Version 5.1 of August 2015

1. Webpart für die Integration von InfoCodex in Sharepoint

1.1 Zweck des InfoCodex-Webparts

Das Auffinden von Dokumenten in Microsoft Sharepoint ist bei kleinen Dokumenten-Kollektionen einfach und klar. Mit zunehmender Grösse wird die Suche langsamer und ineffizient, und es ist schwierig, in einer grossen Menge eine Übersicht zu gewinnen und gezielt die effektiv interessierenden Dokumente zu finden.

Der Einbau des InfoCodex-Webparts in Sharepoint bringt folgende Mehrwerte:

a) Übersichtliche Gliederung der Dokumente und Visualisierung

Die Dokumente werden nach thematischen Gesichtspunkten automatisch gegliedert und in einem "Bücherregal" angeordnet. Verwandte Artikel erscheinen nebeneinander im gleichen Fach. Das grafisch dargestellte "Bücherregal" bietet eine rasche Übersicht und ermöglicht eine gezielte Suche.

Der Anwender steht nicht mehr vor einer schwarzen Wand.

b) Ordnen der Treffer nach Relevanz

Die Dokumente, die der Suchabfrage am ähnlichsten sind, erscheinen zuoberst. Die Relevanz (= objektives, patentiertes Ähnlichkeitsmass) wird grafisch angezeigt.

c) Sprachübergreifende, semantische Suche

Neben der Suche nach fixen Begriffen kann auch nach Synonymen gesucht werden. Dies ist besonders dann wichtig, wenn man Dokumente sucht, *die man nicht selbst geschrieben hat*. Andere Autoren bevorzugen unter Umständen andere Begriffe für dasselbe Subjekt (z.B. Wechselfieber statt Sumpffieber oder Malaria bzw. EWR statt Europäische Wirtschaftsgemeinschaft).

Mit der *Ähnlichkeitssuche* kann gar mit freien Textblöcken oder ganzen Dokumenten gesucht werden. Es brauchen keine Begriffe aus der Suchabfrage mit Begriffen aus den Dokumenten übereinzustimmen. Es kommt nur noch auf die inhaltliche Ähnlichkeit an.

Die Synonymsuche und die Ähnlichkeitssuche sind *sprachübergreifend*. Mit deutschen Suchabfragen werden auch die entsprechenden englischen, französischen, italienischen oder spanischen Dokumente gefunden. (Das geht wesentlich über die üblichen "mehrsprachigen" Systeme hinaus, die einfach die Sprache erkennen und bestenfalls die sprachabhängigen Endungen und Konjugationen kennen.)

d) Automatische Zusammenfassungen und Keyword-Bildung ("Tagging")

Für jedes Dokument wird automatisch eine aussagekräftige Zusammenfassung erstellt und es werden standardisierte Keywords extrahiert ("Tagging"). Dies erleichtert ein rasches Durchforsten der gefundenen Dokumente.

Ausserdem enthalten die Kurzinformationen in den Trefferlisten echte Inhalte (Titel, Kernaussage).

e) Heatmap-Darstellung und Bündelung ähnlicher Dokumente

Dies verschafft einen *raschen grafischen Überblick*, und die Bündelung *verkürzt die Trefferlisten ohne Informationsverlust*. Daraus ergeben sich signifikante *Effizienzsteigerungen* bei der Informationsbearbeitung.

f) Integration von verteilten Informationsquellen

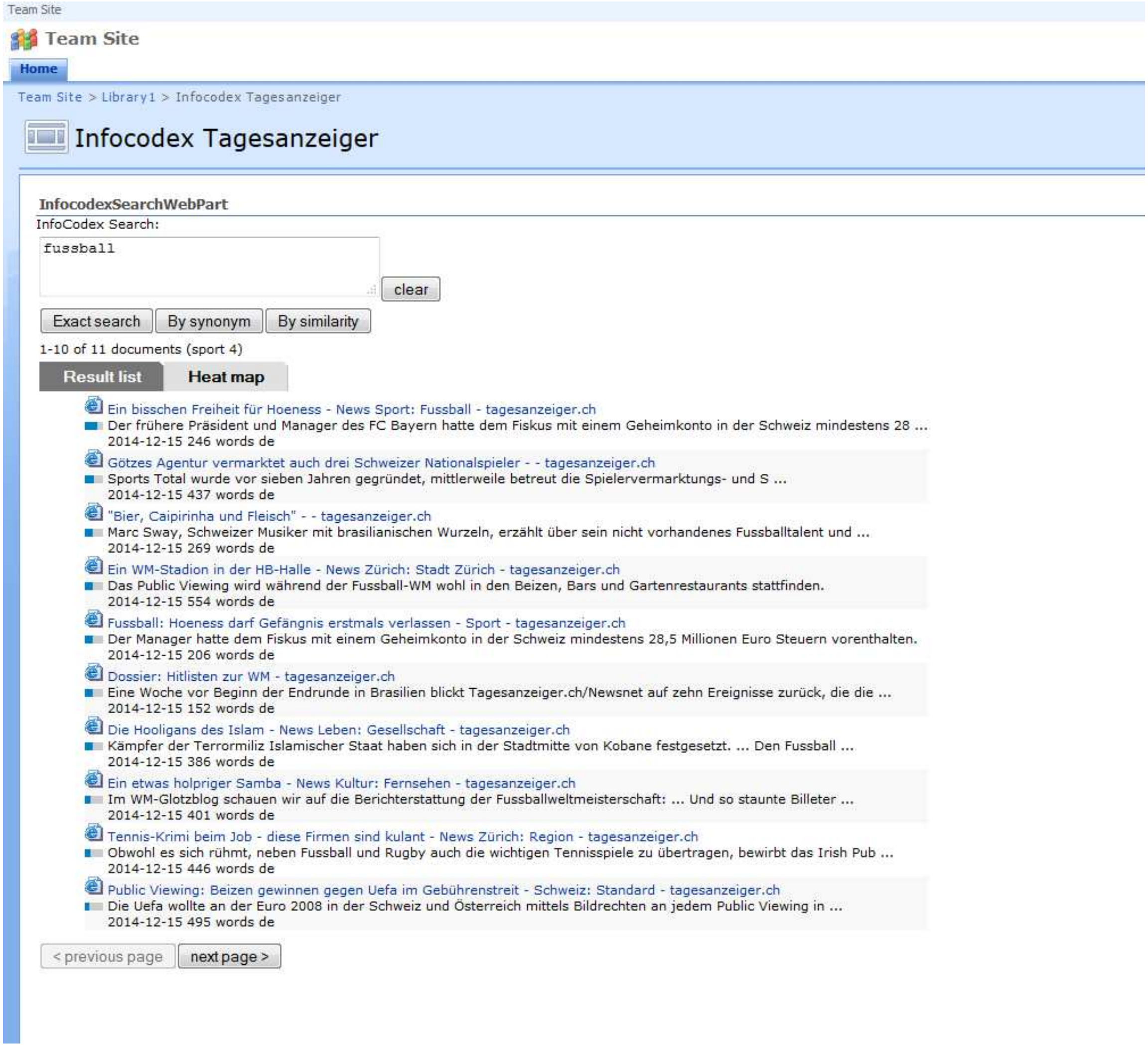
Mit dem InfoCodex-Webpart kann innerhalb ein und derselben Kollektion auf breit gestreute Quellen zugegriffen werden (einschliesslich Internet-WebSites und Suchresultate von Google etc.).

1.2 *Suchen und Finden in Sharepoint mit dem InfoCodex-WebPart*

a) <u>Suchabfrage und Resultatliste</u>

Im vorliegenden Beispiel steht eine erweiterte Suchmaske zur Verfügung, mit welcher neben traditionellen Keyword-Abfragen ("Exact search") auch Semantische Suchen ("By synonym") oder gar die Ähnlichkeits-Suchen ("By similarity") unterstützt werden. Im letzten Falle wird ein freier Textblock (z.B. ein E-Mail-Text) als Abfrage in das Suchfeld hineinkopiert, und es werden die Dokumente gesucht, die dem Suchtext inhaltlich am besten entsprechen.

Resultatliste sortiert nach absteigender Relevanz bezüglich der Suchabfrage:



InfoCodex Search: fussball

1-10 of 11 documents (sport 4)

Ein bisschen Freiheit für Hoeness - News Sport: Fussball - tagesanzeiger.ch
Der frühere Präsident und Manager des FC Bayern hatte dem Fiskus mit einem Geheimkonto in der Schweiz mindestens 28 ...
2014-12-15 246 words de

Götzes Agentur vermarktet auch drei Schweizer Nationalspieler - - tagesanzeiger.ch
Sports Total wurde vor sieben Jahren gegründet, mittlerweile betreut die Spielervermarktungs- und S ...
2014-12-15 437 words de

"Bier, Caipirinha und Fleisch" - - tagesanzeiger.ch
Marc Sway, Schweizer Musiker mit brasilianischen Wurzeln, erzählt über sein nicht vorhandenes Fussballtalent und ...
2014-12-15 269 words de

Ein WM-Stadion in der HB-Halle - News Zürich: Stadt Zürich - tagesanzeiger.ch
Das Public Viewing wird während der Fussball-WM wohl in den Beizen, Bars und Gartenrestaurants stattfinden.
2014-12-15 554 words de

Fussball: Hoeness darf Gefängnis erstmals verlassen - Sport - tagesanzeiger.ch
Der Manager hatte dem Fiskus mit einem Geheimkonto in der Schweiz mindestens 28,5 Millionen Euro Steuern vorenthalten.
2014-12-15 206 words de

Dossier: Hitlisten zur WM - tagesanzeiger.ch
Eine Woche vor Beginn der Endrunde in Brasilien blickt Tagesanzeiger.ch/Newsnet auf zehn Ereignisse zurück, die die ...
2014-12-15 152 words de

Die Hooligans des Islam - News Leben: Gesellschaft - tagesanzeiger.ch
Kämpfer der Terrormiliz Islamischer Staat haben sich in der Stadtmitte von Kobane festgesetzt. ... Den Fussball ...
2014-12-15 386 words de

Ein etwas holpriger Samba - News Kultur: Fernsehen - tagesanzeiger.ch
Im WM-Glotzblog schauen wir auf die Berichterstattung der Fussballweltmeisterschaft: ... Und so staunte Billeter ...
2014-12-15 401 words de

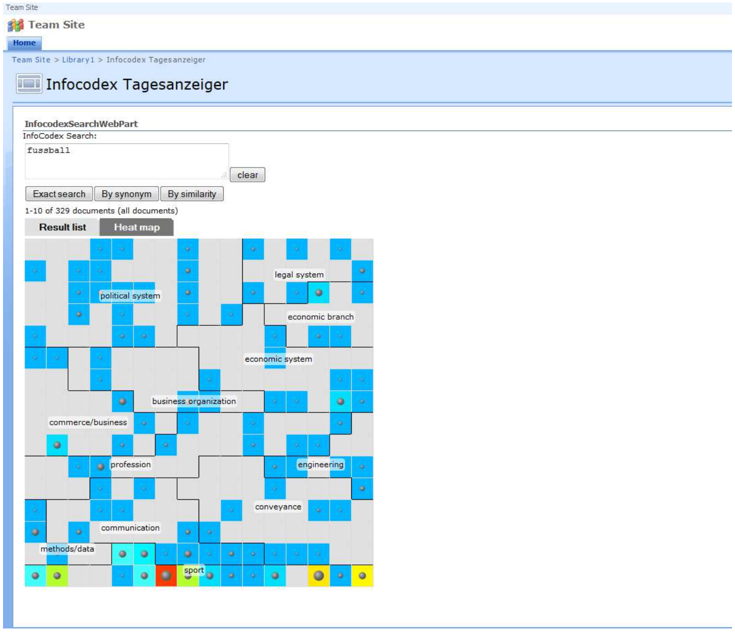
Tennis-Krimi beim Job - diese Firmen sind kulant - News Zürich: Region - tagesanzeiger.ch
Obwohl es sich rühmt, neben Fussball und Rugby auch die wichtigen Tennisspiele zu übertragen, bewirbt das Irish Pub ...
2014-12-15 446 words de

Public Viewing: Beizen gewinnen gegen Uefa im Gebührenstreit - Schweiz: Standard - tagesanzeiger.ch
Die Uefa wollte an der Euro 2008 in der Schweiz und Österreich mittels Bildrechten an jedem Public Viewing in ...
2014-12-15 495 words de

Pro Treffer werden folgende Informationen angezeigt:

- Dokumentensymbol (hier: HTML-Dokument)
- Titel des Dokuments
- Darunter der "wichtigste" Satz des Dokuments (Kernaussage)
- Dokument-Datum und Anzahl Wörter
- Ähnlichkeitsbalken (unterhalb des Dokumentensymbols); gibt die Relevanz des Treffers an (Ähnlichkeit zwischen dem Dokumenteninhalt und der Suchabfrage)

b) Visualisierung der Suchresultate



InfoCodex Search:

fussball

1-10 of 329 documents (all documents)

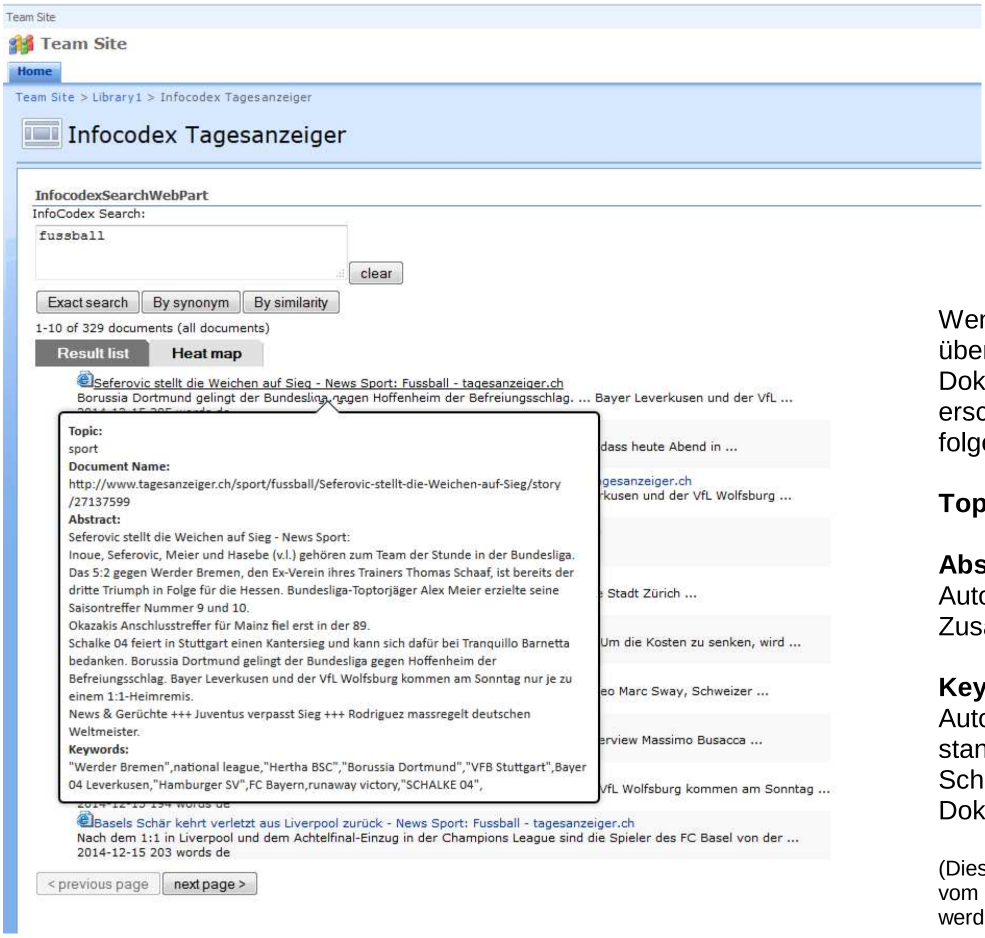
Die Suchresultate können in einer

Heatmap

dargestellt werden. Die rot gefärbten Kästchen enthalten die ähnlichsten Dokumente ("hot spots"), während die blauen Kästchen wenig ähnlich zur Suchabfrage sind.

Bei grossen Resultatlisten ermöglicht diese Visualisierung ein rasches und gezieltes Auffinden der relvanten Dokumente.

c) Anzeigen von Zusatzinformationen (Abstract, Keywords, Thema)



InfoCodex Search:

fussball

1-10 of 329 documents (all documents)

Seferovic stellt die Weichen auf Sieg - News Sport: Fussball - tagesanzeiger.ch
Borussia Dortmund gelingt der Bundesliga gegen Hoffenheim der Befreiungsschlag. ... Bayer Leverkusen und der VfL ...

Topic:
sport
Document Name:
http://www.tagesanzeiger.ch/sport/fussball/Seferovic-stellt-die-Weichen-auf-Sieg/story/27137599
Abstract:
Seferovic stellt die Weichen auf Sieg - News Sport:
Inoue, Seferovic, Meier und Hasebe (v.l.) gehören zum Team der Stunde in der Bundesliga.
Das 5:2 gegen Werder Bremen, den Ex-Verein ihres Trainers Thomas Schaaf, ist bereits der dritte Triumph in Folge für die Hessen. Bundesliga-Toptorjäger Alex Meier erzielte seine Saisontreffer Nummer 9 und 10.
Okazakis Anschlusstreffer für Mainz fiel erst in der 89.
Schalke 04 feiert in Stuttgart einen Kantersieg und kann sich dafür bei Tranquillo Barnetta bedanken. Borussia Dortmund gelingt der Bundesliga gegen Hoffenheim der Befreiungsschlag. Bayer Leverkusen und der VfL Wolfsburg kommen am Sonntag nur je zu einem 1:1-Heimremis.
News & Gerüchte +++ Juventus verpasst Sieg +++ Rodriguez massregelt deutschen Weltmeister.
Keywords:
"Werder Bremen",national league,"Hertha BSC","Borussia Dortmund","VFB Stuttgart",Bayer 04 Leverkusen,"Hamburger SV",FC Bayern,runaway victory,"SCHALKE 04",

dass heute Abend in ...

gesanzeiger.ch
rkusen und der VfL Wolfsburg ...

Stadt Zürich ...

Um die Kosten zu senken, wird ...

eo Marc Sway, Schweizer ...

erview Massimo Busacca ...

VfL Wolfsburg kommen am Sonntag ...

Basels Schär kehrt verletzt aus Liverpool zurück - News Sport: Fussball - tagesanzeiger.ch
Nach dem 1:1 in Liverpool und dem Achtelfinal-Einzug in der Champions League sind die Spieler des FC Basel von der ...
2014-12-15 203 words de

Wenn man mit der Maus über den Titel eines Dokuments fährt, erscheint ein Fenster mit folgendem Inhalt:

Topic (Hauptthema)

Abstract
Automatisch generierte Zusammenfassung

Keywords ("Tags")
Automatisch extrahierte, standardisierte Schlüsselwörter für das Dokument

(Diese Informationen können vom System weiterverwendet werden, z.B. Tagging etc.)

2. Concept Filters for Targeted Market Intelligence

2.1 Objective

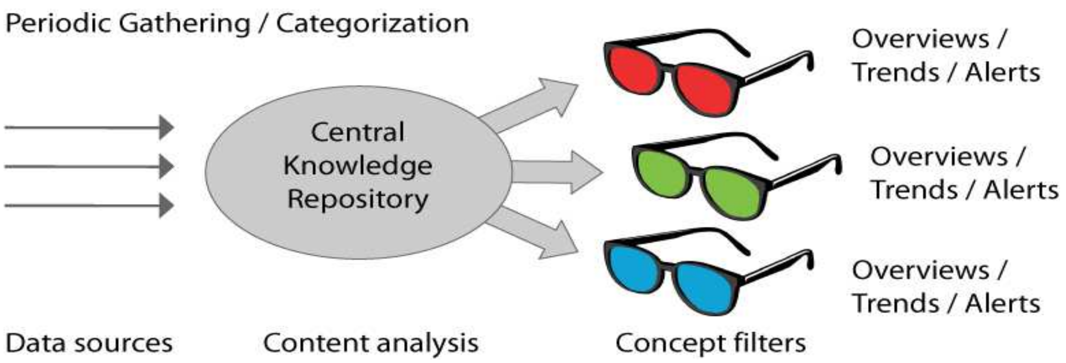
Spidering out into the Web and gathering data on new technologies (or on public tenders or on other topics) and making the relevant data available in an easy accessible way.

2.2 Sources and Process Flow

The InfoCodex application gathers information from up to 10'000 Web sources specified by the system administrator: WebSites, RSS feeds, or results of Web search engines. This is done by means of InfoCodex's spider agents on a central InfoCodex server (see User Manual 1, Section "Creating New Collections").

The collection process is started automatically every evening, and the newly found or the updated documents are added to a central knowledge pool (see User Manual 2, Section "Job Scheduling").

The various business departments are interested only in a subset of the gathered information. They can specify their *fields of interest* (topics) by means of *concept filters* (= textual description of the topics of interest) and they receive just that information which is compliant with their concept filters.


Periodic Gathering / Categorization
Central Knowledge Repository
Overviews / Trends / Alerts
Overviews / Trends / Alerts
Overviews / Trends / Alerts
Data sources
Content analysis
Concept filters

Provision of overviews targeted to the individual user groups (compliant with the fields of interest)

2.3 Setting the Concept Filters and Alerts

Call InfoCodex with the start URL: *yourservername*/infocodex5.html

User / password: .... / ....

Select the collection at the field top left: (if the desired collection is not already loaded)

Then, the Concept Filter Administration can be called via the button "**Admin**" → *Concept Filters*.

The following data can be set in the filter administration:

Concept filters

Retrieval conditions consisting of a textual description of the interesting topic in colloquial language (for a similarity search) and/or
a Boolean search <boole> ... </boole>.

In addition to <boole>, InfoCodex supports also <sources>, <filename> and <language> as "hard conditions".

The concept filters can be grouped into ***topics*** (*fields of interest),* i.e. a topic can be described by 1 to 20 individual concept filters.

Users

Users that should receive the output (alerts), filtered according to their fields of interests. For each user, 0 to n fields of interests can be assigned.

Settings
- output options (RSS feeds, e-mail alerts, link lists)
- tolerance for mutation alerts (a Web document is considered as "changed" if the divergence from the previous document version exceeds the specified tolerance)

Examples of topics (fields of interest) and concept filters are given in Section 6.

2.4 Retrieval Strategies and Display of the Results

Result lists per topic (field of interest)

For each concept filter of a topic, the following retrieval tests are made for all newly added documents after an update ("incremental load"):

a) Are the "hard conditions" fulfilled (<boole> etc.)?
b) If yes, is the matching similarity beween the concept filter text and the considered document greater than the "minimal relevance" specified in the "Concept filter" mask?
c) From the documents fulfilling conditions (a) and (b), the "max. docments" (see "Concept filter" mask) with the highest similarities are selected for the corresponding topic.

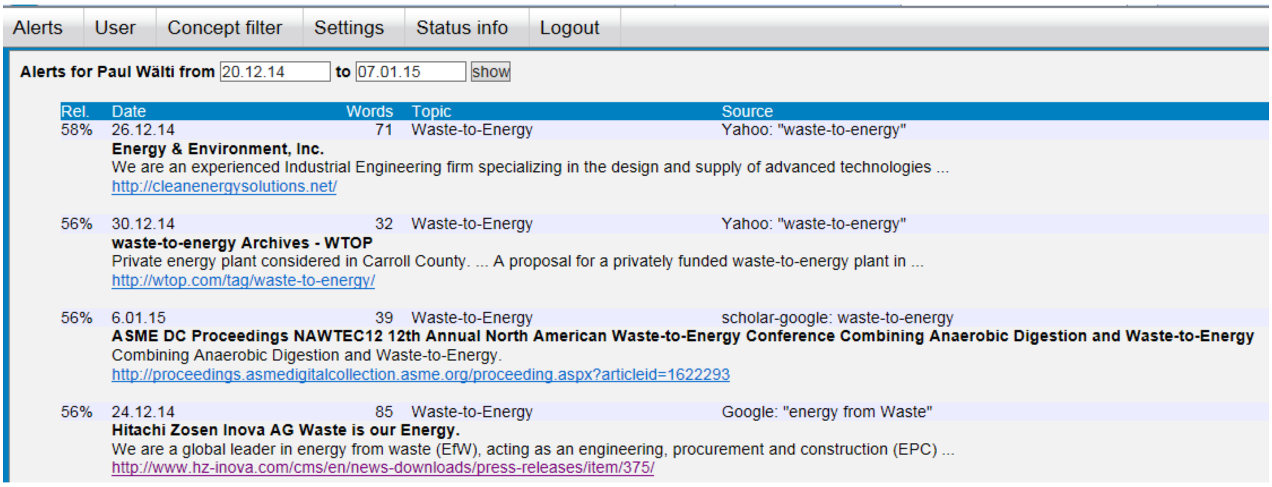
The parameters "minimal relevance" and "max. documents" are set in the mask "**Concept filter**" (see Section 6). The list of the documents selected for a given topic is composed of the selected candidates from the individual concept filters. Documents having passed more than one concept filter are taken just once (no duplicates), and the highest relevance is used in the display of the corresponding document.

The same document can be assigned to more than one topic (but not within one topic).

Storage of the results

The URLs of the documents selected per import (i.e. per daily update) are stored in a central table, together with the relevance and a summary. Depending on the chosen output option (see mask "Settings"), the result lists are forwarded by e-mails or made available as RSS feeds on the InfoCodex server.

Moreover, the results can be seen directly in the concept filter application using the mask "**Alerts**" (see following example).



Alerts for Paul Wälti from 20.12.14 **to** 07.01.15

Rel. Date Words Topic Source

58% 26.12.14 71 Waste-to-Energy Yahoo: "waste-to-energy"
Energy & Environment, Inc.
We are an experienced Industrial Engineering firm specializing in the design and supply of advanced technologies ...
http://cleanenergysolutions.net/

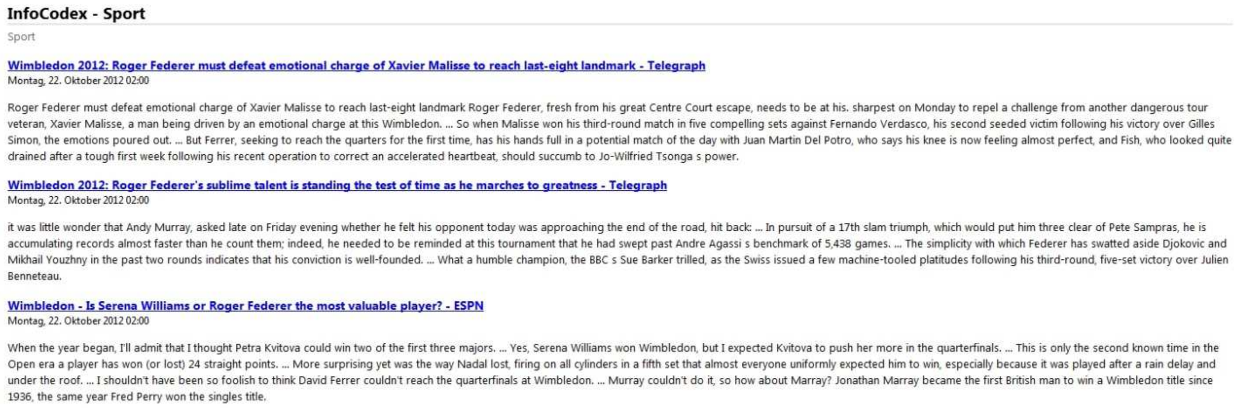
56% 30.12.14 32 Waste-to-Energy Yahoo: "waste-to-energy"
waste-to-energy Archives - WTOP
Private energy plant considered in Carroll County. ... A proposal for a privately funded waste-to-energy plant in ...
http://wtop.com/tag/waste-to-energy/

56% 6.01.15 39 Waste-to-Energy scholar-google: waste-to-energy
ASME DC Proceedings NAWTEC12 12th Annual North American Waste-to-Energy Conference Combining Anaerobic Digestion and Waste-to-Energy
Combining Anaerobic Digestion and Waste-to-Energy.
http://proceedings.asmedigitalcollection.asme.org/proceeding.aspx?articleid=1622293

56% 24.12.14 85 Waste-to-Energy Google: "energy from Waste"
Hitachi Zosen Inova AG Waste is our Energy.
We are a global leader in energy from waste (EfW), acting as an engineering, procurement and construction (EPC) ...
http://www.hz-inova.com/cms/en/news-downloads/press-releases/item/375/

*Example of a filtered retrieval results displayed under "**Alerts**" in the Concept filter application*

Example of a filtered retrieval list represented by RSS feeds



InfoCodex - Sport

Sport

Wimbledon 2012: Roger Federer must defeat emotional charge of Xavier Malisse to reach last-eight landmark - Telegraph
Montag, 22. Oktober 2012 02:00

Roger Federer must defeat emotional charge of Xavier Malisse to reach last-eight landmark Roger Federer, fresh from his great Centre Court escape, needs to be at his. sharpest on Monday to repel a challenge from another dangerous tour veteran, Xavier Malisse, a man being driven by an emotional charge at this Wimbledon. ... So when Malisse won his third-round match in five compelling sets against Fernando Verdasco, his second seeded victim following his victory over Gilles Simon, the emotions poured out. ... But Ferrer, seeking to reach the quarters for the first time, has his hands full in a potential match of the day with Juan Martin Del Potro, who says his knee is now feeling almost perfect, and Fish, who looked quite drained after a tough first week following his recent operation to correct an accelerated heartbeat, should succumb to Jo-Wilfried Tsonga s power.

Wimbledon 2012: Roger Federer's sublime talent is standing the test of time as he marches to greatness - Telegraph
Montag, 22. Oktober 2012 02:00

it was little wonder that Andy Murray, asked late on Friday evening whether he felt his opponent today was approaching the end of the road, hit back: ... In pursuit of a 17th slam triumph, which would put him three clear of Pete Sampras, he is accumulating records almost faster than he count them; indeed, he needed to be reminded at this tournament that he had swept past Andre Agassi s benchmark of 5,438 games. ... The simplicity with which Federer has swatted aside Djokovic and Mikhail Youzhny in the past two rounds indicates that his conviction is well-founded. ... What a humble champion, the BBC s Sue Barker trilled, as the Swiss issued a few machine-tooled platitudes following his third-round, five-set victory over Julien Benneteau.

Wimbledon - Is Serena Williams or Roger Federer the most valuable player? - ESPN
Montag, 22. Oktober 2012 02:00

When the year began, I'll admit that I thought Petra Kvitova could win two of the first three majors. ... Yes, Serena Williams won Wimbledon, but I expected Kvitova to push her more in the quarterfinals. ... This is only the second known time in the Open era a player has won (or lost) 24 straight points. ... More surprising yet was the way Nadal lost, firing on all cylinders in a fifth set that almost everyone uniformly expected him to win, especially because it was played after a rain delay and under the roof. ... I shouldn't have been so foolish to think David Ferrer couldn't reach the quarterfinals at Wimbledon. ... Murray couldn't do it, so how about Marray? Jonathan Marray became the first British man to win a Wimbledon title since 1936, the same year Fred Perry won the singles title.

Example of RSS-Feeds for the topic "Sport"
(the summaries are the abstracts automatically generated by InfoCodex)

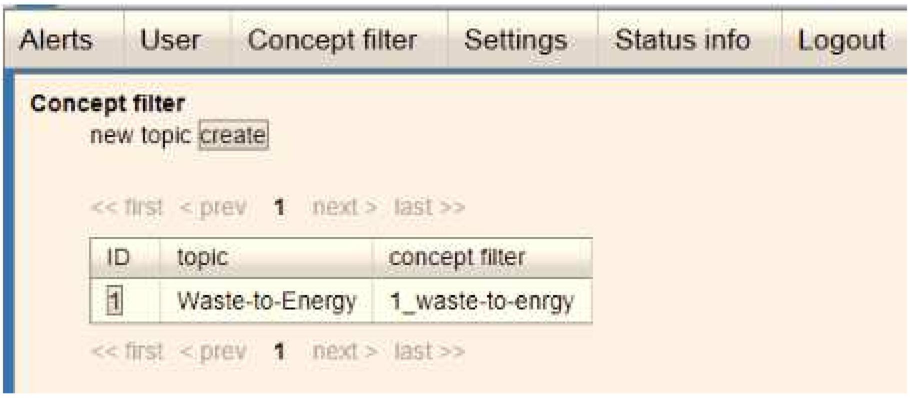
2.5 Calling the Concept Filter Application

Either via the InfoCodex standard interface

yourservername/infocodex5.html → select collection → "**Admin**" → Concept Filter

or directly with a specific URL.

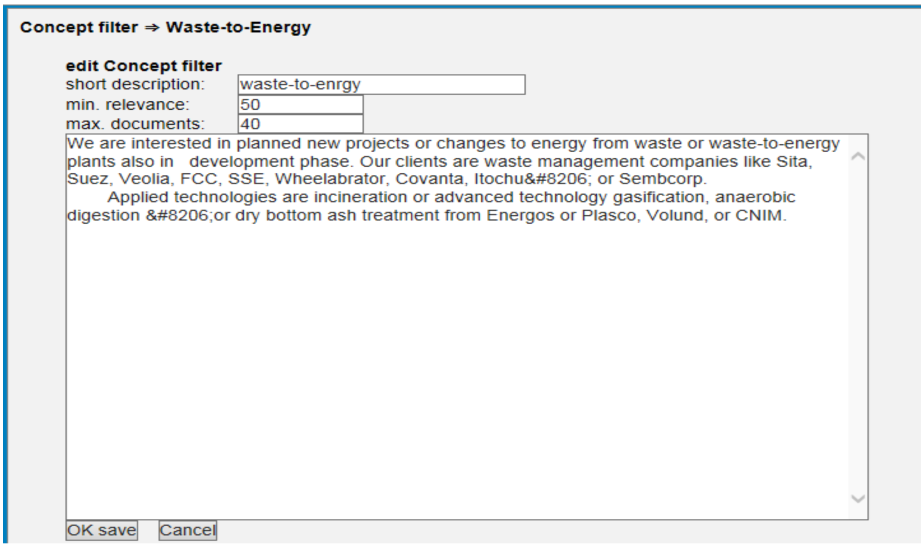
2.6 Examples of Concept Filters and Fields of Interest (Topics)



Concept filter

ID	topic	concept filter
1	Waste-to-Energy	1_waste-to-enrgy

Example of the mask for "Topics" (under the button "**Concept filter**")



Concept filter ⇒ Waste-to-Energy

edit Concept filter

short description: waste-to-enrgy

min. relevance: 50

max. documents: 40

We are interested in planned new projects or changes to energy from waste or waste-to-energy plants also in   development phase. Our clients are waste management companies like Sita, Suez, Veolia, FCC, SSE, Wheelabrator, Covanta, Itochu&#8206; or Sembcorp.

Applied technologies are incineration or advanced technology gasification, anaerobic digestion &#8206;or dry bottom ash treatment from Energos or Plasco, Volund, or CNIM.

Example of the mask for "**Concept filter**"

Examples of entries in the text field of a concept filter

Example 1 (Boolean expression combined with describing text in natural language)

<boole> (Knowledge management or semantic technologies) and
   (enterprise search or economic intelligence)
</boole>
The major challenge faced by Enterprise search is the need to index data and documents from a variety of sources such as: file systems, intranets, document management systems, e-mail, and databases and then present a consolidated list of relevance ranked resources from these various sources. In addition, many applications require the integration of structured data as part of the search criteria and when presenting results back to the users. And of course access controls are vital if users are to be restricted to data and documents which they are granted access by the various document repositories within the enterprise. These major challenges are unique to enterprise search.

Semantic technologies are giving computers the power to understand and categorize documents by thematic content and maximize the impact of information.

Example 2 (Only Boolean expression for exact search; no free text)

<boole>
"Gestion des connaissances" or "base de connaissances" or catégorisation or
"capitalisation des connaissances" or "fouille de texte" or "text mining"
</boole>

Note: The quotation marks mean that these expressions must appear exactly in this form.

Example 3 (Similarity search with describing text in natural language, combined with file name restrictions)

<filename> *sem* </filename>

Semantic technologies are giving computers the power to understand and categorize documents by thematic content and maximize the impact of information.
This is an extremely challenging task - but the patented and award-winning technology behind InfoCodex has proven to manage it, across languages, and across your organization.
Faceted search; content search; content identification; federated search; collaborative tagging; thematic clustering; entity extraction; named entities recognition; enterprise bookmarking

Syntax for <boole>, <sources>, <filename> and <language>

<boole> ... </boole>	Same syntax as in the search mask of InfoCodex; see "**Help**" → Search instructions
<sources> 17; 21 </sources>	Source numbers, delimited with ";". The source numbers can be seen under the link "view sources" in the concept filter mask
<filename> *sem* </filename>	Parts of file names (here: only documents with file names *sem*)
<language> en </language>	Only English documents (possible parameters: de, en, fr, it, es)