoCodex executes automatically all those options given under section 4.3. If the user wishes to influence the categorisation, the generation of descriptors, the extraction of metadata or the setting of access rights, these can be addressed in the full dialog mask under "Advanced Options".

After the establishment of an empty collection, InfoCodex offers a selection of the various data sources for an import of documents. This procedure is described in detail in chapter 5.

Current collection: Pharma

Setup / Delete Collection

collection-ID:	<input type="text" value="10"/>	status: <input type="text" value="-1"/>	
collection name			
german:	<input type="text" value="Pharma"/>		
english:	<input type="text" value="Pharma"/>		mandatory
french:	<input type="text" value="Pharma"/>		
italian:	<input type="text" value="Pharma"/>		
collection-folder	<input type="text" value="c:\icdata\10"/>		
user-groups	<input type="text" value="1,2"/>		mandatory

Optional classification setups

Classification	
Keyword table	<input type="text"/>
Predefined categories	<input type="text"/>
Metadata instructions	<input type="text"/>
Frontend linguistic database	<input type="text"/>
Indexing	<input checked="" type="checkbox"/> incl. indexing of exact terms
Special options	<input checked="" type="checkbox"/> Generation of abstracts <input checked="" type="checkbox"/> Build document families
Import mode	<input checked="" type="checkbox"/> for identical filenames → update old entry
Map formation	
No. of map columns	<input type="text"/> (no. of col. = no. of rows)
No. of main topics	<input type="text"/> (coarse subdivision of the map)

View data sources	Delete collection	OK save	Quit
--------------------------	--------------------------	----------------	-------------

Fig. 11: Advanced Collection Options

User groups	Restriction of access rights to special user groups
Keyword Table	Table containing especially important words/terms and their relative weights. These keywords influence both the categorisation and generation of descriptors (see User Manual Part 2, Section [5.1])
<i>Predefined categories</i>	Tables which control or influence the categorisation (see User Manual Part 2, Section [5.3])
Metadata instructions	A means for the specification of rules for the extraction of special metadata from documents
Front-end linguistic database	A complement to the InfoCodex database (> 3 million entries), exists the possibility of creating a specialized individual database, including taxonomy. It may e.g. contain such things as company-internal abbreviations or specialist terms. The front-end database takes priority over the InfoCodex database. This can have a significant influence over the categorisation and descriptor generation (see User Manual Part 2, Sect. 5.2)
Import mode	When incrementing an existing collection, in the default case, documents are added to the collection even if they were already imported. By checking the special option "Import mode" the registered filename will not be duplicated and only the content will be updated.

5. Adding Documents

5.1 Selection of Data Sources

For any specific collection, a maximum of 200 data sources may be specified. All documents from the selected data sources will be included, in as far as InfoCodex can analyse the documents it finds.

The originals are not copied, but only read, analysed and indexed. The metadata as well as the addresses of the documents are assimilated into the InfoCodex database. The size of the depository of the documents is unrestricted; it has no influence on the “virtual order” that InfoCodex creates.

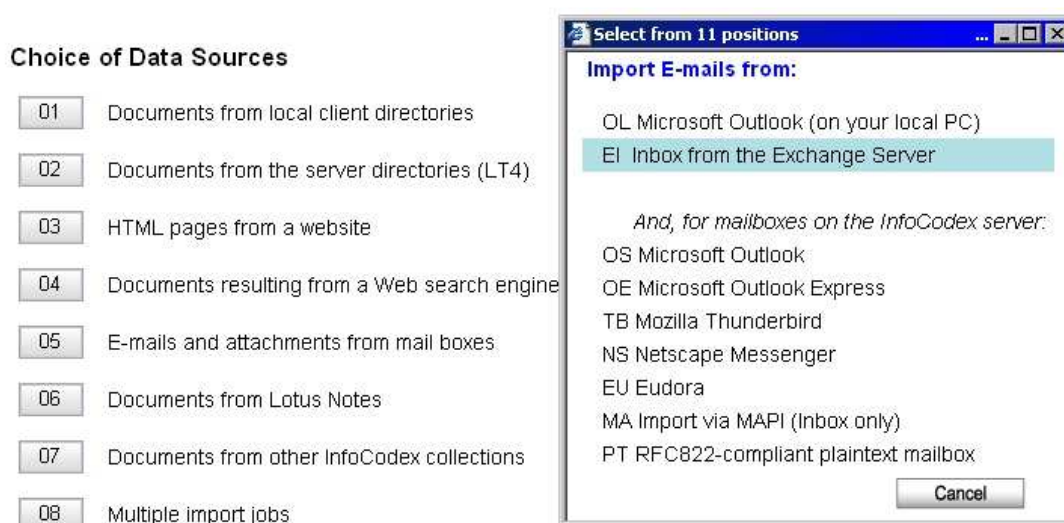


Fig. 12: Choice of Data Sources for Document Import

Documents from one's own client

The InfoCodex application and the browser have no direct access to the documents on the local client machine (for security reasons). To transfer documents from the local client into InfoCodex, an installation of the “file-transfer plugin” is required. This can be downloaded from the InfoCodex start page.

The installation requires system administrator privileges on the client. After the installation one should also ensure that the user has the required access rights to execute the installed programs. When importing from a local client, a directory (folder) is selected and all files (including subdirectories) are assimilated into InfoCodex.

Documents from the server (incl. network)

This is the usual method used to import documents from an internal network. In as far as the user has the corresponding rights, it is possible to access all disk drives or even an entire network.

Websites

All that is needed here is to enter the desired web address, e.g. follows all links to a specified and downloads the located pages for content analysis. All file types supported by InfoCodex can be imported, or, if required, each data source can be restricted to include only certain file formats.

Provision to enter a password is available for the case of password-protected websites.

Results from web search engines

After selecting a search engine (e.g. Google) an input mask appears where the search text is entered in exactly the same way as one would in the search engine itself. InfoCodex downloads all the documents in the resulting hit list for content analysis purposes. It is important to note that most search engines only yield at most 1,000 documents, even though the displayed number of hits is way over this number.

The list of available search engines can vary depending on the installation.

A password can also be submitted for password protected search engines (e.g. Web of Science).

Email and Attachments

InfoCodex supports various mailbox formats like Outlook, Outlook Express, Thunderbird, Eudora etc. (see above). The content of attachments, including ZIP files, are also analysed and InfoCodex treats them as one document .

Before choosing

OL Microsoft Outlook (on one's local PC)

or

EI Inbox from Exchange Server

Outlook must be installed on the server. In the second case, access to the central Exchange Server must be established and the user must give share access to InfoCodex .

The import from local Outlook (**OL**) cannot be performed in batch mode, because the user must confirm permission to access his mailbox every few minutes. It is therefore advantageous to send the local Outlook mails as an Outlook Archive (**PST file**) to the server, and from there to import. InfoCodex recognizes such PST archives and can automatically extract the contained elements.

Documents from Lotus Notes

In order to dock onto Lotus Notes, a Notes client and the InfoCodex module Notes Connect must be installed on the server.

At import, Lotus Notes documents are first selected via Notes Connect and saved temporarily as index files in the NotesConnect directory for processing.

Documents from SharePoint

Documents from SharePoint can be simply imported by providing the URL of a document library. If required, a username and password can also be supplied.

Documents from other InfoCodex collections

If necessary, documents may be used from other existing InfoCodex collections by means of the result of a search text (analogous to the results from web search engines).

Multiple import jobs

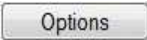
With this option, multiple import jobs like various search engine queries along with documents from websites and local directories can be simultaneously executed. This is an ideal agent for information research.

HTML Pages from a List

This function allows an import of documents from the internet via a text file containing a list of URLs.

5.2 Import execution (Content Analysis / Indexing)

After selecting the data sources to be included into the collection, the automatic execution of the import is started with **"OK Start import"**.

Prior to starting the import, some special adjustments can be set with  e.g. a maximum on the number of documents to be imported or the activation of an OCR interpreter for scanned documents.

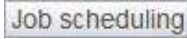


The import consists of the following processes:

1. The InfoCodex spider agents gather the documents; convert these to temporary text files and extract the metadata from the documents (author, title, date. etc).
2. Language recognition: English / German / French / Italien / Spanish; lexical and semantic analysis.
3. True cross-language content analysis, correlation with the taxonomy
4. Construction of a 100 dimensional content space based on the collection and the projection of the documents into this space.
5. Categorisation of the documents by means of a neural network technique (Kohonen map) ↓ Placement in a library bookshelf.
6. Tagging the documents: Generation of descriptors.
7. Indexation for efficient searching.

These processes run in the background. The user need not follow these processes and can concentrate on his/her other tasks..

Job Scheduling

An alternative to the direct execution of an import, import instructions can be set up as batch jobs using . Starting time and periodicity of the batch jobs can also be specified in the dialog mask. The batch instructions can be viewed, edited or deleted at any time:

→ Admin → C2 Display/manage data sources.

Typical uses

Competition monitoring:	Any accessible information over the latest activities of the competition is collected every night. This information is available in the morning in an ordered and clearly arranged form.
-------------------------	--

Updating large collections:

Reload afresh ("Initial Load") the documents of the local network over the weekend. In the meantime, from Monday to Thursday the newest documents that are up to one day are added overnight ("Incremental Load").

Import Options

The following additional import options are available:

Collection: Financial Markets

Imports in Queue

1. Copy website "http://www.news.com.au/finance/markets" (upto 100 doc. from www...

Maximum age of documents in days ,e.g. 30 or 30+
(**30+ means:** include documents with an unknown date)

☐ Save documents as TXT files

☐ Remember non-convertible files

☐ Apply content filter on HTML files

Min. chars per text block for content filter

Group authorization / OCR activation

Import job

Authorized user groups

(:T/20) No. jobs

OCR

☐

OK

Cancel

Maximum age	Maximum age of the documents in days. Only newer documents are imported, which shortens the processing time, especially for periodic "Incremental-Loads".
TXT files	Store documents temporarily as plain text.
Remember non-convertible files	- The system creates a blacklist of non-convertible files which can significantly reduce the processing time for incremental loads of large collections.
Content filters	<p>Beside the actual content, HTML pages contain a lot of additional information like navigation instructions and advertisement. A content filter is useful here in that the actual content text blocks are usually larger than the text blocks of the additional information.</p> <p>A content filter can be assigned per data source by an entry in a CONFILT row (e.g. CONFILT=60) in the instruction file of the batch job.</p>
Authorization	Import jobs consisting of multiple data sources can be assigned independent authorizations for each source.
OCR	Switch on OCR text recognition for graphic documents. In the default case, InfoCodex uses MicroSoft's OCR engine, which is part of MicroSoft Office (since Version 2003). Support for other OCR engines (OmniPage, FineReader) is optionally possible.

6. Analysing Content

The analysis functions are found in the **Content** menu.

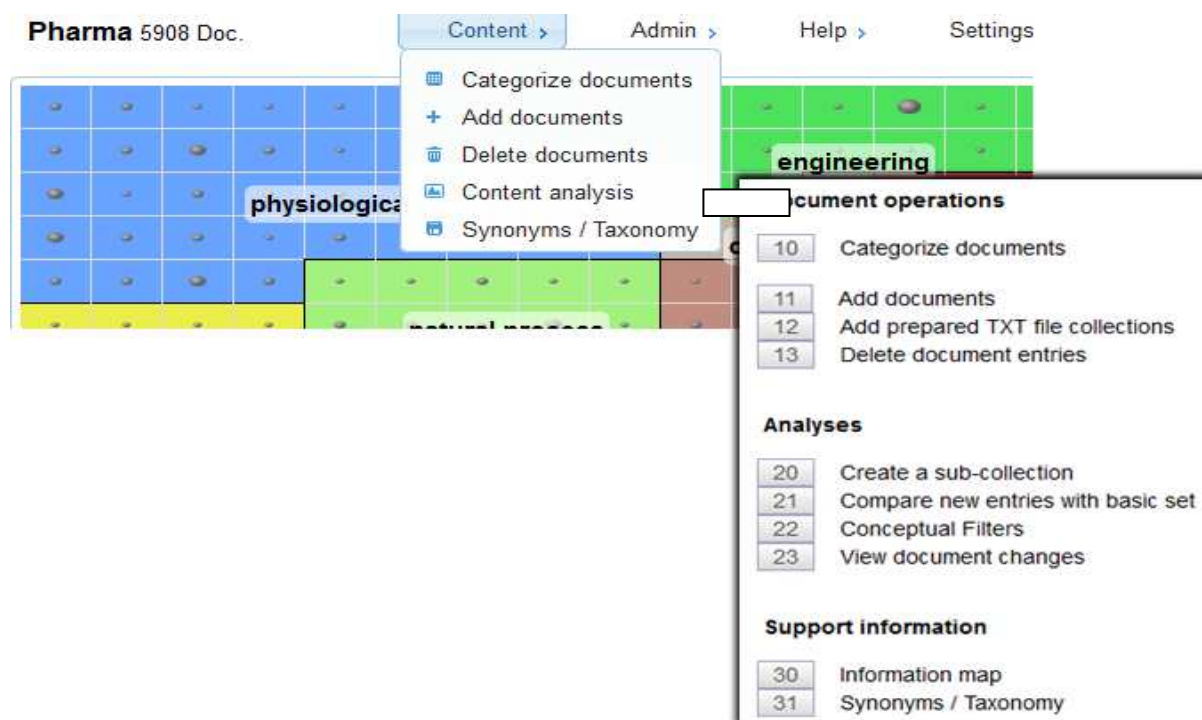


Fig. 13: Content: The actual analyses are available under the point **Content analysis**.

6.1 Categorising Documents

With this function, documents from selectable data sources are allocated to specific compartments of the information map on the basis of their content. Documents within a particular compartment have similar content, and so such an allocation means a *categorisation based on a content perspective*.

This function is also the foundation for Response Management: Incoming documents are categorised and, depending on their allocated category, trigger certain given processes by means of a prepared decision matrix.

Examples of use:

- Automatic redistribution of emails that come into one central address.
- Automatic preprocessing of customer enquiries / complaints with the generation of a suggested reply based on pre-designed text modules.
- Searching for standards and rules for certain problem situations.

Along with the categorization, an automatic tagging of the documents with 18 descriptors takes place. These descriptors, which are generally significant, can be used for the tagging of documents for the incorporation into archives.

Procedure:

1. Start the function "Categorise documents"
2. Select the data sources (the same way as described in section 4.2)
3. Output the results to screen or into an Excel spreadsheet (as a selectable set of document characteristics and descriptors)

By using this function, the analyzed documents are not integrated into the current collection.

6.2 Add Documents

Use this menu point to add new data sources to the current collection; see section 5.1.

6.3 Delete Document Entries

This function allows documents to be removed from the current collection. Only the references to the documents are removed; the original documents remain untouched.

Method:

1. Create a search that will locate the undesired documents.
2. Save the search.
3. The actual documents to be deleted from the search can then be filtered out with a detailed selection.
4. The references to the selected entries are then removed from the collection.

6.4 Create a Sub-Collection (see “Analyses“)

With this function it is possible to create a subset from an existing collection. It can be created from a selection of the main topics; by a saved set of search results; or a combination of both.

6.5 Compare New Entries with Basic Set / Trend Recognition

This function enables the early recognition of changes, e.g. for:

- Competition monitoring / Market observation
- Intelligence services
- Recognition of new trends and technologies

Typically, a collection is based on a particular theme (e.g. a combination of several internet searches). Documents are appended to the collection at periodic intervals.

By comparing the new entries with the existing basis collection the most important changes can be quickly and reliably recognised.

Using the three available evaluation techniques: “Thematic Shifts“, “Heat Map“, “New Subjects“, changes are exposed at the click of a button. For example, with “New Subjects“ those documents are revealed that contain real, new facts compared with the previous documents.

Sydney Morning Herald : Analyse New Entries

The content characteristics of new entries are compared with those of the basic set (thematic shift).

Selection of the basic set and the new entries

Import Date	No. of Documents	Basic Set	New Entries
30.06.15	50	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1.07.15	47	<input type="checkbox"/>	<input type="checkbox"/>
2.07.15	47	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Possible Evaluations

Thematic Shift	List the displacements of the thematic characteristics
Heat Map	The fields (neurons) having the largest relative increments are colored in red
New Subjects	Document list ordered by largest deviations from previous content

Fig. 14: Analysing new entries; display thematic shift

7. Synonyms / Taxonomy

This function gives an insight to the linguistic infrastructure of InfoCodex. It is particularly useful to view the scope of synonym groups and for clarifications in cases of seemingly strange results (in particular, when a document receives an odd descriptor, which at first glance seems irrational or illogical).

N.B: Individual words in a language often have more than one meaning, and the translation into another language and the classification in a synonym group and the binding into the taxonomy is inevitably subjective. The total classification of a document through the neural network is not problematic. Humans also experience some ambiguity reading and ingesting document content.

Synonyms → *show similar words*

enter word, which synonyms are to be displayed in 3 selectable languages

Display in ☒ English ☒ German ☒ French ☐ Italian

word English

or German

or

Hierarchy of meanings → *classification system (taxonomy)*

enter key to see its classification (nothing = overview)

term English

or German

Abb. 15: Query mask for synonyms and hierarchy of meanings

7.1 Synonym Groups

Using the upper part of the mask (**Synonyms**), one can see to which synonym group (*semantic cloud*) a word belongs and where it is ordered in the concept hierarchy. The words are displayed side by side in 3 selectable languages .

A synonym group combines several terms with the same meaning, e.g. “Fahrausweis“, “Führerschein“, “driving licence“, “permis de conduire“ etc.

internet >> internet >> computer network >> information technology >> INFORMATION/ COMMUNICATION

English

internet, internet world, internets
Internet of

German

Internet, Internet 2, Internetz
Internetwelt, Internet der
internetadäquat, internetbasierende
internetbasiert, internetbezogen
internetbezogene, internetgestützt
internetgestützten, internettauglich

French

Internet, Internet des
internet, internet haut
l'internet

Fig. 16: Related terms in English, German and French

7.2 Hierarchy of Meanings

Using the lower part of the mask (**concept hierarchy**) one can view the entire classification system (**Taxonomy**), in fact, top-down to the very synonym groups .

If no constraints are entered, InfoCodex displays the uppermost nodes of the taxonomy tree in both English and German.

If a front-end database has been installed, those structures are also displayed. In this way, the user can check if terms belong to the front-end database or to the standard InfoCodex database.

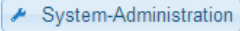
The linking of synonym groups with the taxonomy tree, i.e. the allocation to hypernyms to synonym groups, forms a special *ontology*, correlating the most important meaning of words.

Informatik >> INFORMATION/ KOMMUNIKATION

0	• Informatik	• information technology
1	• Software	• software
1.01	• Business-Software	• business software
1.0101	• Adressbewirtschaftung	• address management
1.0102	• CAD	• computer-aided design
1.0103	• Marketing-Datenbank	• database marketing
1.0104	• Finanzsoftware	• financial software
1.0105	• HR software	• HR software
1.0106	• Wissens-Management	• knowledge management
1.0107	• Suchtechnologie	• search technology
1.0108	• Auftragsbewirtschaftung	• order processing
6.01	• Kommunikations-Software	• communication software
6.02	• Identifikationskarte	• identification cards
6.03	• Computer-Netzwerk	• computer network
6.04	• Bilddaten	• image data
7	• Computer-Netzwerk	• computer network
7.01	• WLAN	• WLAN
7.02	• Internet	• internet
7.03	• Intranet	• intranet
8	• Peripheriegerät	• peripheral equipment
8.01	• Datenspeicherungsgerät	• data storage device
8.0101	• Disk	• disk
8.02	• Verbindungsausrüstung	• interconnection equipment

Fig. 17: Concept hierarchy for the area of "Information/ Communication" , to which the concept "internet" is assigned.

8. System Management (Summary of Part 2 of the Manual)

The system management functions of InfoCodex are accessible with the  menu point. They are divided into:

User Administration	Permitted users and user groups; data protection features; LDAP interface to central user administration; setup of autonomous IC domains
System Administration	Administration of collections; influence over categorisation; system settings



Fig. 18: "Admin" menu

These functions are described in Part 2 of the user manual. Only those features that are also relevant to users are outlined in the following section .

For normal users only some of the functions are available. For example, only the function "Change password", appears under "User administration".

8.1 Data Protection / IC Domain Concept

Data protection is implemented in three different levels:

User Groups

Every InfoCodex user is a member of one or more user groups. Access rights to domains and collections can be set per user and group. The actual rights of a user are determined by his own rights and the rights of all the groups he belongs to.

File System Security

When this option is activated, a user can only find and view documents from data sources in the network that he has access to independently from InfoCodex access rights. All access to files in the context of the user are governed by the access control and protocol of the underlying operating system.

InfoCodex Domains

Access to a collection is determined by the domain in which the collection exists. In this way, for example, a protected domain for the business administration can be set up to which not even the IT staff have access.

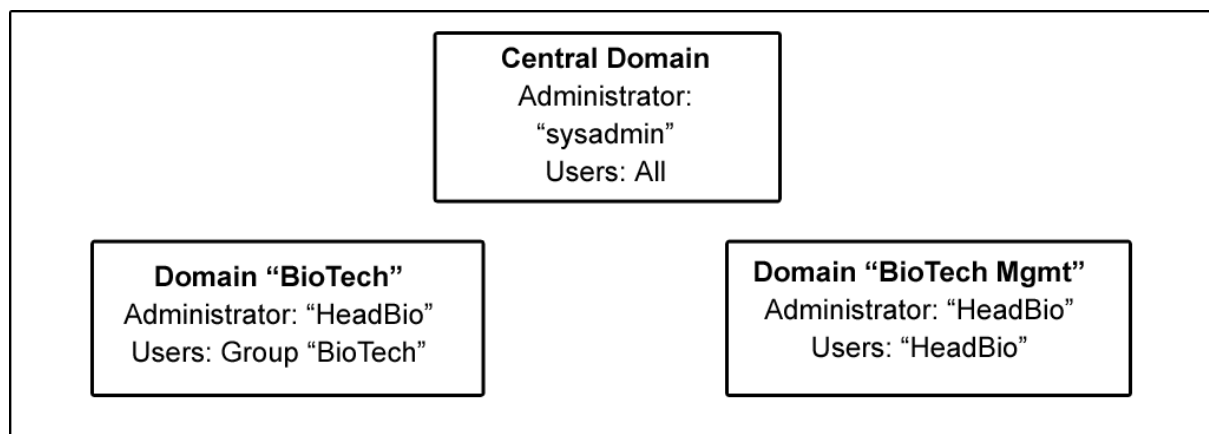


Fig. 19: InfoCodex domains with different user groups

The individual IC domains can only be created by the system administrator of the main domain. A domain administrator who will have full sovereignty over the domain is assigned on domain creation. Even the system administrator has no access to the new domain unless the domain administrator explicitly grants it. In the normal case, the system administrator has only the permission to delete the collection.

8.2 Collection Administration

These functions apply to the setting up and maintenance of the individual collections (see Section 4). They are only available to those users with the corresponding privileges.

Special functions:

Delete collection	Removal of entire collections from the system.
Viewing status	Viewing processing status of a collection (esp. in the import and analysis phases); viewing the processing report
Job Scheduling	If the import of documents is controlled via batch jobs (vgl. Section 5.2), the corresponding scripts can be viewed and modified under the menu point "Display/manage data sources".

8.3 Influencing the Categorisation

The categorisation of the documents (shelving in a virtual book cabinet) and allocation of descriptors to individual documents is substantially determined by the InfoCodex linguistic database and taxonomy. The database contains over 3 million terms in English, German, French, Italian and Spanish, and covers largely all fields of knowledge.

It may, however, be desired that the user introduce his own accents and, for example, special technical expressions or company-internal abbreviations. In this way the categorisation and allocation of descriptors are significantly influenced.

The following possibilities are available:

Keyword Setting

A list of words/expressions can be defined whose significance is heavily weighted at the content analysis phase.

Front-end Database

A customer-specific linguistic database (with links into the taxonomy tree) can be prepared containing e.g. special technical expressions or company-internal abbreviations. Such a front-end database takes priority over the standard InfoCodex database. If a word is encountered during the content analysis phase that occurs in the front-end database, then its meaning and significance, instead of that from the InfoCodex database, comes into effect.

Predefined Categories

This is a means to influence the grouping into the information map (i.e. the "book cabinet").. In many cases, the simple and interactive "Ad hoc categorisation" can achieve very much. A further possibility exists to prescribe fixed categories and to neutralise the automatism of the neural network. In this way, for example, a company-internal classification could be integrated into InfoCodex.

9. Selected Examples

9.1 Information Research

Problem description

There should be a periodic search for findings on the theme "secondary contamination in raptors after the usage of poison baits against *Arvicola terrestris*".

Search terms: rodenticide, Bromadiolone, raptor, *Arvicola terrestris*

Procedure

1. Create a new collection.
2. Add documents: Menu point "Results from a web search engine", e.g. Google
3. Enter search terms: rodenticide, Bromadiolone, raptor, "*Arvicola terrestris*"
4. Click on "Preview". The Google search results in zero hits.
5. Click on "Extended queries" and select suitable synonyms to the individual search terms (perhaps, also in other languages).
6. Start the import.

This will generate multiple Google searches with different synonym combinations.

Refinements

Perform the first search as described above. Afterwards, set up a batch job for the same same query (Job scheduling):

- weekly execution,
- use option "2 = Add documents to the existing collection".

9.2 Finding Available Know-how

Problem description

A researcher wants to open up a new research field. During his research he stumbles across an interesting document (e.g. a paper). He wants to find out what knowledge is already available in his own research institute which could be relevant to this specific paper.

Procedure

1. Copy the contents of the paper into the clipboard (Word document, PDF, HTML or even email).
2. Select the collection "Internal Network" in InfoCodex.
3. Paste the clipboard contents into the search field text area.
4. Start the search.

Result

InfoCodex retrieves the documents available on the local network most similar to the given paper. The terms from the paper are not necessarily present in the document; because the search was for thematic content similarity.