

# **InsumaScout Booklet**

Konkurrenzbeobachtung / Marktanalyse

Erweiterung von Wissensdomänen

Thematische, lernfähige und automatisierte Web-Suche

**InsumaScout ist ein lernfähiges, thematisches Such- und Liefersystem. Es basiert auf einem Crawler mit Benutzer-adaptivem Filter. Der Filter wird sukzessive durch den Benutzer selbst und dessen Bewertungen eingehender Dokumente trainiert. InsumaScout findet so immer relevantere Dokumente im Web, ohne dass der Benutzer ständig sich wiederholende Suchanfragen stellen muss. Ein voll abgeschlossener Zyklus aus Dokumentenlieferung-Bewertung-Lieferung generiert eine Fachsuchmaschine aus dem Stand. Der gelieferte hoch-relevante Content wird schließlich im Routinebetrieb täglich mit aktuellen Dokumenten ergänzt. Der Content kann dazu benutzt werden, bestehende Inhalte (eines Portals, Content Management Systems, etc.) bzw. eine Wissensdomäne eines Unternehmen-Intranets sinnvoll zu erweitern. Auf Wunsch können registrierten Benutzern relevante Inhalte per Email zugesandt werden (täglich oder als wöchentliche Zusammenfassung).**

## **Inhalt:**

1. InsumaScout Control Center	2
2. Administrations-Seite	3
3. Definition des Kern-Themas	4
4. Definition der Start-URLs	5
5. Freie Definition von Kategorien	6
6. Regelmäßige Lieferung relevanter Dokumente (Fetch Rounds)	7
7. Bewertung der gelieferten Dokumente	8
8. Freshness-Detection und Routine-Phase	9
9. Profi-Suche über erfolgreich bewertete Dokumentenmenge	10
10. Hilfe-System	11

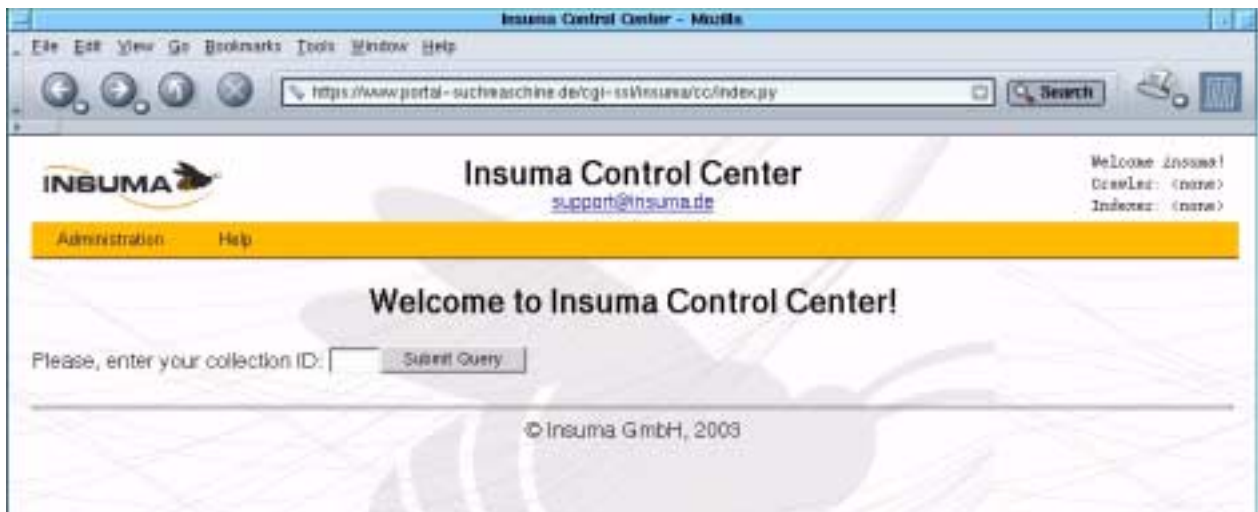
# InsumaScout Control Center



Login-ID

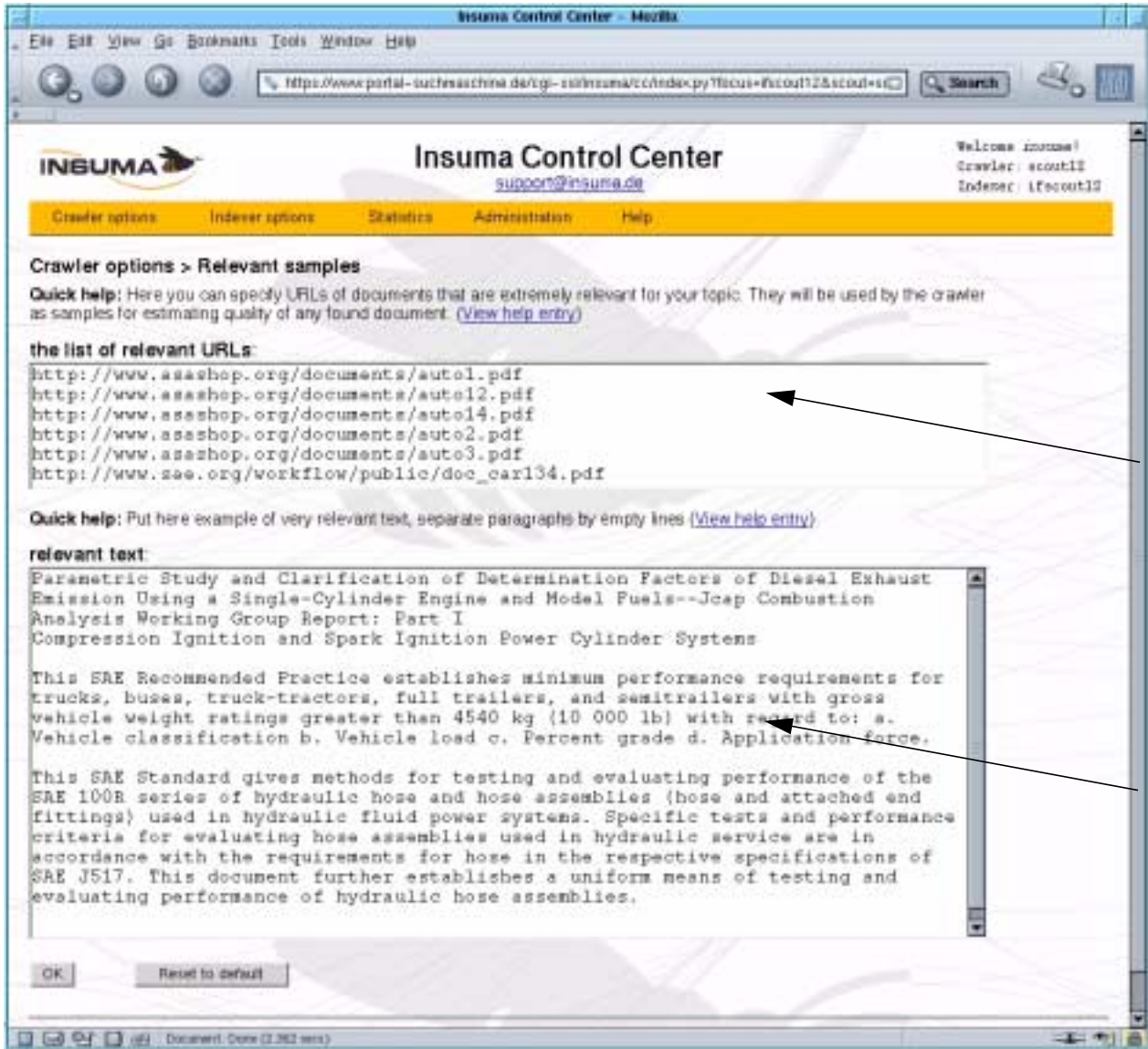
Das **Control-Center** erlaubt die Administration mehrerer Themenbereiche (genannt: *Collections*) zur gleichen Zeit. Jedem Themenbereich lässt sich - ähnlich einem Autorensystem - ein dedizierter Administrator mit erweiterten Zuständigkeiten sowie mehrere untergeordnete Benutzer (Content Analyser) zuordnen. Das Einloggen in das InsumaScout-Control Center erfolgt von der Insuma-Startseite.

## Administrations-Seite



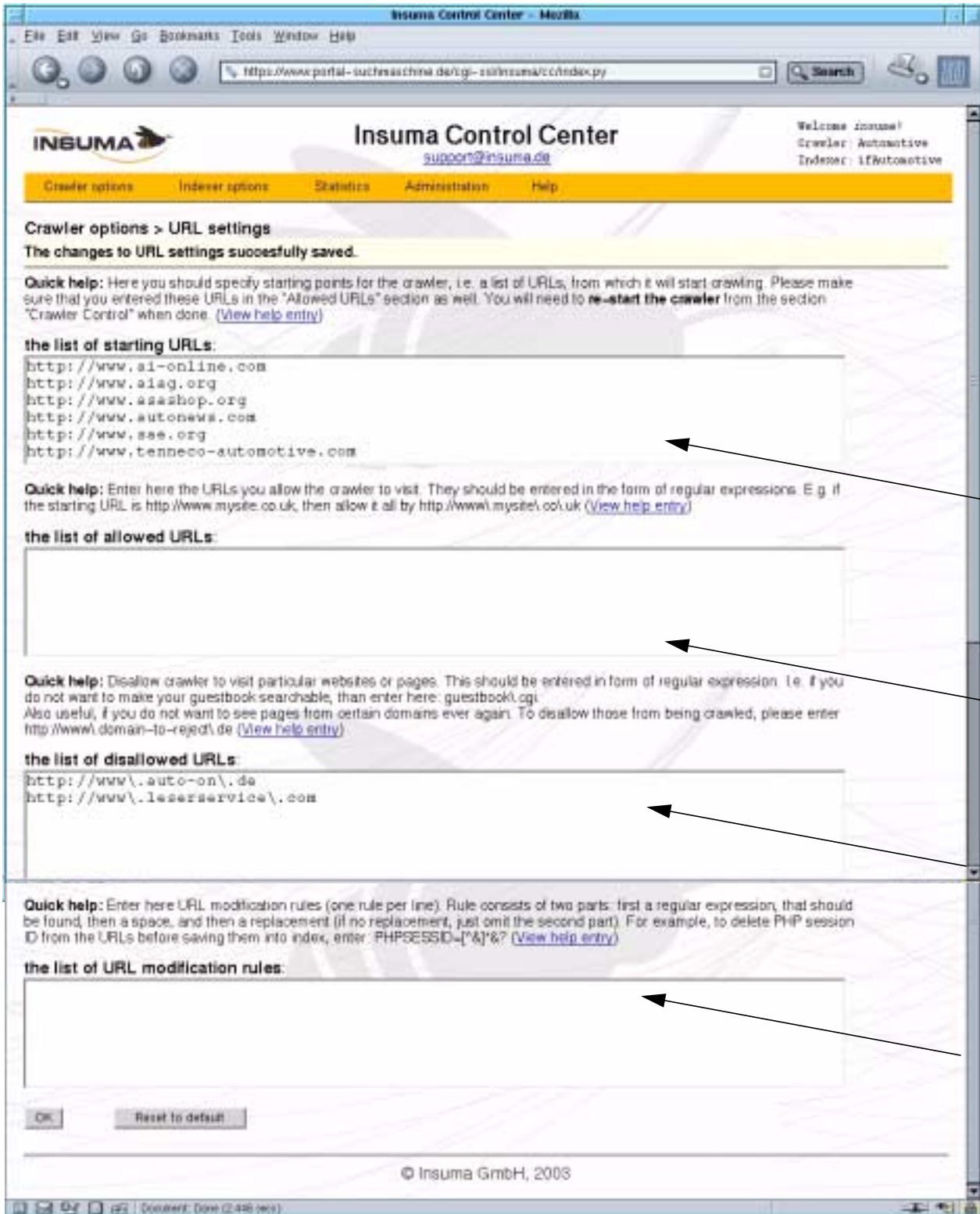
Die **Administrations-Seite** erlaubt den Zugriff auf alle weitere Bearbeitungs-Funktionalitäten, wie Bewertungs-Funktion, Profi-Suche, Control-Center für das Collection Set-Up, etc.

## Definition des Kern-Themas



Das **Kern-Thema** erlaubt dem Benutzer, das Thema seiner Wahl frei zu definieren. Er kann dies zudem problemlos über bereits vorhandene Dokumente (doc, rtf, pdf, txt, html) realisieren, die in das System upgeloadet werden. Die Themen-Collections sind sprach-spezifisch.

# Defintion der Start-URLs



Start-URLs

Einschränkung der Suche

Ausschluss von Sub-Seiten

Ausschluss von Session-IDs

Die **Start-URLs** bzw. Start-Adressen dienen dazu, dem System Startpunkte der Suche vorzugeben. Mit Hilfe des Ausschlusses von **Sub-Seiten**, können bei Bedarf Unterseiten von den zu crawlenden Seiten ausgeschlossen oder exakter spezifiziert werden. Die letzte Form behandelt Adressen, welche **Session\_IDs** beinhalten.

## Freie Definition von Kategorien

The screenshot shows the Insuma Control Center web interface. The page title is 'Indexer options > URL to category mapping'. The 'Quick help' section explains that users can sort crawled web pages to categories based on their URL. The 'Format' section shows the format: `URL_mask category_value`. The 'Examples' section lists two examples: `http://www.mysite.com/path/category1/ My_category_1` and `http://www.mysite.com/path/news/ News`. Below the examples is a table with two rows of data:

<code>http://www.asashop.org/eshop/automotive</code>	Shop
<code>http://www.tenneco-automotive.com</code>	Industry

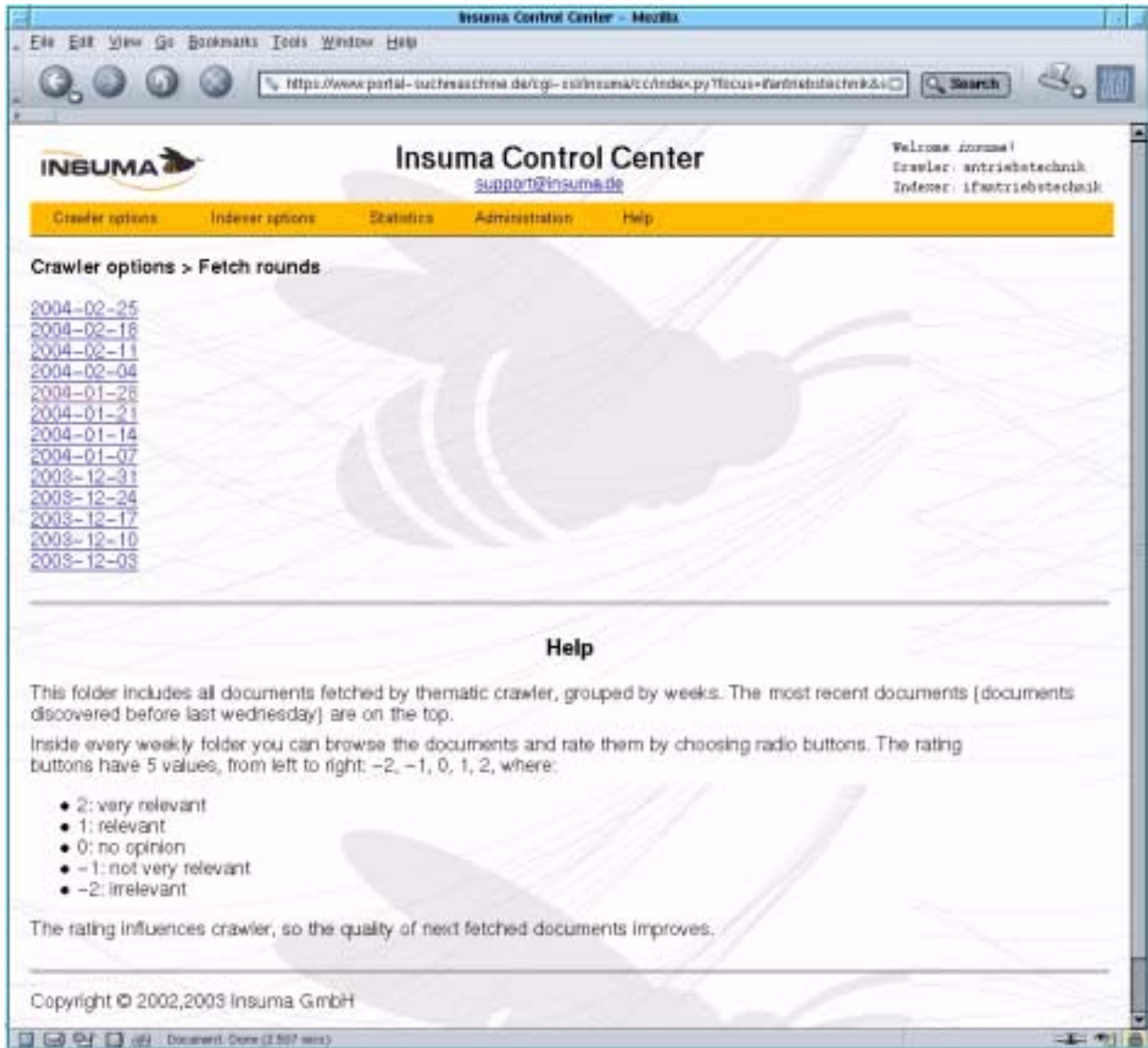
An arrow points to the 'Industry' category in the table, labeled 'Kategorien'. The page also includes a navigation menu with 'Crawler options', 'Indexer options', 'Statistics', 'Administration', and 'Help'. The footer shows '© Insuma GmbH, 2003'.

**Kategorien** können in beliebiger Anzahl frei durch den Benutzer definiert werden.

InsumaScout erlaubt seinen Benutzern, so viele Kategorien zu definieren, wie sie möchten. Eingehende noch unannotierte Dokumente werden automatisch den existierenden Benutzer-Kategorien zugeteilt. Die Zuteilung geschieht über ausgeklügelte semantische Ähnlichkeitsanalysen.

Der gesamte Kategorisierungs-Prozess benötigt keine *manuelle* Arbeiten durch den Benutzer!

## Regelmäßige Lieferung relevanter Dokumente (Fetch Rounds)



Die von InsumaScout **gelieferten Dokumente** können auf dieser zentralen Seite durch den Benutzer (Content Analyser) bewertet werden. Dem Benutzer stehen für jedes gelieferte Dokument *Bewertungen* zwischen vollkommen irrelevant (-2) und sehr relevant (+2) zur Verfügung. Die gelieferten Dokumente werden in (wöchentliche) sogenannte **Fetch-Rounds** gegliedert.

## Bewertung der gelieferten Dokumente

The screenshot shows a web browser window titled "Search results - Media". The address bar contains the URL "https://www.insuma-scout.de/g...". The page displays search results for a query. Four results are visible, each with a title, a brief description, a URL, and a list of categories. Annotations with arrows point to specific features:

- Dok.-Bewertungsschema:** Points to the "Rate:" field below the first result, which shows a scale from -2 to +2.
- Dok.-Kategorisierung:** Points to the "Assign categories:" dropdown menu for the second result, which is open and shows a list of categories.
- Dok.-Archivierungssystem:** Points to the "Assign categories:" dropdown menu for the third result.
- Dok.-Annotierungssystem:** Points to the "Annotations" link next to the title of the third result.

Die von Insuma-Scout **gelieferten Dokumente** können auf dieser Fetch-Round Seite durch den Benutzer (Content Analyser) bewertet werden. Dem Benutzer stehen für jedes gelieferte Dokument *Bewertungen* zwischen vollkommen irrelevant (-2) und sehr relevant (+2) zur Verfügung. Zugleich lassen sich die Dokumente in die zuvor definierten Kategorien einteilen. Das *Dokumenten-Annotierungssystem* erleichtert den Austausch von weiteren Bemerkungen zwischen den beteiligten Mitarbeiter-Gruppen (Groupware-Lösung).

## Eingang in die Routine-Phase



In die **Routine-Phase** wird nach ca. 4 Fetch-rounds eingetreten bzw. nachdem der Crawler ausreichend trainiert ist. In die Routine-Phase geht InsumaScout über, nachdem der Benutzer die sogenannte **Freshness-Detection** auf 1 setzt (für die *tägliche* Benachrichtigung eingehender relevanter Dokumente).

## Profi-Suche über erfolgreich bewertete Dokumentenmenge

Test search page - Mozilla

File Edit View Go Bookmarks Tools Window Help

https://www.portal-suchmaschine.de/ocsit/search\_antrieb.html Search

### Search

Search, browse, categorise, rate:

Search in body (can be left empty):

— AND —

limit by category:

limit by fetch round:  AND

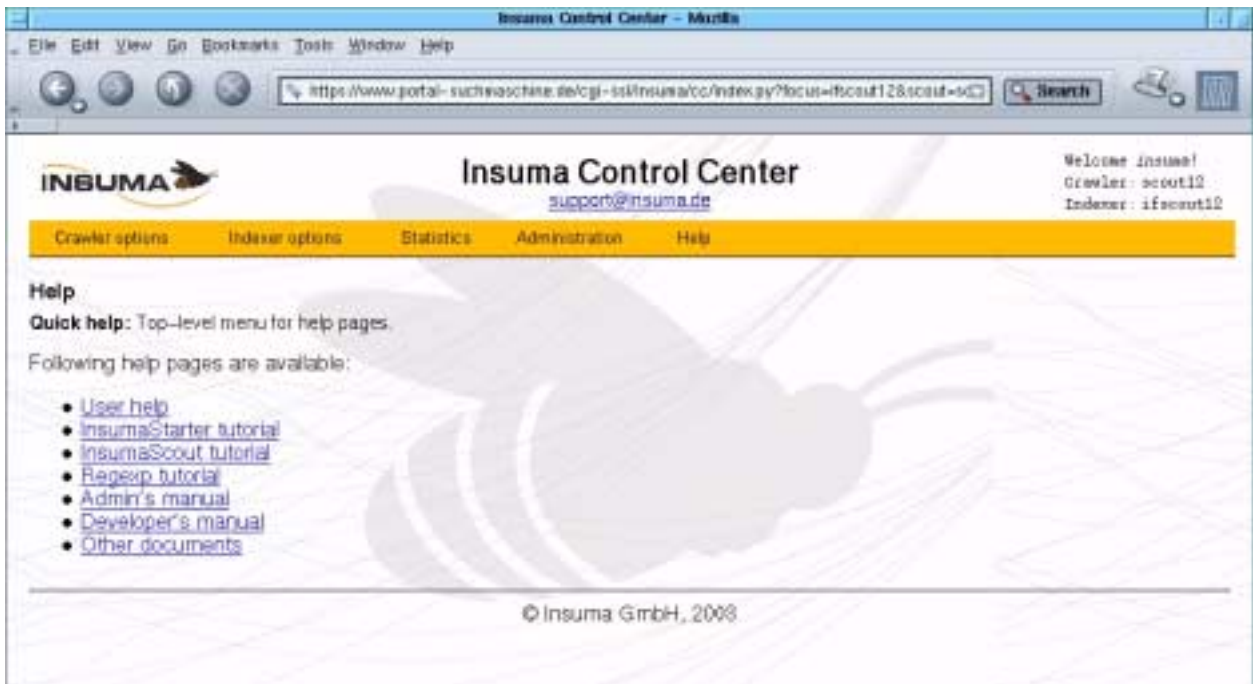
Last-modified from (YYYY-MM-DD)  until  (from or/and until can be left empty)

Show  results per page

show categories  
 show relevance

Schließlich kann durch die **Profi-Suche** über die Dokumentenmenge der bereits bewerteten und gespeicherten Dokumente gesucht werden. Dabei werden selbstverständlich die zugewiesenen *Kategorien* sowie das *Änderungsdatum* der Dokumente mit berücksichtigt..

# Hilfe-System



Dem Benutzer und Administrator steht jederzeit ein **Online Hilfe-System** sowie mehrere weiterführende Artikel zum Download zur Verfügung.

## Kontakt

Insuma GmbH  
Sand 13  
72076 Tübingen  
Tel.: (07071) 3 659 610  
Fax.: (07071) 3 659 614  
Email: [info@insuma.de](mailto:info@insuma.de)