Automating Production of Cross Media Content for Multi-channel Distribution

www.AXMEDIS.org

DE3.1.2.2.8
Specification of AXMEDIS CMS Crawling Capabilities, first update of part of DE3.1.2

Version: 1.3
Date: 20-04-2006
Responsible: DSI - Focuseek (revised and approved by coordinator)

Project Number: IST-2-511299
Project Title: AXMEDIS
Deliverable Type: report
Visible to User Groups: yes
Visible to Affiliated: yes
Visible to the Public: yes

Deliverable Number: DE3.1.2.2.8
Contractual Date of Delivery: M18
Actual Date of Delivery: Specification of AXMEDIS CMS Crawling Capabilities, first update of part of DE3.1.2
Title of Deliverable: Document
Work-Package contributing to the Deliverable: WP3.1
Task contributing to the Deliverable: WP3, WP2
Nature of the Deliverable: report
Author(s): DSI, Focuseek

Abstract: this part includes the specification of components, formats, databases and protocol related to the AXMEDIS Framework area including the crawlers of Focuseek.

Keyword List: CMS, Content management systems, Crawling, ODBC, Tamino, XML, ORACLE, MySQL, etc.

2. For the avoidance of doubt the licensee should be aware that although the Copyright in the documents is given under warranty this warranty does not extend to the content of any document which may contain references or specifications or technologies that are covered by patents (also of third parties) or that refer to other standards. AXMEDIS is not responsible and does not guarantee that the information contained in the document is fully proprietary of AXMEDIS consortium and/or partners.
3. Licensee hereby undertakes to the Licensor that he will, without prejudice to any other right of action which the Licensor may have, at all times keep the Licensor fully and effectively indemnified against all and any liability (which liability shall include, without limitation, all losses, costs, claims, expenses, demands, actions, damages, legal and other professional fees and expenses on a full indemnity basis) which the Licensor may suffer or incur as a result of, or by reason of, any breach or non-fulfillment of any of his obligations in respect of this License.

3. **INFRINGEMENT**

   1. Licensee undertakes to notify promptly the Licensor of any threatened or actual infringement of the Copyright which comes to licensee notice and shall, at the Licensor's request and expense, do all such things as are reasonably necessary to defend and enforce the Licensor's rights in the Copyright.

4. **GOVERNING LAW AND JURISDICTION**

   1. This Licence shall be subject to, and construed and interpreted in accordance with Italian law.
   2. The parties irrevocably submit to the exclusive jurisdiction of the Italian Courts.

Please note that:

- You can become affiliated with AXMEDIS. This will give you the access to a huge amount of knowledge, information and source code related to the AXMEDIS Framework. If you are interested please contact P. Nesi at nesi@dsi.unifi.it. Once affiliated with AXMEDIS you will have the possibility of using the AXMEDIS specification and technology for your business.
- You can contribute to the improvement of AXMEDIS documents and specification by sending the contribution to P. Nesi at nesi@dsi.unifi.it
- You can attend AXMEDIS meetings that are open to public, for additional information see WWW.axmedis.org or contact P. Nesi at nesi@dsi.unifi.it

Table of Content

1 Executive Summary and Report Scope

The full AXMEDIS specification document has been decomposed in the following parts:

DE number	Deliverable title	responsible
DE3.1.2.2.1	Specification of General Aspects of AXMEDIS framework, first update of DE3.1.2 part A AXMEDIS-DE3-1-2-2-1-Spec-of-AX-Gen-Asp-of-AXMEDIS-framework-upA-v1-0.doc	DSI
DE3.1.2.2.2	Specification of AXMEDIS Command Manager, first update of DE3.1.2 part B AXMEDIS- DE3-1-2-2-2-Spec-of-AX-Cmd-Man-upB-v1-0.doc	DSI
DE3.1.2.2.3	Specification of AXMEDIS Object Manager and Protection Processor, first update of DE3.1.2 part B AXMEDIS-DE3-1-2-2-3-Spec-of-AXOM-and-ProtProc-upB-v1-0.doc	DSI
DE3.1.2.2.4	Specification of AXMEDIS Editors and Viewers, first update of DE3.1.2 part B AXMEDIS-DE3-1-2-2-4-Spec-of-AX-Editors-and-Viewers-upB-v1-0.doc	DSI
DE3.1.2.2.5	Specification of External AXMEDIS Editors/Viewers and Players, first update of DE3.1.2 part B AXMEDIS-DE3-1-2-2-5-Spec-of-External-Editors-Viewers-Players-upB-v1-0.doc	EPFL
DE3.1.2.2.6	Specification of AXMEDIS Content Processing, first update of DE3.1.2 part C AXMEDIS-DE3-1-2-2-6-Spec-of-AX-Content-Processing-upC-v1-0.doc	DSI
DE3.1.2.2.7	Specification of AXMEDIS External Processing Algorithms AXMEDIS-DE3-1-2-2-7-Spec-of-AX-External-Processing-Algorithms-v1-0.doc	FHGIGD
DE3.1.2.2.8	Specification of AXMEDIS CMS Crawling Capabilities, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-8-Spec-of-AX-CMS-Crawling-Capab-v1-0.doc	DSI
DE3.1.2.2.9	Specification of AXMEDIS database and query support, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-9-Spec-of-AX-database-and-query-support-v1-0.doc	EXITECH
DE3.1.2.2.10	Specification of AXMEDIS P2P tools, AXEPTool and AXMEDIS, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-10-Spec-of-AXEPTool-and-AXMEDIA-tools-v1-0.doc	CRS4
DE3.1.2.2.11	Specification of AXMEDIS Programme and Publication tools, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-11-Spec-of-AX-Progr-and-Pub-tool-v1-0.doc	UNIVLEEDS
DE3.1.2.2.12	Specification of AXMEDIS Workflow Tools, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-12-Spec-of-AX-Workflow-Tools-v1-0.doc	IRC
DE3.1.2.2.13	Specification of AXMEDIS Certifier and Supervisor and networks of AXCS, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-13-Spec-of-AXCS-and-networks-v1-0.doc	DSI
DE3.1.2.2.14	Specification of AXMEDIS Protection Support, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-14-Spec-of-AX-Protection-Support-v1-0.doc	FUPF
DE3.1.2.2.15	Specification of AXMEDIS accounting and reporting, first update of part of DE3.1.2 AXMEDIS-DE3-1-2-2-15-Spec-of-AX-Accounting-and-Reporting-v1-0.doc	EXITECH

1.1 This document concerns

In this document specification of the CMS Crawling subsystem are presented.

1.2 List of Modules or Executable Tools Specified in this document

A module is a component that can be or it is reused in other cases or points of the AXMEDIS framework or of other AXMEDIS based solutions.
The modules/tools have to include effective components and/or tools and also testing components and tools.

Module/tool Name	Module/Tool Description and purpose, state also in which other AXMEDIS area is used	Standards exploited if any
Crawler Collector Indexer	Implemented by the focuseek platform, it is responsible for gathering and indexing contents from CMSs	
Crawler Results Integrated database	It a component of focuseek platform containing the local copy of original content.	
Watch manager	Notify when new contents are gathered from CMS	
CMS Missing Interface(s)	Gathering plug-ins for custom CMSs. (Tamino/Lobster and generic Web Service)	
Crawler Query Adapter / Collector Engine Query Support Interface	Interface to query Crawler database	
Fast Access DB Interface	An interface module to transfer contents from the Crawler Results Integrated Database	

1.3 List of Formats Specified in this document

A format can be (i) an XML content file for modeling some information, (ii) a file format for storing information, (iii) a format that is manipulated by the tools described in this document, etc...

Format Name	Format Description and purpose, state also in which other modules is used	Standards exploited if any
FFF	Internal focuseek searchbox XML reprerentation format for documents and associated metadata	

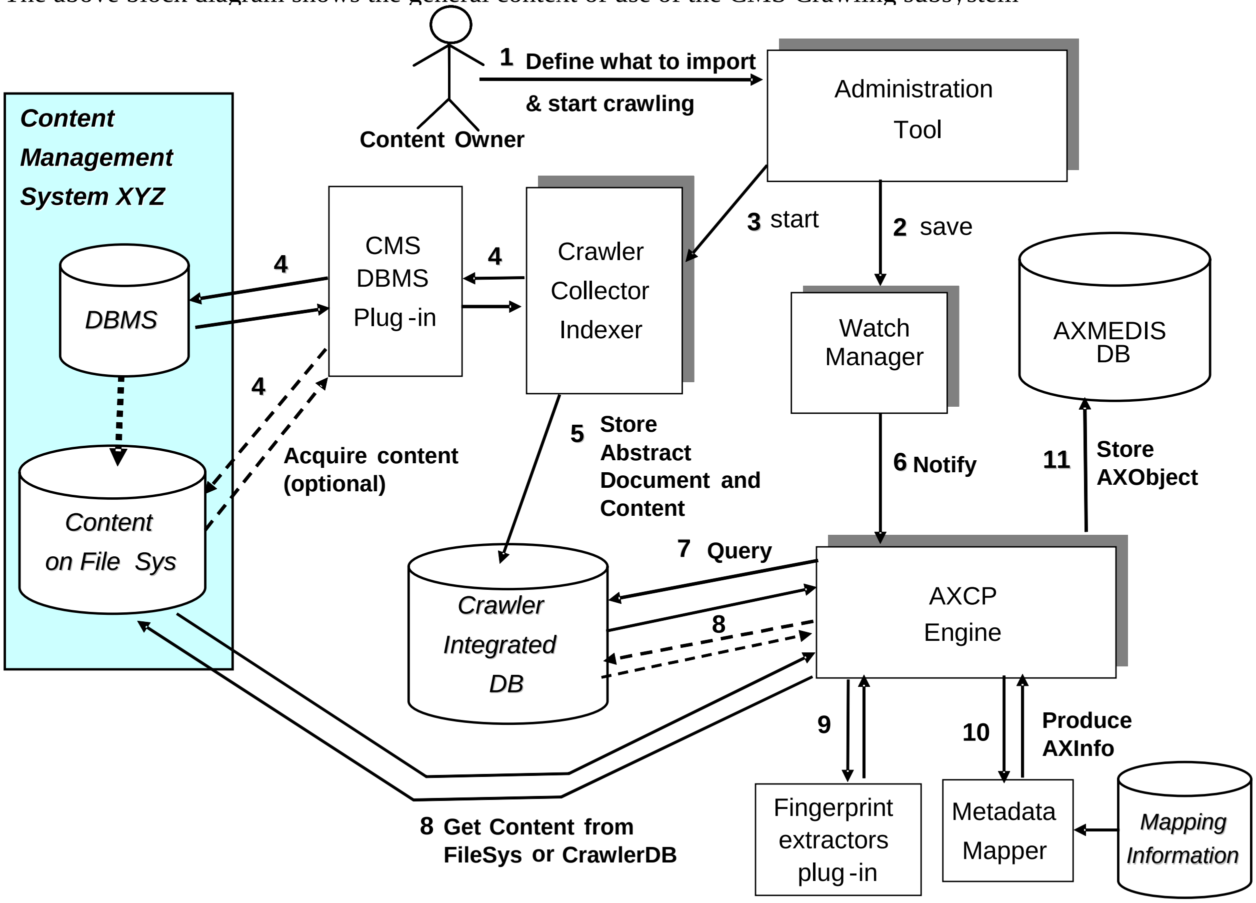
1.4 List of Protocols Specified in this document

A protocol is a communication modality among distinct processes that can be located or not on different computers.

Protocol Name	protocol Description and purpose, state also in which other modules is used	Who is the master and who is the slave	Standards exploited if any
SOAP	Focuseek searchbox is a Web Service completely accessible using a SOAP API		SOAP, .NET
REST	Used to implement the Fast Access Db Interface		

2 General architecture and relationships among the modules produced

The above block diagram shows the general context of use of the CMS Crawling subsystem



3 Executable Tool – Collector Indexer

Tool Profile		
Collector Indexer		
Responsible Name	Baldini	
Responsible Partner	Focuseek	
Status (proposed/approved)	approved	
Implemented/not implemented	Implemented	
Status of the implementation	100%	
Executable or Library/module (Support)	Executable	
Single Thread or Multithread	Multithread	
Language of Development	C++	
Platforms supported	Windows, Linux, Mac OS X	
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for public download		
Address for accessing to WebServices if any, add accession information (user and Passwd) if any		
Test cases (present/absent)	absent	
Test cases location		
Usage of the AXMEDIS configuration manager (yes/no)	no	
Usage of the AXMEDIS Error Manager (yes/no)	no	
Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)

Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
libcurl	Latest stable	CURL License
expat	Latest stable	Expat license
Henry Spencer regex library	Latest stable	Henry Spencer regex license
Apache Lucene	Latest stable	Apache license
SQLite	Latest stable	Public domain
libwv2	Latest stable	LGPL license
EasySOAP++	Latest stable	LGPL license
wxWidget	Latest stable	LGPL license
GCJ Runtime	Latest stable	LIBGCJ license
neon	Latest stable	LGPL
samba	Latest stable	GPL license
boost	Latest stable	BOOST license
iksemel	Latest stable	LGPL license
libxml2	Latest stable	MIT license
libxslt	Latest stable	MIT license

raptor	Latest stable	LGPL license
uwimap	Latest stable	University of Washington's Free-Fork License
zlib	Latest stable	ZLIB license

3.1 General Description of the Module

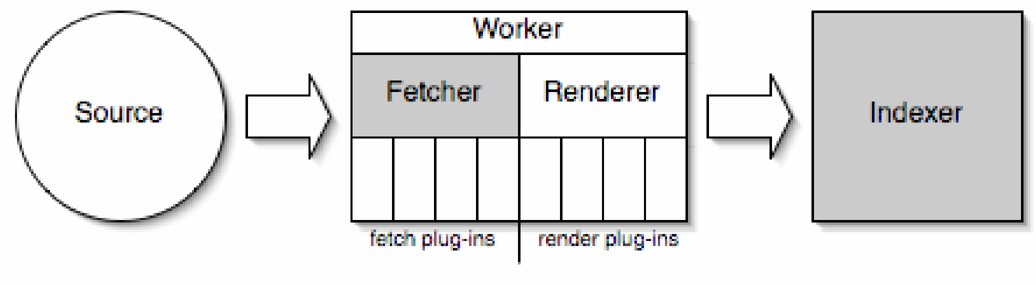
Focuseek is used to crawl into the CMSs and the databases of the content providers, integrators, and distributors to access content and indexing it. Specific interfaces for accessing to these databases will be defined and developed. We will design and develop a framework (based on Focuseek) which allows collecting and indexing different media sources.

The idea is to access already available information and metadata into CMSs and transfer it into a cache of the content. Those may have information related to the content, their metadata, their technical details, DRM aspects, licensing aspects, possible products that use the content, multilingual information, etc…

This tool is called in the general architecture as Crawler Collector Indexer. It may navigate on any kind of source by means of special plug-ins that will be developed.

This task will be performed by acquiring the license of a product called Focuseek and integrating it in the AXMEDIS framework. The license will come with editable manual to customize the installation to the AXMEDIS solution.

This component is already integrated into the focuseek platform and it is represented by the union of the fetching agent (fetcher) and of the indexing service (indexer).



In the previous picture with light grey color the subcomponent of the focuseek architecture involved in this portion of AXMEDIS system.

Note that if the Renderer component of Worker has no active plug-ins configured it is excluded from the overall process.

The other type of plug-in (fetch plug-in in the picture) are those used to let the fetcher to gather directly contents from sources. Custom plug-ins for specific CMS will be developed, focuseek supports by defaults HTTP, NNTP, SMB, FTP, local FS, ODBC, WebDav, Gopher and SMB protocols.

3.2 Module Design in terms of Classes

This module is part of the standard focuseek commercial package, a class diagram is not available.

3.2.1 Focuseek plug-in architecture

All plug-ins in the focuseek architecture are managed by a single component called Worker. Such component expose a standard C API and all Plug-Ins can be implemented as standard DLL.

focuseek searchbox supports three plugin types:

- Protocol plugin. Plugins implementing a standard protocol (e.g. http) or a custom one.
- Parser plugin. Plugins that parses documents (e.g. HTML documents).
-

Protocol plugins

Protocol plugins basically accept an URL specifying a document and retrieve it. Their output comprises an error code (e.g. "Document not found"), the document contents as a binary blob of data and optionally the MIME type for the document if the protocol supports it [1].

Parser plugins

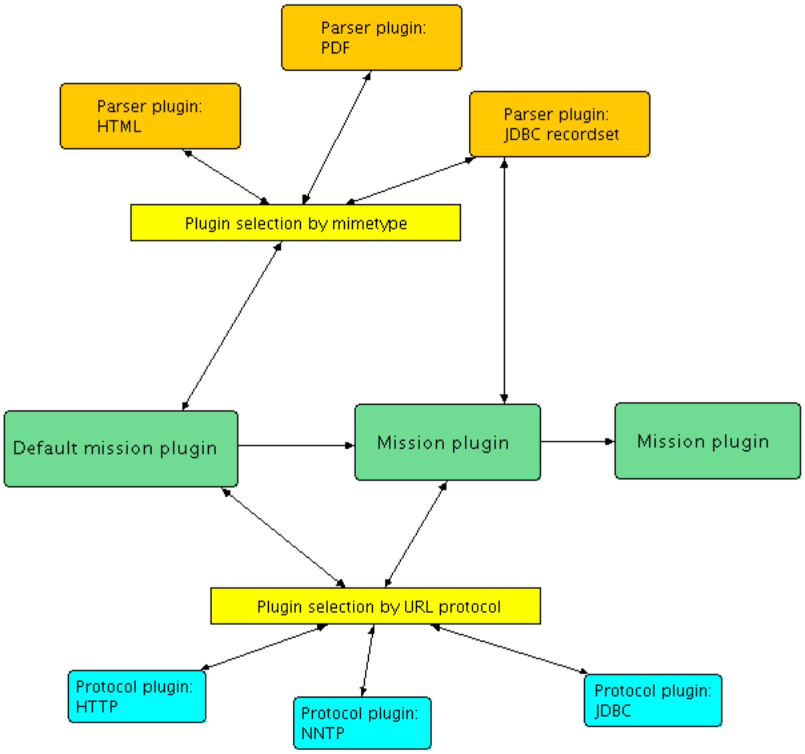
Parser plugins receive the original document contents, i.e. a blob of binary data, and extract textual and typographical information from it. A parser plugin input contains info about the document, such as the document URL, MIME type and contents. It also contains a list of info about "related documents": documents that are required to correctly process the original document. For example a plugin for the HTML format could require a CSS style sheet, referred in the HTML document itself.

The plugin outputs a list of requests for related documents and a list of parsed documents. Parsed documents are expressed in the FFF format and will be further processed by other plugins and by searchbox itself.

The plugin is invoked repeatedly until no more related documents are required or no more related documents can be successfully retrieved. At this point the plugin must emit the processed documents or signal an error condition.

The plugin is also passed a *userdata* field that can be used to store temporary parser states.

The round trip



Everything starts when searchbox needs to proces a document from a source. It retrieves the mission plugin pipeline for the source and passes the document URL as input to the first mission plugin. Each mission plugin performs its computations, maybe requiring retrieval and/or parsing of other documents and then passes the results on to the next mission plugin in the pipeline. After all the mission plugins ran searchbox checks if some of them required further processing. If it did then the pipeline is fed *the whole output* of the last plugin and a new pipeline processing cicle starts. If no work needs to be done then searchbox retrieves the results and stores them in its index.

Altough mission plugins usually schedule work for protocol and parser plugins this is not required; a trivial mission plugin might add a fixed metadata to all the documents it process just to mark them.

3.2.2 Crawler Results Integrated Database(WP4.2.1: DSI with subcontract)

This component is part of focuseek architecture and it is called DocumentStore. It organized as a multilevel document cache which is able to store the original documents fetched from source in their original format. Such cache can be used in cases where the original source is non always online. Such DocumentStore can be set-up in three different ways:

- to contain only the CRC of the original content
- to contain CRC and all metadata and a pointer to the original content
- to contain CRC, metadata and the content in its original format

The documentstore is accessible through the SOAP interface of the focuseek platform.

The specific data structure to be filled is:

```
struct ArchiveConfiguration
{
	string Name;
	boolean Historicize;
	string AdministrativeContact;
	integer Source;
	integer Auth;
	Frequency SchedulingFrequency;
	integer SchedulingAttribute;
	integer AccessTime;
	integer PagesLimit;
	integer GarbageLimit;
	Cache DocumentCache;
	AccessSpec[] ACL;
}
```

Used to define parameters of an archive, it has the following fields:

- **string Name -** Archive name, used by client.
- **boolean Historicize -** Enable historicization. The value that is initially set with the AddArchive call is retained throughout all the lifecycle of the archive, i.e. you cannot change its value with the SetArchiveConfiguration call.
- **string AdministrativeContact -** Email address of administrative contact.
- **integer Source -** ID of the source configuration used to crawl documents for this archive.
- **integer Auth -** ID of the authentication configuration used to crawl documents for this archive when an authentication type other than AUTH_NONE is specified in the source configuration. Use 0 if AUTH_NONE is used in the source configuration.
- **Frequency SchedulingFrequency -** Base archive automatic refresh frequency.
- **integer SchedulingAttribute -** Base refresh frequency multiplication factor. If the base frequency is HOURLY or DAILY, the multiplication stands for the interval in hours or day between two crawls.
- **integer AccessTime -** Start of crawl. If the base frequency is HOURLY it can have a value between 0 and SchedulingAttribute and sets the time of the first daily crawl, expressed as an offset from 00:00 GMT. If the base frequency is DAILY it can have a value between 0 and 23 and sets the GMT time of crawl.
- **integer PagesLimit -** Maximum number of documents to fetch during a crawl session. "0" mean no limit. When during a crawl this limit is reached the crawling session is terminated.

- **integer GarbageLimit** - Minimum age of a document before garbage collection. "0" mean no garbage collection. The age is expressed in seconds and is measured from the last time the document was fetched.
- **Cache DocumentCache** - Caching level to be used for this archive.
- **AccessSpec[] ACL** - Access control list for this configuration.

Where DocumentCache can be:

- FULLCACHE **-** Retain full document cache
- CONTEXTCACHE **-** Retain minimal cache needed for context extraction
- NOCACHE **-** Don't retain any cache

For a full specification of current focuseek SOAP API see *searchbox 2.0:* Reference Manual

3.3 User interface description

No specific user interface, see Module Administration Tool

3.4 Technical and Installation information

The installation procedure is the standard one for the focuseek searchbox 2.1 product. For details please see the “Getting Started” section of the User’s Guide at the address: http://www.focuseek.com/manuals/User/gettingstarted.html#id2529594

References to other major components needed	N.A.
Problems not solved	none
Configuration and execution context	

3.5 Draft User Manual

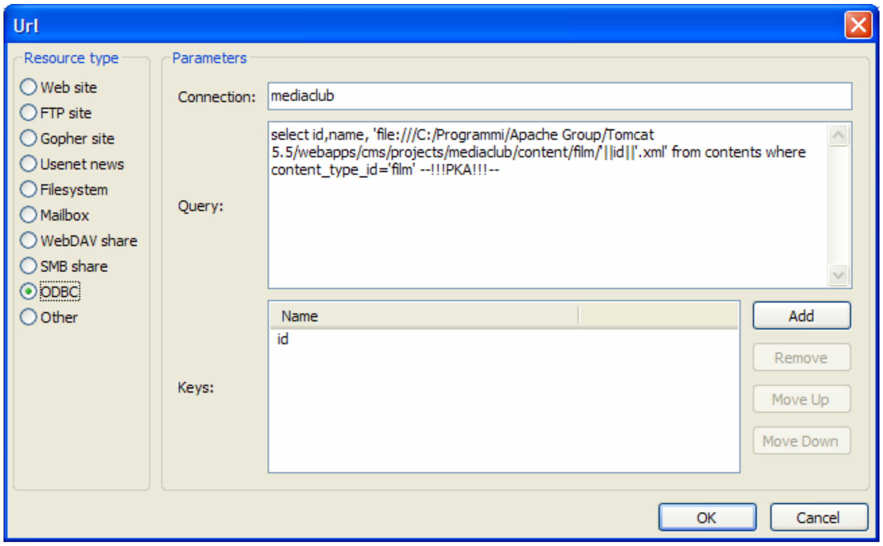
See the standard focuseek searchbox 2.1 documentation at the address: http://www.focuseek.com/manuals/User/index.html or in the attached documentation

3.6 Examples of usage

See the “Gathering” section of the standard focuseek searchbox documentation at the address: http://www.focuseek.com/manuals/User/gathering.html or in the attached documentation

The following screenshots show the complete configuration of an odbc and XML source.

STEP1: add a new ODBC source



Resource type:
○ Web site
○ FTP site
○ Gopher site
○ Usenet news
○ Filesystem
○ Mailbox
○ WebDAV share
○ SMB share
● ODBC
○ Other

Connection: mediaclub

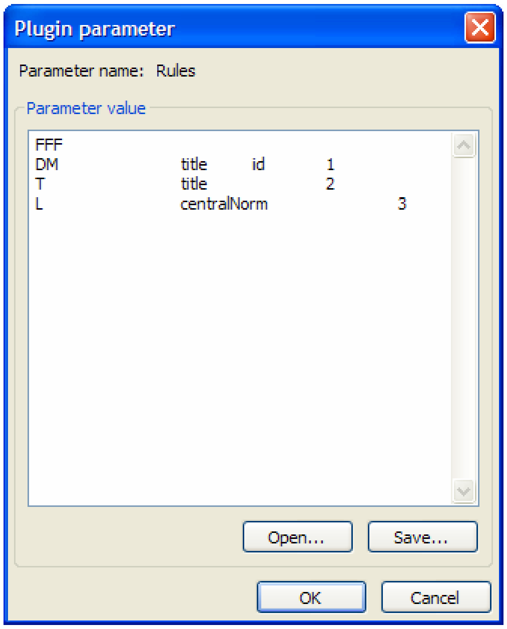
Query:

```
select id,name, 'file:///C:/Programmi/Apache Group/Tomcat
5.5/webapps/cms/projects/mediaclub/content/film/'||id||'.xml' from contents where
content_type_id='film' --!!!PKA!!!--
```

Keys:

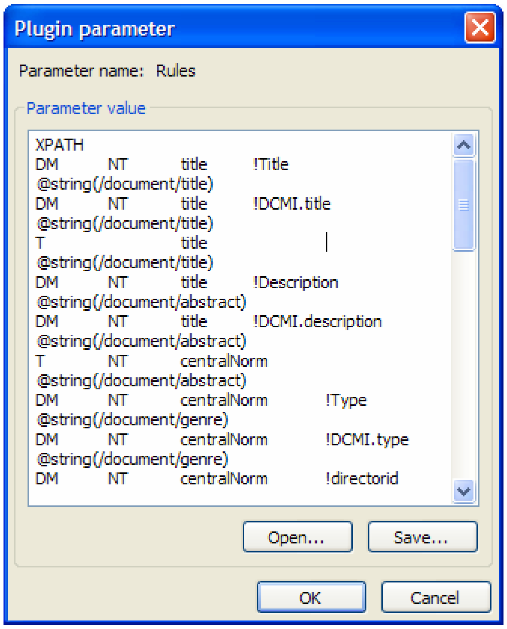
Name
id

STEP 2: configure the ODBC and XML plugins



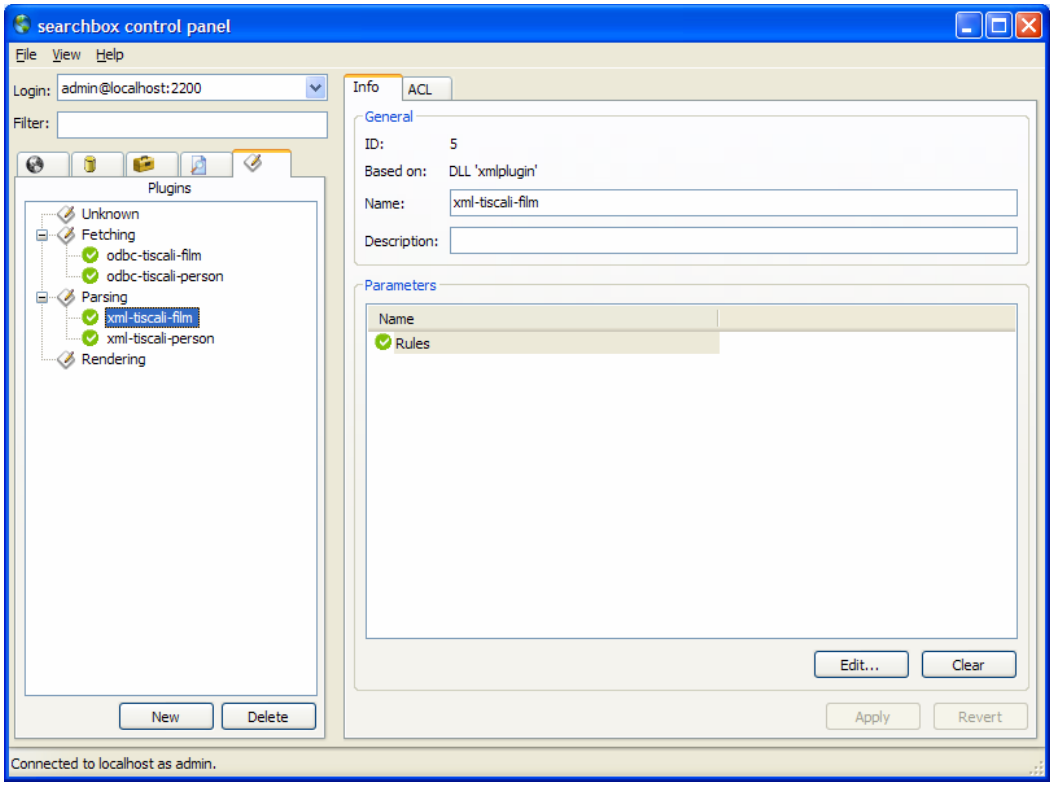
Parameter name: Rules

```
FFF
DM                  title      id          1
T                   title                  2
L                   centralNorm                      3
```



Parameter name: Rules

```
XPATH
DM        NT        title       !Title
@string(/document/title)
DM        NT        title       !DCMI.title
@string(/document/title)
T                   title
@string(/document/title)
DM        NT        title       !Description
@string(/document/abstract)
DM        NT        title       !DCMI.description
@string(/document/abstract)
T         NT        centralNorm
@string(/document/abstract)
DM        NT        centralNorm          !Type
@string(/document/genre)
DM        NT        centralNorm          !DCMI.type
@string(/document/genre)
DM        NT        centralNorm          !directorid
```



Login: admin@localhost:2200

Filter:

- Unknown
- Fetching
  - odbc-tiscali-film
  - odbc-tiscali-person
- Parsing
  - xml-tiscali-film
  - xml-tiscali-person
- Rendering

ID: 5

Based on: DLL 'xmlplugin'

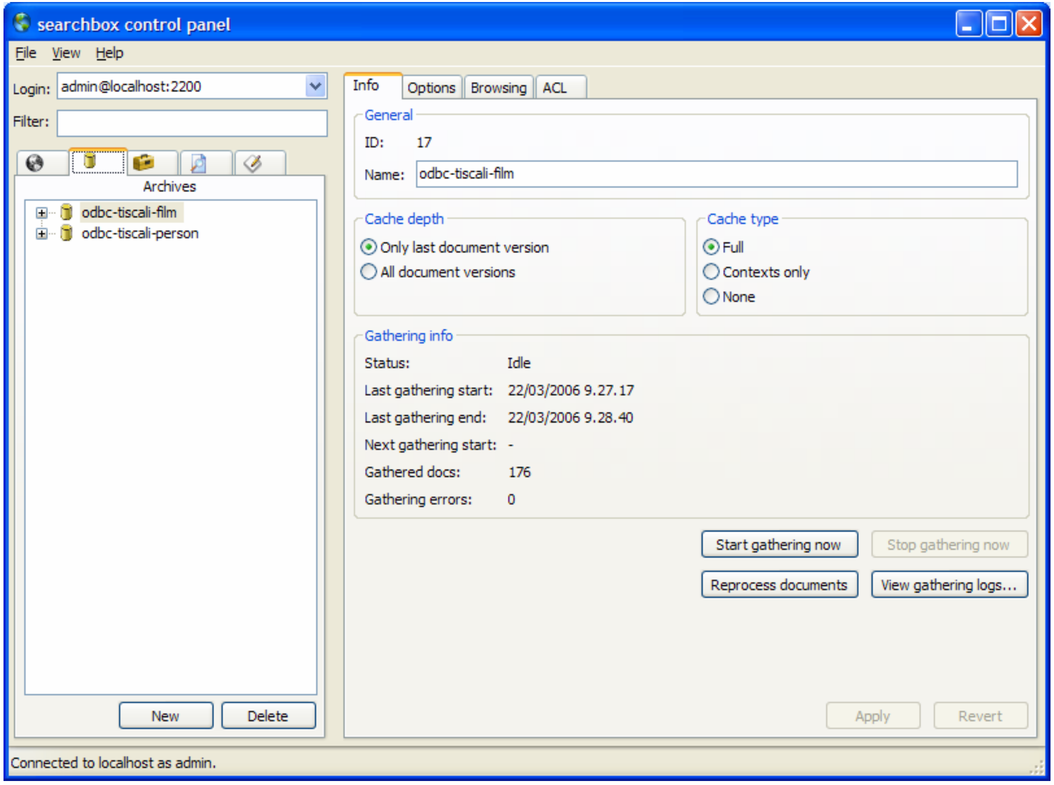
Name: xml-tiscali-film

Description:

Name
✓ Rules

Connected to localhost as admin.

STEP 4: start gathering



Login: admin@localhost:2200

Filter:

- odbc-tiscali-film
- odbc-tiscali-person

ID: 17

Name: odbc-tiscali-film

Cache depth: ● Only last document version ○ All document versions

Cache type: ● Full ○ Contexts only ○ None

Status: Idle

Last gathering start: 22/03/2006 9.27.17

Last gathering end: 22/03/2006 9.28.40

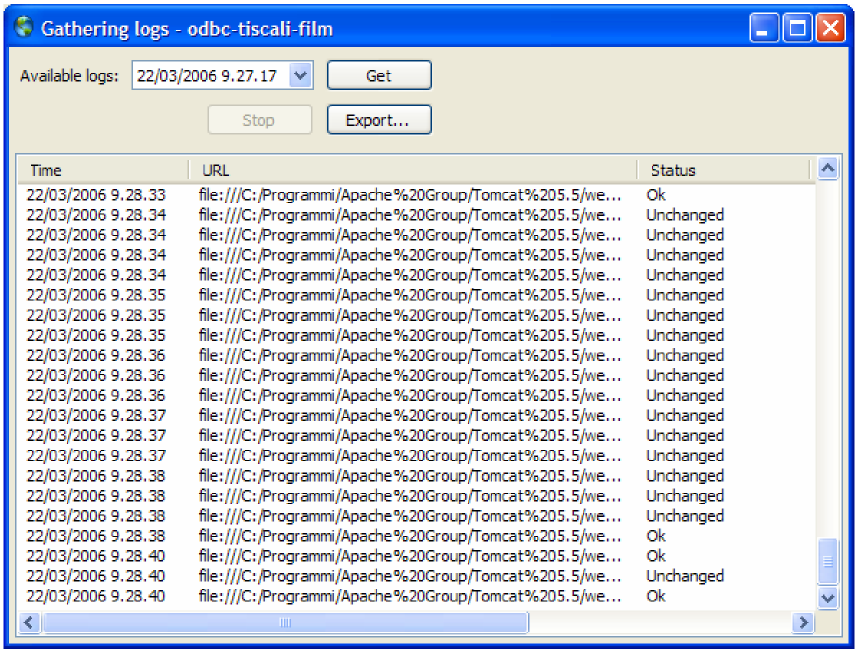
Next gathering start: -

Gathered docs: 176

Gathering errors: 0

Connected to localhost as admin.

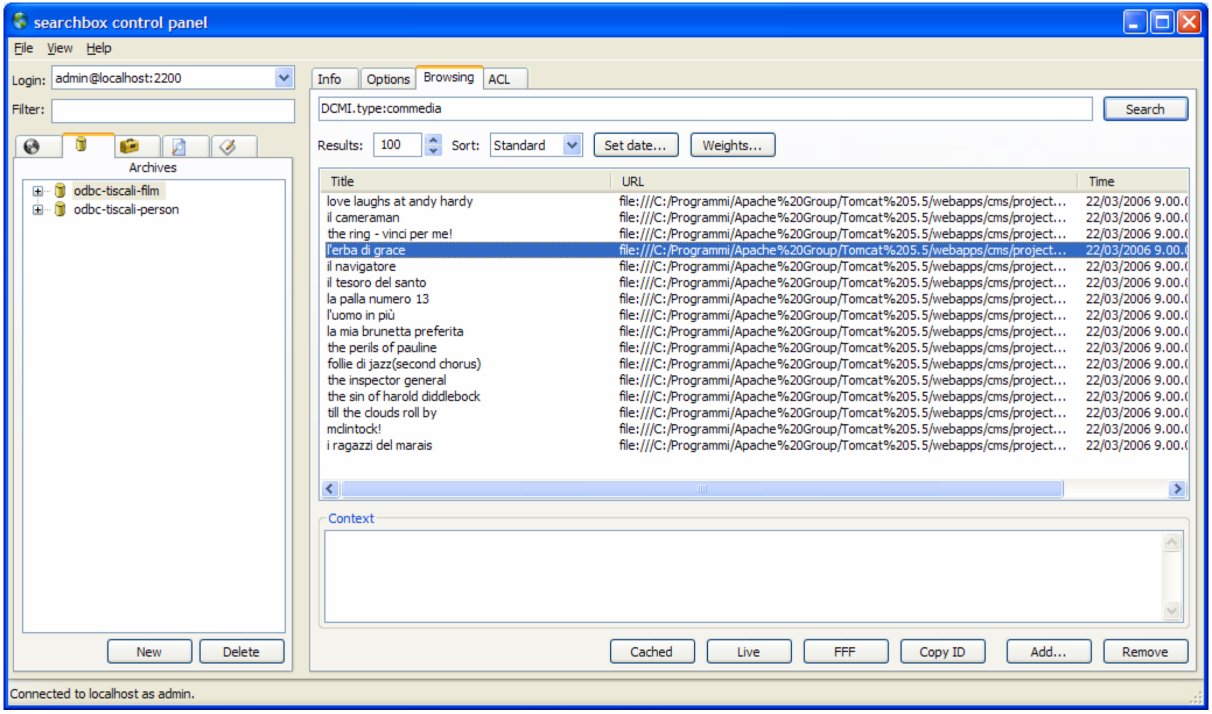
STEP 5: check gathering logs



Available logs: 22/03/2006 9.27.17

Time	URL	Status
22/03/2006 9.28.33	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Ok
22/03/2006 9.28.34	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.34	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.34	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.34	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.35	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.35	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.35	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.36	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.36	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.36	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.37	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.37	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.37	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.38	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.38	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.38	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.38	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Ok
22/03/2006 9.28.40	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Ok
22/03/2006 9.28.40	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Unchanged
22/03/2006 9.28.40	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/we...	Ok

STEP 6: perform a query (on metadata)



Login: admin@localhost:2200

Filter:

odbc-tiscali-film
odbc-tiscali-person

DCMI.type:commedia

Results: 100 Sort: Standard

Title	URL	Time
love laughs at andy hardy	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
il cameraman	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
the ring - vinci per me!	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
l'erba di grace	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
il navigatore	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
il tesoro del santo	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
la palla numero 13	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
l'uomo in più	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
la mia brunetta preferita	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
the perils of pauline	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
follie di jazz(second chorus)	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
the inspector general	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
the sin of harold diddlebock	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
till the clouds roll by	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
mclintock!	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.
i ragazzi del marais	file:///C:/Programmi/Apache%20Group/Tomcat%205.5/webapps/cms/project...	22/03/2006 9.00.

Connected to localhost as admin.

3.7 Integration and compilation issues

searchbox comes in a binary package available for all three supported platforms (Windows, Linux, Mac OS X)

3.8 Configuration Parameters

The main configuration parameters are defined in the focuseek.cfg configuration file as described in http://www.focuseek.com/manuals/User/administration.html#id906358 or in the attached documentation

3.9 Errors reported and that may occur

A list of all possible crawling errors is available at:
http://www.focuseek.com/manuals/User/gathering.html#id870604 or in the attached documentation

4 Module – Crawler Results Integrated database

<table>
<tr><th colspan="3">Module Profile</th></tr>
<tr><th colspan="3">Crawler Results Integrated database</th></tr>
<tr><td>Responsible Name</td><td colspan="2">Baldini</td></tr>
<tr><td>Responsible Partner</td><td colspan="2">Focuseek</td></tr>
<tr><td>Status (proposed/approved)</td><td colspan="2">approved</td></tr>
<tr><td>Implemented/not implemented</td><td colspan="2">Implemented</td></tr>
<tr><td>Status of the implementation</td><td colspan="2">100%</td></tr>
<tr><td>Executable or Library/module (Support)</td><td colspan="2">Executable</td></tr>
<tr><td>Single Thread or Multithread</td><td colspan="2">Multithread</td></tr>
<tr><td>Language of Development</td><td colspan="2">C++</td></tr>
<tr><td>Platforms supported</td><td colspan="2">Windows, Linux, Mac OS X</td></tr>
<tr><td>Reference to the AXFW location of the source code demonstrator</td><td colspan="2">https://cvs.axmedis.org/repos/Applications/collector_indexer</td></tr>
<tr><td>Reference to the AXFW location of the demonstrator executable tool for internal download</td><td colspan="2">https://cvs.axmedis.org/repos/Applications/collector_indexer</td></tr>
<tr><td>Reference to the AXFW location of the demonstrator executable tool for public download</td><td colspan="2"></td></tr>
<tr><td>Address for accessing to WebServices if any, add accession information (user and Passwd) if any</td><td colspan="2"></td></tr>
<tr><td>Test cases (present/absent)</td><td colspan="2"></td></tr>
<tr><td>Test cases location</td><td colspan="2">http:///////////////////</td></tr>
<tr><td>Usage of the AXMEDIS configuration manager (yes/no)</td><td colspan="2">no</td></tr>
<tr><td>Usage of the AXMEDIS Error Manager (yes/no)</td><td colspan="2">no</td></tr>
<tr><td>Major Problems not solved</td><td colspan="2">--
--</td></tr>
<tr><td>Major pending requirements</td><td colspan="2">--
--</td></tr>
<tr><td></td><td colspan="2"></td></tr>
<tr><td>Interfaces API with other tools, named as</td><td>Name of the communicating tools References to other major components needed</td><td>Communication model and format (protected or not, etc.)</td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
SQLite	Latest stable	Public domain
zlib	Latest stable	ZLIB license

4.1 General Description of the Module

It a component of focuseek platform containing the local copy of original content. For further details see the Collector Indexer component.

4.2 Module Design in terms of Classes

This module is part of the standard focuseek commercial package, a class diagram is not available.

4.3 User interface description

No specific user interface.

4.4 Technical and Installation information

The installation procedure is the standard one for the focuseek searchbox 2.1 product. For details please see the “Getting Started” section of the User’s Guide at the address: http://www.focuseek.com/manuals/User/gettingstarted.html#id2529594 or in the attached documentation

References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

4.5 Draft User Manual

See the standard focuseek searchbox 2.1 documentation at the address: http://www.focuseek.com/manuals/User/index.html or in the attached documentation

4.6 Examples of usage

See the “Gathering” section of the standard focuseek searchbox documentation at the address: http://www.focuseek.com/manuals/User/gathering.html or in the attached documentation

4.7 Integration and compilation issues

searchbox comes in a binary package available for all three supported platforms (Windows, Linux, Mac OS X)

4.8 Configuration Parameters

The main configuration parameters are defined in the focuseek.cfg configuration file as described in http://www.focuseek.com/manuals/User/administration.html#id906358 or in the attached documentation

4.9 Errors reported and that may occur

A list of all possible crawling errors is available at:
http://www.focuseek.com/manuals/User/gathering.html#id870604 or in the attached documentation

5 Module – Crawler Query Adapter / Collector Engine Query Support Interface

Module Profile	
Crawler Query Adapter / Collector Engine Query Support Interface	
Responsible Name	Baldini
Responsible Partner	Focuseek
Status (proposed/approved)	approved
Implemented/not implemented	Implemented
Status of the implementation	100%
Executable or Library/module (Support)	Module

Single Thread or Multithread	Multithread	
Language of Development	C++	
Platforms supported	Windows, Linux, Mac OS X	
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for public download		
Address for accessing to WebServices if any, add accession information (user and Passwd) if any		
Test cases (present/absent)	absent	
Test cases location		
Usage of the AXMEDIS configuration manager (yes/no)	no	
Usage of the AXMEDIS Error Manager (yes/no)	no	
Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		

User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
libcurl	Latest stable	CURL License
expat	Latest stable	Expat license
Henry Spencer regex library	Latest stable	Henry Spencer regex license
Apache Lucene	Latest stable	Apache license
SQLite	Latest stable	Public domain
libwv2	Latest stable	LGPL license
EasySOAP++	Latest stable	LGPL license
wxWidget	Latest stable	LGPL license
GCJ Runtime	Latest stable	LIBGCJ license
neon	Latest stable	LGPL
samba	Latest stable	GPL license
boost	Latest stable	BOOST license
iksemel	Latest stable	LGPL license
libxml2	Latest stable	MIT license
libxslt	Latest stable	MIT license
raptor	Latest stable	LGPL license
uwimap	Latest stable	University of Washington's Free-Fork License
zlib	Latest stable	ZLIB license

5.1 General Description of the Module

It a component of focuseek platform endowed of a specific interface (Crawler Query Adapter) to receive, from the Query Support, queries for searching into the Crawler Results Integrated Database and sending back the information to the AXMEDIS database manager. the local copy of original content. For further details see the Collector Indexer component.

5.2 Module Design in terms of Classes

This module is part of the standard focuseek commercial package, a class diagram is not available.

5.3 User interface description

No specific user interface.

5.4 Technical and Installation information

The installation procedure is the standard one for the focuseek searchbox 2.1 product. For details please see the “Getting Started” section of the User’s Guide at the address: http://www.focuseek.com/manuals/User/gettingstarted.html#id2529594 or in the attached documentation

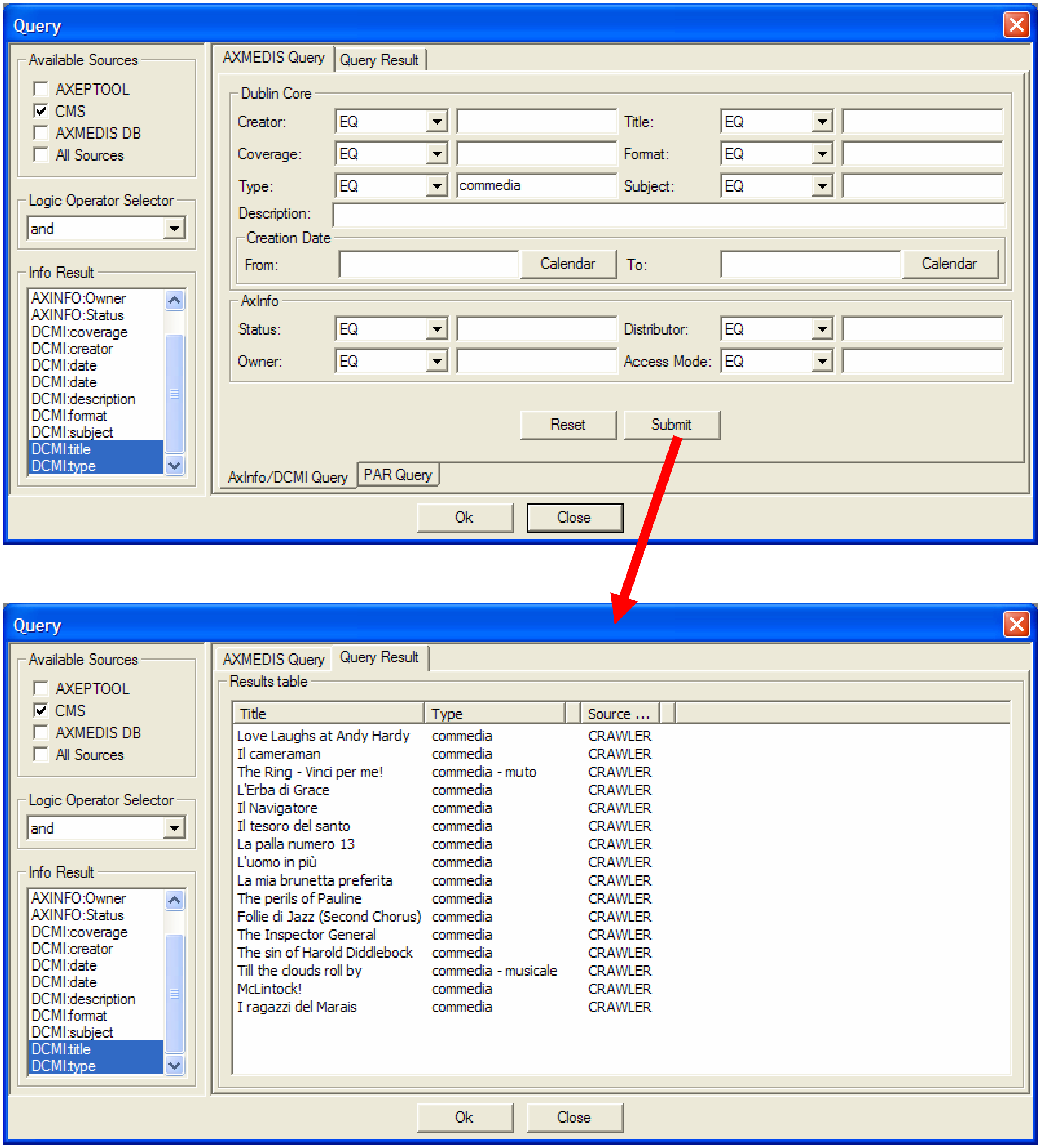
References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

5.5 Draft User Manual

This module implements the standard Axmedis Query Adapter inteface. Such interface is accessible from the http://HOST:2200/axsoap endpoint.

5.6 Examples of usage

The following screenshot shows an example of query from the axeditor and the obtained results.



Available Sources

- ☐ AXEPTOOL
- ☑ CMS
- ☐ AXMEDIS DB
- ☐ All Sources

Logic Operator Selector: and

Info Result: AXINFO:Owner, AXINFO:Status, DCMI:coverage, DCMI:creator, DCMI:date, DCMI:date, DCMI:description, DCMI:format, DCMI:subject, DCMI:title, DCMI:type

Dublin Core

Creator:	EQ		Title:	EQ	
Coverage:	EQ		Format:	EQ	
Type:	EQ	commedia	Subject:	EQ	

Description:

Creation Date — From: To:

AxInfo

Status:	EQ		Distributor:	EQ	
Owner:	EQ		Access Mode:	EQ	

Results table

Title	Type	Source ...
Love Laughs at Andy Hardy	commedia	CRAWLER
Il cameraman	commedia	CRAWLER
The Ring - Vinci per me!	commedia - muto	CRAWLER
L'Erba di Grace	commedia	CRAWLER
Il Navigatore	commedia	CRAWLER
Il tesoro del santo	commedia	CRAWLER
La palla numero 13	commedia	CRAWLER
L'uomo in più	commedia	CRAWLER
La mia brunetta preferita	commedia	CRAWLER
The perils of Pauline	commedia	CRAWLER
Follie di Jazz (Second Chorus)	commedia	CRAWLER
The Inspector General	commedia	CRAWLER
The sin of Harold Diddlebock	commedia	CRAWLER
Till the clouds roll by	commedia - musicale	CRAWLER
McLintock!	commedia	CRAWLER
I ragazzi del Marais	commedia	CRAWLER

5.7 Integration and compilation issues

searchbox comes in a binary package available for all three supported platforms (Windows, Linux, Mac OS X)

5.8 Configuration Parameters

None.

5.9 Errors reported and that may occur

None.

6 Module - Watch Manager

<table>
<tr><th colspan="3">Module/Tool Profile</th></tr>
<tr><th colspan="3">Watch Manager</th></tr>
<tr><td>Responsible Name</td><td colspan="2">Baldini</td></tr>
<tr><td>Responsible Partner</td><td colspan="2">Focuseek</td></tr>
<tr><td>Status (proposed/approved)</td><td colspan="2">approved</td></tr>
<tr><td>Implemented/not implemented</td><td colspan="2">implemented</td></tr>
<tr><td>Status of the implementation</td><td colspan="2">100%</td></tr>
<tr><td>Executable or Library/module (Support)</td><td colspan="2">Internal searchbox module</td></tr>
<tr><td>Single Thread or Multithread</td><td colspan="2">Multithread</td></tr>
<tr><td>Language of Development</td><td colspan="2"></td></tr>
<tr><td>Platforms supported</td><td colspan="2">Windows, Linux, Mac OS X</td></tr>
<tr><td>Reference to the AXFW location of the source code demonstrator</td><td colspan="2">https://cvs.axmedis.org/repos/Applications/collector_indexer</td></tr>
<tr><td>Reference to the AXFW location of the demonstrator executable tool for internal download</td><td colspan="2">https://cvs.axmedis.org/repos/Applications/collector_indexer</td></tr>
<tr><td>Reference to the AXFW location of the demonstrator executable tool for public download</td><td colspan="2"></td></tr>
<tr><td>Address for accessing to WebServices if any, add accession information (user and Passwd) if any</td><td colspan="2"></td></tr>
<tr><td>Test cases (present/absent)</td><td colspan="2">absent</td></tr>
<tr><td>Test cases location</td><td colspan="2"></td></tr>
<tr><td>Usage of the AXMEDIS configuration manager (yes/no)</td><td colspan="2">no</td></tr>
<tr><td>Usage of the AXMEDIS Error Manager (yes/no)</td><td colspan="2">no</td></tr>
<tr><td>Major Problems not solved</td><td colspan="2">--
--</td></tr>
<tr><td>Major pending requirements</td><td colspan="2">--
--</td></tr>
<tr><td></td><td colspan="2"></td></tr>
<tr><td>Interfaces API with other tools, named as</td><td>Name of the communicating tools References to other major components needed</td><td>Communication model and format (protected or not, etc.)</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
expat	Latest stable	Expat license
Henry Spencer regex library	Latest stable	Henry Spencer regex license
EasySOAP++	Latest stable	LGPL license
neon	Latest stable	LGPL
boost	Latest stable	BOOST license
iksemel	Latest stable	LGPL license
libxml2	Latest stable	MIT license
libxslt	Latest stable	MIT license
raptor	Latest stable	LGPL license

6.1 General Description of the Module

This component notifies about new contents to be imported in the AXMEDIS DB by a special object called Watch and available from focuseek. A Watch is defined by a set of rules (persistent queries on the index of crawled contents) so that every time a new content match this set of rules the Watch can notify to another

component that there is a new content to be imported. This is a native feature of focuseek so that only a configuration task must be performed using the standard Crawler User Interface.
This module has been customized in order to satisfy the AXMEDIS requirements so that in the current version is possible to send the notification to a generic Web Service. Such customizations are now integrated into the standard searchbox distribution. For details please se the focuseek documentation about Watch and SOAP notification at the address: http://www.focuseek.com/manuals/User/publishing.html#id2557105 or in the attached documentation

6.2 User interface description

No specific user interface for this module. The Watch administration interface can be found into the searchbox User's Guide manual at: http://www.focuseek.com/manuals/User/publishing.html#id2557105
or in the attached documentation

6.3 Technical and Installation information

No installation procedure required. This module comes packaged with the focuseek searchbox platform.

References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

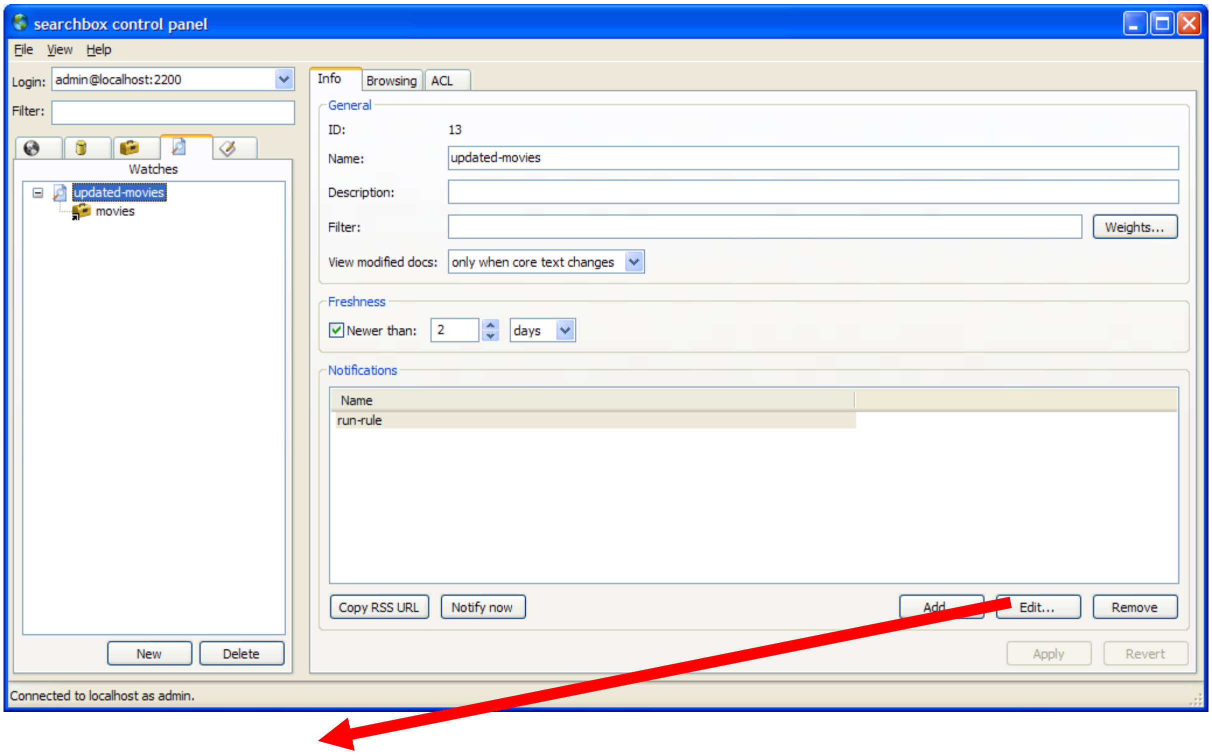
6.4 Draft User Manual

See the standard focuseek searchbox 2.1 documentation at the address:
http://www.focuseek.com/manuals/User/publishing.html#id2557105 or in the attached documentation

6.5 Examples of usage

See the “Publishing” section of the standard focuseek searchbox documentation at the address: http://www.focuseek.com/manuals/User/publishing.html#id2556814 (or in the attached documentation) where the procedure of creating a new watch is described.

The following screenshots show the configuration of a watch on the movies archive, which notifies any changes on the documents to the AXCP Scheduler giving the AXCP rule to be used for AXMEDIS object update.



Login: admin@localhost:2200

Filter:

updated-movies
movies

General

ID: 13

Name: updated-movies

Description:

Filter:

View modified docs: only when core text changes

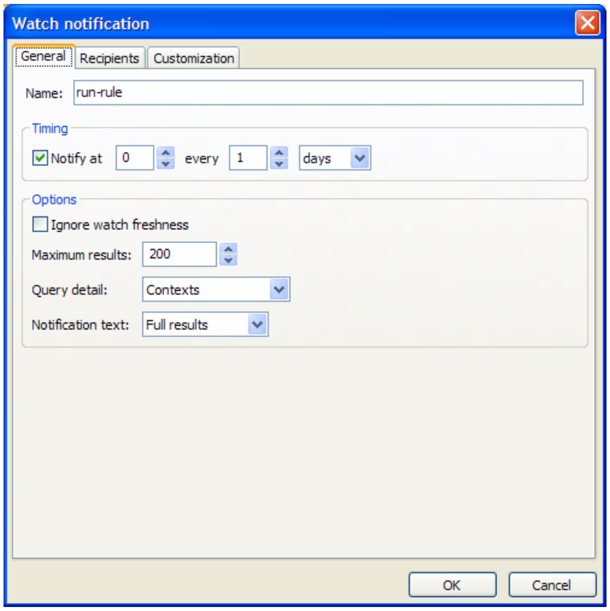
Freshness

☑ Newer than: 2 days

Notifications

Name
run-rule

Connected to localhost as admin.



Name: run-rule

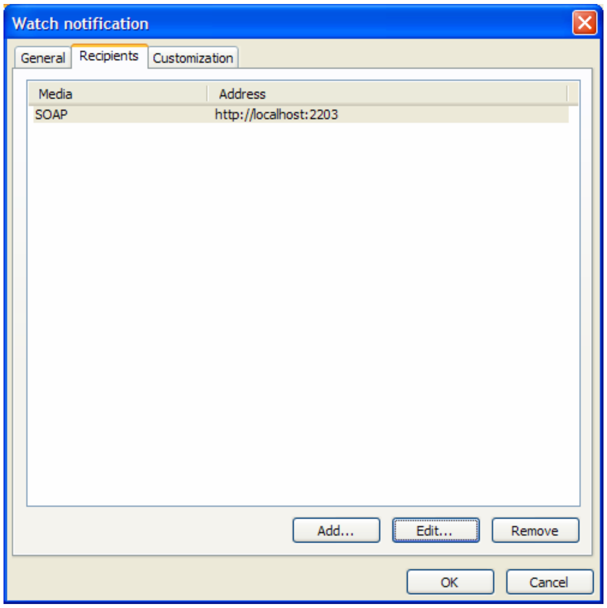
☑ Notify at 0 every 1 days

□ Ignore watch freshness

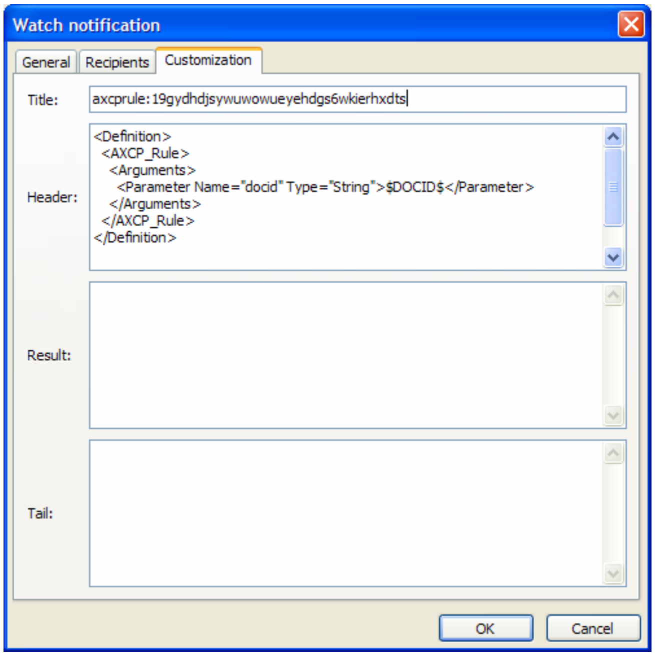
Maximum results: 200

Query detail: Contexts

Notification text: Full results



Media	Address
SOAP	http://localhost:2203



Title: axcprule:19gydhdjsywuwowueyehdgs6wkierhxdts

Header:

```
<Definition>
 <AXCP_Rule>
  <Arguments>
   <Parameter Name="docid" Type="String">$DOCID$</Parameter>
  </Arguments>
 </AXCP_Rule>
</Definition>
```

Result:

Tail:

In the above picture the message skeleton is defined giving the AXCP rule ID to apply and the related arguments (as XML structure), the macro $DOCID$ will be replaced with the focuseek internal document id of the updated document.

6.6 Configuration Parameters

The main configuration parameters are defined in the focuseek.cfg configuration file as described in http://www.focuseek.com/manuals/User/administration.html#id906358 or in the attached documentation

6.7 Errors reported and that may occur

None.

7 Executable Tool – Administration Tool

Module/Tool Profile	
Administration Tool	
Responsible Name	Nicola Baldini
Responsible Partner	focuseek
Status (proposed/approved)	approved
Implemented/not implemented	implemented
Status of the implementation	100%
Executable or Library/module (Support)	Executable
Single Thread or Multithread	Multithread
Language of Development	C++
Platforms supported	Windows, Linux, Mac OS X

Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer
Reference to the AXFW location of the demonstrator executable tool for public download	
Address for accessing to WebServices if any, add accession information (user and Passwd) if any	
Test cases (present/absent)	absent
Test cases location	
Usage of the AXMEDIS configuration manager (yes/no)	
Usage of the AXMEDIS Error Manager (yes/no)	
Major Problems not solved	-- --
Major pending requirements	-- --

Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		

User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
expat	Latest stable	Expat license
EasySOAP++	Latest stable	LGPL license
wxWidget	Latest stable	LGPL license
libxml2	Latest stable	MIT license
libxslt	Latest stable	MIT license

7.1 General Description of the Module

It is a client application to remotely manage using the SOAP interface the focuseek searchbox functionalities.

7.2 User interface description

The current focuseek searchbox Control Panel is fully compatible with the AXMEDIS requirements so that the official documentation can be used as reference for AXMEDIS too.
The searchbox User's Guide is available in the searchbox packege or online at the address: http://www.focuseek.com/manuals/User/index.html or in the attached documentation

7.3 Technical and Installation information

The Control Panel application is automatically installed with focuseek searchbox.

References to other major components needed	Collector Indexer, Watch Manager
Problems not solved	None
Configuration and execution context	

7.4 Draft User Manual

See the official focuseek doumentation at the address: http://www.focuseek.com/manuals/User/index.html or in the attached documentation

7.5 Examples of usage

See the official focuseek doumentation at the address: http://www.focuseek.com/manuals/User/index.html or in the attached documentation

7.6 Configuration Parameters

Config parameter	Possible values
Endopoint of searchbox WebService	An URL
PORT	The number of port used by the searchbox Web Service as configured into the focuseek.cfg configuration file
Username	A username
Password	A password

7.7 Errors reported and that may occur

Error code	Description and rationales
INTERNAL	The server reported an internal error
ACCESS_VIOLATION	You haven't the needed rights to modify this object
IN_USE	You cannot remove this item because it is currently in use
MAX_NUMBER_ARCHIVES	You have reached the maximum number of archives allowed by your license
CANT_LOAD_PLUGIN_DLLS	Server error loading plugin dlls
INVALID ENTITY	Invalid entity
ERROR_WRITING_ENTITY	Error writing configuration entity
ERROR_READING_ENTITY BAD_MAGIC UNSUPPORTED_MIMETYPE FILTER_MISMATCH MISSING_PARAMETER MAX_DOCUMENTS EXTPLUGINCONFIG DUPLICATED DISKFULL	Error reading configuration entity The object has been modified by another user Unsupported MIME type Filter mismatch Missing parameter Document corpus limit reached Server error configuring extended plugin Entity name already used Disk full

8 Module - Fast Access DB Interface

Module/Tool Profile	
Fast Access DB Interface	
Responsible Name	Nicola Baldini
Responsible Partner	Focuseek
Status (proposed/approved)	approved
Implemented/not implemented	implemented
Status of the implementation	100%
Executable or Library/module (Support)	Internal searchbox module
Single Thread or Multithread	Multithread
Language of Development	C++
Platforms supported	
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/....................

Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.	
Reference to the AXFW location of the demonstrator executable tool for public download		
Address for accessing to WebServices if any, add accession information (user and Passwd) if any		
Test cases (present/absent)		
Test cases location	http:///////////////////	
Usage of the AXMEDIS configuration manager (yes/no)	no	
Usage of the AXMEDIS Error Manager (yes/no)	no	
Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		

User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
expat	Latest stable	Expat license
EasySOAP++	Latest stable	LGPL license
wxWidget	Latest stable	LGPL license
zlib	Latest stable	ZLIB license

8.1 General Description of the Module

In order to retrieve large files from the documents cache a more efficient REST interface has been added to searchbox Engine.
Such interface is reachable at the endpoint: http://ENDPOINT/doc/DOCID where ENDPOINT is the address of the searchbox server (hostname and port) and DOCID is the document ID of a document stored by searchbox. For more information about the REST interface of searchbox see http://www.focuseek.com/manuals/Programmer/query-xml.html or in the attached documentation

8.2 User interface description

No user interface provided for this module.

8.3 Technical and Installation information

The module is installed by default with the focuseek searchbox package.

References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

8.4 Draft User Manual

See http://www.focuseek.com/manuals/Programmer/query-xml.html or in the attached documentation

8.5 Configuration Parameters

None.

8.6 Errors reported and that may occur

Standard HTTP errors.

9 Module - TAMINO/Lobster plugin

Module/Tool Profile

TAMINO/Lobster Plugin

Responsible Name	Nicola Baldini	
Responsible Partner	Focuseek	
Status (proposed/approved)	Approved	
Implemented/not implemented	Implemented	
Status of the implementation	Complete	
Executable or Library/module (Support)	Library	
Single Thread or Multithread	N.A.	
Language of Development	C++	
Platforms supported	Windows, linux, osx	
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for public download		
Address for accessing to WebServices if any, add accession information (user and Passwd) if any		
Test cases (present/absent)	absent	
Test cases location		
Usage of the AXMEDIS configuration manager (yes/no)	No	
Usage of the AXMEDIS Error Manager (yes/no)	No	
Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
Lobster CMS		
focuseek searchbox		
Formats Used	Shared with	format name or reference to a section

Protocol Used	Shared with	Protocol name or reference to a section
SOAP		
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
EasySOAP++	EasySOAP++ 20050326	LGPL
expat	Expat 1.95.7	Expat (MIT-like)
libxml2	Libxml2 2.6.19	MIT

9.1 General Description of the Module

The plugin enables focuseek searchbox to gather data from Lobster content management. To do so it extends searchbox tu understand a new class of URLs belonging to the lobs: scheme.

9.2 Module Design in terms of Classes

The implementation of this module is fully procedural. For details please refers directly to the provided source code.

9.3 User interface description

The plugin expands searchbox to intrpret a new class of URLs. These URLs are made of the following parts:

- The scheme, must be lobs:// or slobs:// The former instructs searchbox to contact the Lobster web service using SOAP/HTTP, the latter using SOAP/HTTPS.
- The hostname, optional port and path to the lobster endpoint. E.g. axmedis.learnexact.com/LobsterWebService/Lobster.wsdl
- A literal exclamation point (!)
- The Lobster domain (e.g. AXMEDIS)
- A literal colon (:)

- An XPath query for Lobster to perform. It must return one or more <manifest> tag each describing a Lobster package

The plugin, whose searchbox type is “lobster-plugin” requires no specific configuration and implements no searchbox plugins configuration parameters.

9.4 Technical and Installation information

The plugin is a DLL and is installed by copying it into the searchbox plugins diretroy on the searchbox server:

- on Windows: C:\Program Files\Focuseek\Searchbox\plugins
- on linux: /opt/searchbox/lib/platform/plugins
- on OSX: /Applications/searchbox/lib/platform/plugins

When the plugin is installed the user must instruct searchbox to use it; see “Configuration of a fetching plugin” in searchbox manual.

References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

9.5 Draft User Manual

After the plugin is installed the user must:

- Configure a searchbox source and enable the plugin on it
- Add a lobs: seed to the source
- Enable plain text authentication for the source
- Create an archive derived from the source
- Set the username and password for the archive to the right credentials required for Lobster access.

Source and archive configuration and the whole searchbox gathering process is described in detail in searchbox User’s Guide.

9.6 Examples of usage

To access the axmedis trial lobster account and extract the manifest file for the package with identifier “monet” use the following url:
lobs://axmedis.learnexact.com/LobsterWebService/Lobster.wsdl!AXMEDIS:manifest[@identifier=”monet”]

9.7 Integration and compilation issues

This plugin has been implemented using the standard SDK packaged with any focuseek searchbox distribution. A plugin is a standard DLL compiled with the default OS compiler (Microsoft Visual Studio 2003 for Windows and GCC 4 for Linux and Mac OS X) that include the focuseek-plugin.h file.

9.8 Configuration Parameters

See the above “Draft User Manual” and the Lobster Web Service documentation.

9.9 Errors reported and that may occur

A list of all possible errors is available at: http://www.focuseek.com/manuals/User/gathering.html#id870604 or in the attached documentation

10 Module - WebService Crawling Plugin

WebService Crawling Plugin		
Responsible Name	Nicola Baldini	
Responsible Partner	Focuseek	
Status (proposed/approved)	approved	
Implemented/not implemented	Implemented	
Status of the implementation	100%	
Executable or Library/module (Support)		
Single Thread or Multithread	N.a.	
Language of Development		
Platforms supported	Windows, linux, osX	
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer	
Reference to the AXFW location of the demonstrator executable tool for public download		
Address for accessing to WebServices if any, add accession information (user aNd Passwd) if any		
Test cases (present/absent)	absent	
Test cases location		
Usage of the AXMEDIS configuration manager (yes/no)	no	
Usage of the AXMEDIS Error Manager (yes/no)	no	
Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
focuseek searchbox		
Formats Used	Shared with	format name or reference to a section

Protocol Used	Shared with	Protocol name or reference to a section
SOAP		
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not
EasySOAP++	EasySOAP++ 20050326	LGPL
expat	expat 1.95.7	expat (MIT-like)

10.1 General Description of the Module

This module allows focuseek searchbox to gather data from a suitable SOAP web service. To do so it extends searchbox tu understand a new class of URLs belonging to the soap: scheme.

10.2 Module Design in terms of Classes

The implementation of this module is fully procedural. For details please refers directly to the provided source code.

10.3 User interface description

The plugin expands searchbox to interpret a new class of URLs. These URLs are made of the following parts:

- The scheme, must be soap://
- The SOAP endpoint to call, %-encoded. E.g. localhost:8080
- A slash (/)
- The BASE parameter for the service, %-encoded.
- A slash (/)

The plugin, whose searchbox type is “ax-soap-plugin” requires no specific configuration and implements no searchbox plugins configuration parameters.

10.4 Technical and Installation information

The plugin is a DLL and is installed by copying it into the searchbox plugins diretroy on the searchbox server:

- on Windows: C:\Program Files\Focuseek\Searchbox\plugins
- on linux: /opt/searchbox/lib/platform/plugins
- on OSX: /Applications/searchbox/lib/platform/plugins

When the plugin is installed the user must instruct searchbox to use it; see “Configuration of a fetching plugin” in searchbox manual.

References to other major components needed	
Problems not solved	•
Configuration and execution context	

10.5 Draft User Manual

The called service must implement just two calls: EnumerateDocuments and GetDocument.

In the following discussion please refer to ax-crawlable-service.wsdl for call names and parameters.

The EnumerateDocuments call

The aim of this call is to get a [subset of the full] list of the documents accessible through the service. This call returns only enough information to let allow subsequent requests to retrieve the documents and not the documents contents themselves; in a sense the returned information is the *name* of the document. The input parameter *base* is an opaque string the service can choose to ignore or interpret in a custom way (i.e. to select one among multiple document domains). The output parameter *keys* is a list of strings, each one suitable to be passed to the GetDocument call as a document key.

The GetDocument call

The GetDocument call retrieves the document contents. It accepts the same *base* parameter the EnumerateDocuments call does and a single string, *key*, requesting a specific document. The implementor **must not** assume that only keys returned by a previous EnumerateDocuments call are passed to GetDocument, although this is certainly the most common case.

The GetDocument call returns a single string, *xmldoc*. This is an XML representation of the document contents and the associated metadata. See the document “The FFF document format” for more details on this format.

10.6 Examples of usage

To connect to a SOAP service whose endpoint is “http://myserver:8080/myservice” passing it “MYBASE” as the base use the following url:

soap://myserver%3a8080%2fmyservice/MYBASE/

10.7 Integration and compilation issues

This plugin has been implemented using the standard SDK packaged with any focuseek searchbox distribution. A plugin is a standard DLL compiled with the default OS compiler (Microsoft Visual Studio 2003 for Windows and GCC 4 for Linux and Mac OS X) that include the focuseek-plugin.h file.

10.8 Configuration Parameters

See the above “Draft User Manual” and the Lobster Web Service documentation.

10.9 Errors reported and that may occur

A list of all possible errors is available at: http://www.focuseek.com/manuals/User/gathering.html#id870604 or in the attached documentation

11 Module - MIME type Parsing Plugin

Module/Tool Profile	
MIME type Parsing Plugin	
Responsible Name	Nicola Baldini
Responsible Partner	
Status (proposed/approved)	
Implemented/not implemented	Implemented
Status of the implementation	Complete
Executable or Library/module (Support)	
Single Thread or Multithread	N.a.
Language of Development	
Platforms supported	Windows, linux, osX
Reference to the AXFW location of the source code demonstrator	https://cvs.axmedis.org/repos/Applications/collector_indexer
Reference to the AXFW location of the demonstrator executable tool for internal download	https://cvs.axmedis.org/repos/Applications/collector_indexer
Reference to the AXFW location of the demonstrator executable tool for public download	
Address for accessing to WebServices if any, add accession information (user aNd Passwd) if any	
Test cases (present/absent)	absent
Test cases location	
Usage of the AXMEDIS configuration manager (yes/no)	no
Usage of the AXMEDIS Error Manager (yes/no)	no

Major Problems not solved	-- --	
Major pending requirements	-- --	
Interfaces API with other tools, named as	Name of the communicating tools References to other major components needed	Communication model and format (protected or not, etc.)
focuseek searchbox		
Formats Used	Shared with	format name or reference to a section
Protocol Used	Shared with	Protocol name or reference to a section
Used Database name		
User Interface	Development model, language, etc.	Library used for the development, platform, etc.
Used Libraries	Name of the library and version	License status: GPL. LGPL. PEK, proprietary, authorized or not

11.1 General Description of the Module

This module allows the user to force focuseek gathering even for documents belonging to a MIME type that searchbox cannot parse.

11.2 Module Design in terms of Classes

The implementation of this module is fully procedural. For details please refers directly to the provided source code.

11.3 User interface description

The user configures the plugin through searchbox controlpanel and provides a single configuration parameter, “MIMEList”, containing the list of allowed MIME types, one on each line. For details see the chapter “Configuring a Fetching plugin” in searchbox user manual.

11.4 Technical and Installation information

The plugin is a DLL and is installed by copying it into the searchbox plugins diretroy on the searchbox server:

- on Windows: C:\Program Files\Focuseek\Searchbox\plugins
- on linux: /opt/searchbox/lib/platform/plugins
- on OSX: /Applications/searchbox/lib/platform/plugins

When the plugin is installed the user must configure the allowed MIME types (see above) and instruct searchbox to use it; see “Configuration of a fetching plugin” in searchbox manual.

References to other major components needed	Collector Indexer
Problems not solved	none
Configuration and execution context	

11.5 Draft User Manual

After the plugin is installed the user must:

- Configure a searchbox source and enable the plugin on it or
- Enable the plugin on an existing source

Source and archive configuration and the whole searchbox gathering process is described in detail in searchbox user manual.

11.6 Examples of usage

If the user want to force searchbox to gather images in gif, tiff and jpeg formats she can use the following value for the MIMEList configuration parameter:

```
image/gif
image/jpeg
image/tiff
```

A suitable name for this configuration could be “Gather images”

11.7 Integration and compilation issues

This plugin has been implemented using the standard SDK packaged with any focuseek searchbox distribution. A plugin is a standard DLL compiled with the default OS compiler (Microsoft Visual Studio 2003 for Windows and GCC 4 for Linux and Mac OS X) that include the focuseek-plugin.h file.

11.8 Configuration Parameters

Config parameter	Possible values
MIMEList	A list of MIME types searchbox should gather anyway, one for each line.

11.9 Errors reported and that may occur

A list of all possible errors is available at: http://www.focuseek.com/manuals/User/gathering.html#id870604

12 Provided API named SOAP

See the searchbox Programmers’s Guide at:
http://www.focuseek.com/manuals/Programmer/integrating.html#soap-api or in the attached documentation

13 Formal description of format FFF

See the searchbox Programmers’s Guide at: http://www.focuseek.com/manuals/Programmer/fff-xml.html or in the attached documentation

14 Formal description of communication protocol SOAP

See the official specification at: http://www.w3.org/TR/2002/WD-soap12-part1-20020626/

15 Formal description of communication protocol REST

Some resurces can be fond at: http://en.wikipedia.org/wiki/Representational_State_Transfer

16 Bibliography

- Searchbox User’s Guide at: http://www.focuseek.com/manuals/User/index.html
- Searchbox Programmer’s Guide at: http://www.focuseek.com/manuals/Programmer/index.html