



## **Software Product Description**

AMI Knowledge Discovery version 5.0

SPD-AMIKD-50-ENG v1.0

February 2008

# Summary

1	Approval.....	4
2	New name .....	4
3	Principal Developments .....	4
4	Introduction .....	7
5	Functional Description .....	8
5.1	Taking into account the sources of information.....	8
5.1.1	Standardisation and Federation.....	8
5.1.2	Interrogation of Sources .....	9
5.1.3	Source Structure and Document Format .....	11
5.1.4	Security and Confidentiality.....	12
5.1.5	Dynamic Information .....	13
5.1.6	External Sources .....	14
5.2	Search .....	14
5.2.1	Search Creation or Request.....	14
5.2.2	Restrictions .....	15
5.2.3	Request Parameters .....	15
5.2.4	Search By Example .....	17
5.3	Retrieval.....	18
5.3.1	Results Presentation .....	18
5.3.2	Relevance.....	19
5.3.3	Correlated Information.....	20
5.3.4	Categorisation and Clusterisation .....	21
5.3.5	Retrieval of Document Information.....	22
6	Architecture .....	22
6.1	Overview .....	22
6.2	Services .....	23
6.3	Compatibility.....	24
6.4	Supported Environments .....	24
7	Glossary.....	24
8	Documentation and Contacts .....	26

## **Copyright**

©Copyright 2003-2008 Go Albert SA

This document is protected by copyright. The product is distributed with a license that limits its use, copy, distribution and re-compilation.

This document may not be reproduced, in whole or in part in any form and in any manner whatsoever, without the prior approval of AMI.

This document is subject to updates without advance notification.

## **Trade Marks and Patents**

albert®, the logo albert, AMI™, Albert Meaning Interpreter™, What you mean is what you get™ and meaning bus are trade marks of Go Albert SA

## **Patents**

U.S. Patent n° 6-446-064 Enhancing e-commerce using natural language interface .

U.S. Patent n° 6-594-657 Enhancing Online Support.

U.S. Patent n° 6-598-039 Natural Language Interface.

All other product names are trademarks of their respective owners.

## ➔ Overview

This document constitutes a functional description of the product **AMI Knowledge Discovery v5.0. (reference SPD-AMIKD-50-ENG)**

### 1 Approval

Approved by	Function	Date
Alain Beauvieux	CEO	12 <sup>th</sup> February 2008
Eric Fourboul	CTO	12 <sup>th</sup> February 2008

### 2 New name

**AMI Knowledge Discovery version 5** is the culmination of research and development work that started back in 2000. Successive developments have led us to design a software application for the management and use of knowledge which is no longer just a search engine.

The version 5 product has, up until now, been called AMI Enterprise Discovery and this upgrade represents an important phase in the products development reflected in the name change from "AMI Enterprise Discovery" to "AMI Enterprise Knowledge" and acknowledging that the problem is no longer just about searching for information. By maintaining the sequential version numbering AMI makes a statement that the product "Knowledge Discovery" continues to build on a legacy of more than seven years of work in this area

### 3 Principal Developments

**AMI Knowledge Discovery v5.0** includes a host of significant functional additions and improvements to the earlier version 4 that was released in May 2006

In particular this new version offers fully integrated compatibility with the AMI Enterprise Intelligence v4.0 suite which is described in the document SPD-AMEI-40-ENG available for download via our website at [www.amisw.com](http://www.amisw.com) (see "Documentation" tab)

New features of **AMI Knowledge Discovery v5.0** are described in this section. Other sections of this document are unchanged.

## Connection and Indexation

**AMI Knowledge Discovery v5.0** offers a unique indexing technology based on our patented “document signature” calculation technology.

While most search engines can index documents **AMI Knowledge Discovery v5.0** goes further in its capacity as a meta-engine operator in connection to external sources.

This technology, widely used in applications such as Competitive Intelligence solutions, provides great flexibility in the way in which external sources are managed or if target content, for whatever reason, cannot otherwise be easily indexed. Hence, rather than developing complex mechanisms to retrieve data often with no real guarantee of perfect synchronisation **AMI Knowledge Discovery v5.0** can use existing search engines provided by the target content management system.

By combining the two options, indexing and high-level connection to sources (without the need for indexing) **AMI Knowledge Discovery v5.0** offers many opportunities in terms of the exploitation of information sources, whether internal or external. In particular, each user has a single point of entry from which to search both inside and outside the organisation.

In addition the **AMI Knowledge Discovery v5.0** Administrator can build clusters of sources to provide a search environment that is both universal and “secured”.

## Personalised Interfaces

**AMI Knowledge Discovery v5.0** allows the creation of personalised dashboards presenting information from multiple sources, both internal and external, based very specifically on a users centres of interest and making “comparative” reading easy. In the same way, if sources referenced are external then **AMI Knowledge Discovery v5.0** can acts as a meta-engine to complement other sources already indexed. Results can be presented in any number of ways, by document type, by source etc.

## Automatic Search Agents

Benefiting from AMI's substantial experience in the field of automated data collection **AMI Knowledge Discovery v5.0** offers an optional module allowing a user to request a query to be executed repeatedly. **AMI Knowledge Discovery v5.0** manages the execution of this request at user defined time intervals. Results can be presented via a dashboard as described above.

## Automatic Profiling

Based on AMI's patented "document signature" technology **AMI Knowledge Discovery v5.0** offers a unique feature to automatically model individual user interests from defined actions such as the recommendation of a document. In this way **AMI Knowledge Discovery v5.0** is able to propose new documents to a user as soon as they are indexed.

This feature is in itself a very important development since the engine is no longer just a "searcher" of documents but provides information of interest to the reader automatically without the need for the user to ask. Such a feature is essential in collaborative work projects where often the sharing of information is based on a prior knowledge of its existence. In delivering a system of automatic profiling **AMI Knowledge Discovery v5.0** provides a very effective solution to this problem.

## Modular and Scalable Interface

**AMI Knowledge Discovery v5.0** provides an interface built around functional objects that can be easily assembled and configured to match a user or organisation's exact requirements. A large library of tools is available to facilitate navigation within Results: extracting named entities, sorting by date, source, language, level of relevance etc with the ability to query in natural language or Boolean. Bespoke interfaces delivering user defined function requirements can be constructed within a few hours.

## Back Office

The "Log Manager" system has been completely revised and expanded. It is now possible for the **AMI Knowledge Discovery v5.0** Administrator to use AMI's statistical analysis tools to build dashboards for monitoring system performance and follow-up actions

These dashboards may involve elements of semantic analysis on submitted requests themselves and the language employed by users allowing the Administrator to better understand user expectations in terms of content.

## 4 Introduction

### ***Search for Information across the Enterprise***

With the exponential growth in information available to an employee in any type of organisation comes the requirement to have at their disposal the necessary tools for finding information. The requirements of Search tools, once limited by the volume of data they could handle and accessibility issues, are now multiple.

- ✓ Information that may be internal or external to the Enterprise ;
- ✓ multiple source types that are geographically dispersed ;
- ✓ multiple languages ;
- ✓ multiple existing information systems with varying degrees of current integration ;
- ✓ differences in file formats ;
- ✓ quality and compatibility of content across disparate systems ;
- ✓ multiple levels of security ;
- ✓ categorisation of documents and results.

Not only must the structure of Sources and information contained within them be taken into account, but also the organisation and presentation of the results or the choice of proposed results must also be adapted.

As Intranets and Portals become places of research rather than mere information repositories the ability to find and manage information intelligently is essential. The purely technical approach consisting of providing lists of url's corresponding to keywords is no longer acceptable. As the volume of information available online continues to increase at an unprecedented rate and accessibility to certain key sources becomes harder the Organisation must now implement central solutions that are capable of dealing effectively with information retrieval both now and looking forwards.

**AMI Knowledge Discovery v5.0** is the solution described in this document.

### **Organisation of this document**

This document is organised in the following way :

- ✓ The first part, a functional description, details the needs identified by the company regarding the search for information, and each solution delivered by **AMI Knowledge Discovery v5.0**. This first part is itself subdivided into three categories.
  - ⇒ “taken-into-account-of” environment : existing, sources of information, indexation, etc.
  - ⇒ search : language of interrogation and method.
  - ⇒ delivery of results to the user : relevance and classification.
- ✓ In the second part the more technical questions are answered..
- ✓ A glossary is included allowing quick reference to the subjects which may interest the reader more specifically.

## AMI Knowledge Discovery v5.0

**AMI Knowledge Discovery v5.0** is a software application that can be integrated into a Portal or an Intranet for the purpose of providing a universal and intelligent means by which to index and search via an existing or customised user interface.

### Website Access

**AMI Knowledge Discovery v5.0** offers all users federated access to all sources of information within the company. Users can be described as anyone who uses the Organisation's internal information systems such as an Intranet or Document Management System for example.

It is possible that these sources may also contain information that the organisation wants to make available to people externally: e.g. the general public, customers, or partners. The Organisation may also have a website that takes the form an extranet on which it publishes internal information, possibly including external information or data flows from a third party. (e.g. newsfeeds)

With **Website Access**, all the advanced functions offered by **AMI Knowledge Discovery v5.0** can also be made available to website visitors. **Website Access** is the Web-optimised version of the software which is adapted specifically to take account of the different conditions of use to that of an Intranet: e.g. very large numbers of users, not always via a login, cookie management, etc

Via the analysis of **AMI Knowledge Discovery v5.0's** request "logs" **Website Access** allows a much better understanding of customer and visitor behaviour, a clearer understanding of their requirements and the installation of preferential pages if required which overall can deliver greatly improved levels of service and customer satisfaction.

## 5 Functional Description

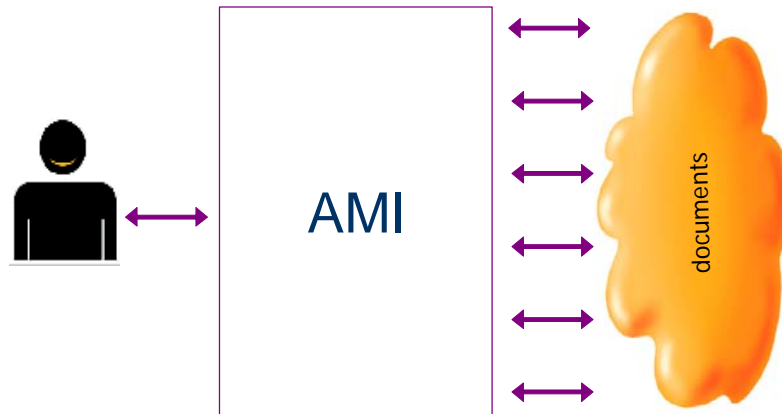
### 5.1 Taking into account the sources of information

Regarding searches of unstructured information the potential sources are innumerable:

- ✓ Dedicated applications;
- ✓ Multiple legacy systems and file formats ;
- ✓ Public and Protected sources ;
- ✓ Static and Dynamic information ;
- ✓ Internal and External sources...

#### 5.1.1 Standardisation and Federation

To allow all types of sources to be taken into account in a universal way, **AMI Knowledge Discovery v5.0** provides a standardising element which makes it possible to regard all sources in a single way whilst preserving their specific characteristics. This element is called AMI, the Automatic Meaning Interpreter™



A user of **AMI Knowledge Discovery v5.0** talks exclusively with AMI. It is AMI which assumes responsibility for accessing the diversity of information types and sources available.

Once the access point to information is standardised, AMI can be presented internally as the single point of access to all information within the Organisation. The user only needs one interface to reach and benefit from all the potential value held within each source.

### 5.1.2 Interrogation of Sources

In order to determine if information is useful to the user the system will first question the source. However if there are sources which offer their own native mechanism of interrogation there is also a possibility that the source may not allow external indexing.

In this case, **AMI Knowledge Discovery v5.0** will create an index for one or more sources, and will question this index.

#### 5.1.2.1 Indexation

##### Indexer and Synchroniser

The Indexer and the Synchroniser are the indexing tools which feed the index tables.

- ✓ The Indexer is **AMI Knowledge Discovery v5.0's** active indexing mechanism. It can cover multiple addresses and update an existing reversed index. It is active insofar as it carries out the indexing process dynamically.
- ✓ The Synchroniser is **AMI Knowledge Discovery v5.0's** passive indexing element. It is a service applied to an http server and allows easy updating of the index.

The AMI index allows the user not only direct access to information, but also the addition of metadata assigned to each document on which semantics and extraction of key entity and concept data is generated; e.g. origin of the document, creation date, category, etc...

*Example* An application whose documents content match with a users search requests can be referenced as a source using the Synchroniser. Its indexing is made possible by means of a dedicated program called an agent which will include, within the AMI index, all the references held within the source application useful to the user.

### **Site Crawler**

**AMI Knowledge Discovery v5.0** includes functionality to “crawl” external information sources and supports the principal current protocols: http, https, ftp, nntp, file.

*Example* A Web site, a file containing files, are sources of this type. Function and control of the AMI Crawler are described in the paragraph entitled “The Generic Connector”.

### **Acquisition and preservation of the cache**

AMI can also, as an option, preserve in its indices a textual representation of indexed information (the cache) capable of being exploited at the end of a search. This can be very useful when the information at point of origin is no longer accessible or for any other reason not available. Note however that use is within the limits permitted by the laws on copyright and royalties (see paragraph 3.1.6, page 11 on external sources).

#### **5.1.2.2 Interrogation**

When the source offers a means of interrogation itself it is not necessary to index it: in this case AMI will connect directly to the source’s search function to question it. The AMI component used is called a “Connector”.

*Example* A typical example might be as follows:

#### **Yahoo Connector!**

Via our international partnership with Yahoo, AMI can question the Yahoo engine in a number of different ways, for instance;

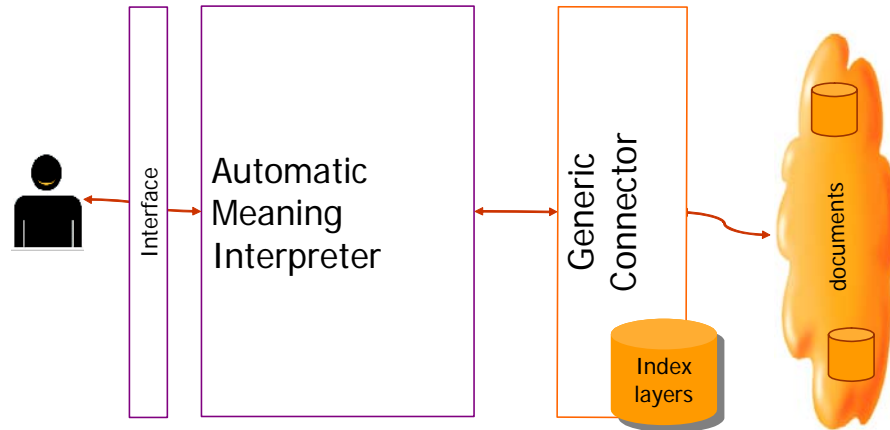
- Searching the world wide web
- Connection to dynamic news feed currently referencing more than 7000 sources.

#### **The Generic Connector**

The Generic Connector is a protocol. It provides a standardised http/xml flow to allow dialogue between AMI and a remote system likely to provide relevant information.

The use of scripts conforming to this protocol guarantees the ability of the system to be connected to sources such as the “invisible” Web or specific applications such as Content and Document Management systems.

To date several hundred connectors have been catalogued across a large number of projects carried out by AMI Partners and Customers.



## Agents

When a source of information cannot be indexed directly by the AMI crawler, and when it cannot be questioned by a connector, an agent can be created which will insert within an index the elements of indexing of each document.

The generic connector and respective “agents” are at the centre of possibilities for federated access to information offered by AMI. All sources of online information are potentially accessible whether the information is unstructured (e.g. Websites, Content Management Systems) or structured (e.g. Databases).

### 5.1.3 Source Structure and Document Format

Original documents are converted during the process of indexing into a standard recognised format on which various linguistic analysis routines can be applied.

Document formats most commonly used by AMI include: Microsoft Word, TXT, XML, HTML, PDF and Postscript however, in total, AMI recognises more than 360 different file formats including all Office formats as well as the majority of legacy file formats.

### Analysis of Content

Specific routines or scripts can be applied at the time of filtering in order to guide the analysis carried out, influence the indexing or to allocate to a document metadata which will be available during the compilation of the results.

## Signature Calculation

For each document indexed by AMI its signature is inserted into the index. The signature of a document is an internal representation, calculated by AMI, used to determine the essential textual contents of the document, i.e. the subject which the document really “speaks” about. Thus, the signature is used in the calculation of the relevance of a document compared to a request.

## HTTP Sites and HTML Pages

The mechanism of a source Descriptor make it possible to give the indexer precise directives as to the way in which sources are crawled thus making it possible to address questions that other systems simply cannot such as: access to “pop-up” pages, how Javascript is treated and the optimisation of indexing times.

The Descriptor also makes it possible to dictate the way in which AMI manages processes such as “links” found in source pages (e.g. to follow the links or not, to index the page or not etc).

In the same way, the case of HTML documents is particular: this type of document is rarely standard in structure; on the contrary, its page-format is often intended to promote, from within the page, different but distinctly linked elements for example guided towards associated advertising information.

However the actual subject of research, the important information, is very often only the central textual part of the overall content that may be present on a page.

This is why **AMI Knowledge Discovery v5.0** provides a mechanism of page “description” which is used to specify the parts of the pages that need to be taken into account.

Descriptors also make it possible to store login’s and passwords that may be required to access certain sites such that any searches conducted can include, in a universal way, the often more valuable information contained on “invisible” sources, i.e. those sources that are beyond the reach of public search engines.

Filters are applicable to all the various types of content, the Descriptors tending to be reserved more for use in relation to HTML pages.

## Languages

As standard **AMI Knowledge Discovery v5.0** implements algorithms that automatically index in English, French, German, Spanish, Portuguese, Italian and Dutch.

The core AMI technology is developed in respect of Unicode standard encoding which allows easy integration of new languages including non-Western.

### 5.1.4 Security and Confidentiality

When documents are publicly available they are easy to reach via an address (URL, file name) returned by AMI.

In many cases however not all users of the system have the right to view all documents. N.B.; When a user does not have the right to access a document AMI does not bring its existence to their attention.

**AMI Knowledge Discovery v5.0** offers wide ranging flexibility allowing an Administrator to adapt each configuration to suit many different situations permitting a user access only to those documents for which they have the rights to view. AMI respects all existing system and user rights of access.

The two approaches most often adopted are as follows:

#### **Known Rights, not changing often**

When the various levels of access rights to a document are known in advance and the content is not subject to regular changes it is prudent to include this level in the index.

The application implementing AMI will of course needs to identify the user who submits the request and to apply a restriction to the metadata applied specifically to the rights to show only the permitted/authorised results.

This is possible when the indexing is conducted by AMI.

#### **Secure or Closed System**

In the case of a secure or closed system it is possible to set up a service which manages, for a given document (e.g. id, key, url ) and a given user, the rights of the user to access and view a document.

This service will be called during the analysis of the search results to add, or not as the case may be, the relevant documents to the list of results returned.

This principle can, in particular, be reserved for cases where contents are not indexed by AMI.

The principle of Search is detailed in paragraph 3.2.

#### **5.1.5 Dynamic Information**

In the simplest cases AMI indexes documents that have a clear and defined physical existence in the form of file. At the end of a search process AMI returns the address of this file and a native protocol makes it possible to access it (e.g. url, http).

However it is often the case also where AMI will index contents which are created dynamically at the time of the indexing (database records, composite documents coming from various sources...). In this case, the index will not contain a URL but a single identifier (key) which will be returned at the end of a search and provided to the application which will recreate the contents of it if the user wishes to access it.

Example: Documents stored in a Document Management System are often in this form. AMI can fully integrate with a Content and Document system considerably enhancing the indexing and usability performance whilst at the same time providing a single point of search access.

### 5.1.6 External Sources

Given AMI's approach and all the possibilities it offers the fact that the sources are external to the Organisation changes nothing regarding the way in which authorisation is applied. The same techniques are used, the same protocols are respected and applied to the information retrieved.

#### Authors Rights and Royalties

Note: Information accessed or collected that is produced and published by third parties can be subject to copyright restrictions and/or laws that may be in force regarding copyright material, its use and distribution. These may vary from country to country and it is the user's responsibility to comply with any stated requirements or restrictions.

## 5.2 Search

If the ability to connect to specific sources of information is an essential precondition then Search is at the heart of the product and offers many different options.

Search is a combination of the creation of a search query with optional structured selection criteria acting on the indexed metadata.

**AMI Knowledge Discovery v5.0** endeavors to identify in the request if "expressions" are used. Thus, rather than just simply finding documents that contain a series of words, it prioritises those which really "speak about" the subjects that are specifically of interest to the user.

### 5.2.1 Search Creation or Request

#### Syntax

Search requests can be expressed in "everyday language" or using Boolean terminology.

Using "everyday language" the user expresses an idea or a subject of research using terms, phrases or even a longer descriptions that AMI, in turn, analyses. It is even possible to use an entire paragraph of text as the search term in order to find within the group (or groups) of documents searched all those relating closely to the required subject.

Example: Food industry and the World food Program

In Boolean terms AMI understands the instructions AND, OR and NOT as well as quotation marks, parenthesis and truncation. It adds to it an AMI "operator" (the question mark) which consists of asking the system to generate all variations of the same term starting from the contents of the Knowledgebase (section: 3.2.3). Example: ("food industry" AND fair trade?) NOT (packaging\* OR plastic\*)

#### Options

In addition to the language identified for each search request the AMI interface makes it possible to specify, if required, the generation of hypotheses (equivalent to the ? operator in Boolean language) and whether, for example, the requirement is to search "all the terms" of the request or "at least one of the terms" of the request.

## Language

AMI automatically recognises English, French, German, Spanish, Portuguese, Italian and Dutch as standard. AMI conforms to Unicode in its development and is not limited to Iso-Latin encoding hence further non-Western languages can be easily integrated.

Various algorithms provide automatic recognition the user's language. These algorithms are based on the morphology of languages as with the history of user requests. A score of confidence in terms of recognition is applied. Automatically recognising the language of the request makes it also possible to further generate hypotheses.

### 5.2.2 Restrictions

Data within sources is sometimes further characterised by additionally accessible information in the form of metadata.

**AMI Knowledge Discovery v5.0** can interact with this metadata, whether it is in an AMI index or a third party index, and thus allow the selection of documents answering only certain criteria. AMI's interrogation syntax conforms fully to the XCQL standard.

### 5.2.3 Request Parameters

In addition to answering the selection criteria (request, metadata) AMI further applies all its capabilities to the following parameters:

#### The Knowledgebase

The *Knowledgebase* contains the data memorised by the system during its use: Enrichment of the *Knowledgebase* is done gradually by a process of automatic learning. It can also be further enhanced by the insertion of vocabulary learned during the indexing processes and by manual insertion of terms by the user.

It's contents are made up of :

- Simple terms (words or groups of words)
- Relationships between terms: In addition to the traditional relationships of synonymy can also be added non-symmetrical and weighted, general or specific relationships (e.g. cat "is a type of" mammal).

The *Knowledgebase* can thus, in its own right, become a true and very accurately defined thesaurus, equally it can be connected to an existing pre-defined thesaurus if required. Because it does not specifically pre-require the use of an existing thesaurus **AMI Knowledge Discovery v5.0** allows continual use and re-use all the investment made in the creation and development of the *Knowledgebase*.

Often referred to as "generation of hypotheses" the purpose is to enrich a user request by referencing the existing content of the knowledgebase. Examples:

User Types	AMI generates the Request
FSA	FSA OR "Financial Services Authority"
Atlantic Ocean	Atlantic Ocean

The *Knowledgebase* can be questioned using either Boolean syntax or free text across the user selected sources and applications that AMI Knowledge Discovery v5.0. is integrated with.

For example, the interface will propose an option of activating, or not, the generation of hypotheses and will allow the sorting of results by date or by language.

The screenshot shows the search interface for AMI Enterprise Discovery v4.0. At the top, there are language selection buttons for 'Français' and 'English', and a breadcrumb trail 'Home > Search'. Below this is a search bar labeled 'Query' containing the text '"Royal Mail"'. Underneath the search bar is an 'Advanced Options' section. This section contains several settings: 'Search Mode' with radio buttons for 'Free Text' (selected) and 'Boolean'; 'With Hypothesis' with a checked checkbox; 'All terms' with a checked checkbox; 'Results per source/result set' with a dropdown menu set to '20'; 'Updated since' with a dropdown menu set to 'Two weeks'; and 'Language' with a dropdown menu set to 'English'. At the bottom of the 'Advanced Options' section is a large orange 'Search' button.

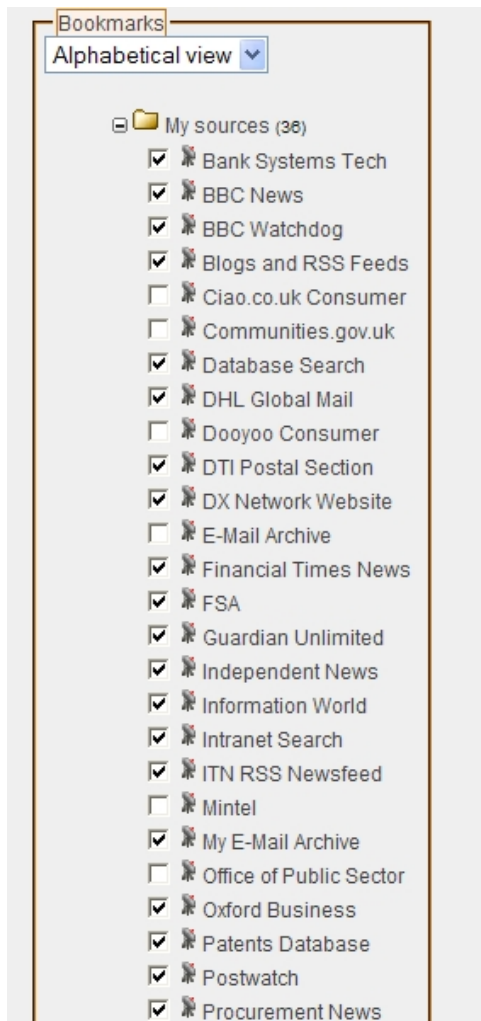
AMI Enterprise Discovery v4.0 allows the user to apply a full range of pre-search filters and conditions to the Search process:

- Free Text or Boolean
- With Hypothesis
- Number of Results per source
- Historic period over which to Search
- Language

### Activation of Sources

Once an information source is connected to the system it can then be set to be indexed daily or at a time frequency to suit user requirements.

It is natural to propose that **AMI Knowledge Discovery v5.0** should address all sources systematically however giving a user the ability to chose selected sources is far more useful, in terms of both the accuracy and comprehension of results returned, and engaging in terms of usability. In connecting to more than one source AMI functions as a genuine meta-engine.



Sources across which to perform a Search can be individually selected by the user and can be of any type, typical examples include;

External	(Public Websites, RSS feeds, Blogs)
Internal	(Intranet, E-Mail Archive, Database)
Professional :	(Subscription, Patent, Library services)
Specialist :	(Consumer Opinion, Legislation)

### Number of Results

The posting of results is controlled by the Retrieval, see below page 15. The maximum number of results required per source can be specified by the user at the time of request.

However It is not just a question of retrieving a fixed number of results but also the threshold of relevance applied which can influence the results that are, or equally, are not posted.

### Other Parameters

Generally, all AMI's configurable parameters can be applied for use within the application automatically or alternatively in a way that allows the user to choose each parameter individually such as the "weighting" applied to each source, choices of Knowledgebase or of other parameters such as a list of "stopwords".

#### 5.2.4 Search By Example

It was mentioned earlier that a real language request term can be a whole paragraph of text. In fact any text can be submitted to AMI which will then subsequently find the documents whose "direction" are semantically closest to that of the required subject text.

On short texts the entire document can be used.

On longer texts **AMI Knowledge Discovery v5.0** is able to extract from the document the most significant sentences; this summary of key sentences can then be submitted to the search engine to find corresponding documents.

This “Similar Documents” process is described in the paragraph entitled “Retrieval”.

The technology used to identify and locate sections of a text most representative of the subject that the document describes called G-Mil (**G**eneration of **M**arkers **I**ndependent of **L**anguage) is based on the extraction of syntax and is patented.

## 5.3 Retrieval

The process of retrieving results is similar to the process of search itself. However it also uses characteristics which are unique to **AMI Knowledge Discovery v5.0**.

### 5.3.1 Results Presentation

Results generated by **AMI Knowledge Discovery v5.0** appear in a single view representing the most accurate and relevant derived from cross-referencing all the results provided by the various sources. Each document is positioned according to its relevance independent of its origin.

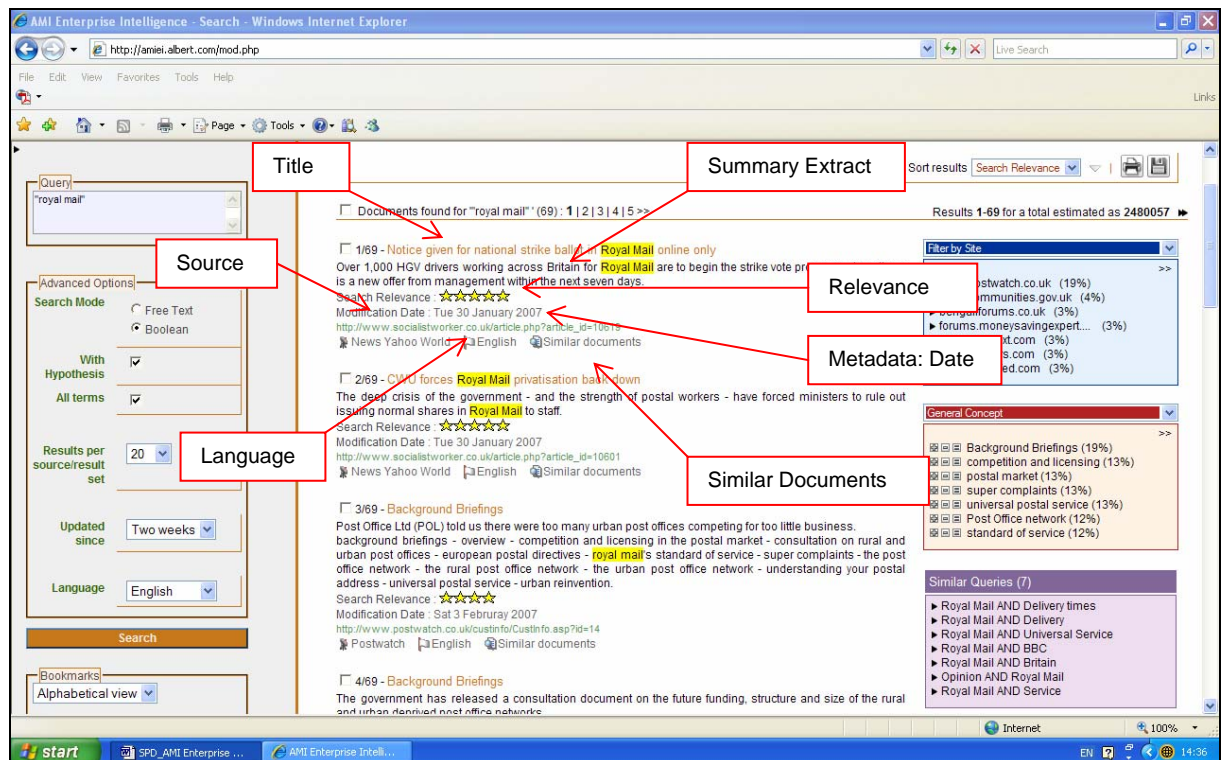
Of course, it is also entirely possible to separate the results emanating from each source respectively if required.

Results are generated in a native XML flow enhanced by information regrouped in two sets: 1) answers to the search query and 2) related information.

The results to the search query consist of an extract of the relevant document such as its title and any associated metadata acquired during indexing. Moreover **AMI Knowledge Discovery v5.0** returns the unique location id of the document which will be used for complete retrieval.

In the simplest cases this identifier is a URL which allows the opening of the file within the browser (HTML page, Office file). In other cases, the key created by AMI will be used to read information in a database and to format it before presenting it to the user.

Under default settings AMI's user interface presents results in the following form:



Any other presentation configuration can be considered fully utilising acquired information contained within the XML flow returned by AMI.

### 5.3.2 Relevance

Relevance is a key central element of **AMI Knowledge Discovery v5.0**. On sources of information capable of containing hundreds of thousands of documents the list of all those simply containing a word or an expression can be very long and completely unusable. This is why AMI prioritises the results that are both necessary and sufficient to meet the search criteria.

Documents are classified, by default, in order of relevance decreasing.

The calculation of relevance (post ranking) carried out by **AMI Knowledge Discovery v5.0** is based on the frequency of appearance of the terms of the request within the documents considered. It also takes into account the position of the term within the document and the relevance of the words within the corpus of documents considered when this information is available.

The level of relevance applied by **AMI Knowledge Discovery v5.0** extends on a scale from 1 to 100. It is thus possible to compare the relevance of a response to one request with another and subsequently set up a threshold of relevance.

## Source Relevance

The level of relevance of documents retrieved can depend on the source or “confidence” level applied to it. A specific weighting parameter can be allocated to each source which is used by the various algorithms to form relevance calculations.

## Other possibilities for Filtering and Sorting

It is possible to sort the list of results on any metadata present in the set of answers in either an increasing or decreasing way.

## Preferential Pages

**AMI Knowledge Discovery v5.0** offers a function called preferential pages which allows the administrator to:

- Associate a document with keywords which the document does not necessarily contain.
- Give to all documents of a specified type a higher weighing than those of the rest of the group

Thus ensuring that a prescribed answer to a specific query will be presented independent of the relevance computing process.

### 5.3.3 Correlated Information

The flow of results also contains additional information, automatically acquired and calculated by **AMI Knowledge Discovery v5.0**, allowing unique and powerful functions to be used as required.

This information is as follows:

Origin of each document: on a multi-index search or on an index which points to several sources, the documents are characterised by origin (e.g. field name in the case of url's, http)

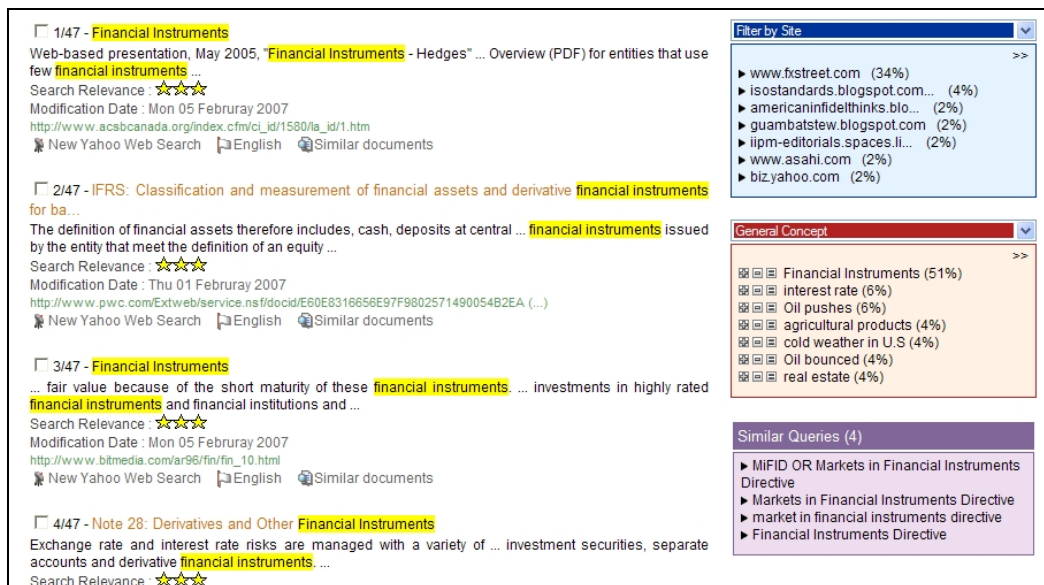
Extracted Nominated Entities:

AMI will automatically identify Locations, People and Organisations contained within the results and also further identifies “general concepts” capable of being used for clusterisation of the results (see below, section 3.3.4).

Similar Requests:

**AMI Knowledge Discovery v5.0** will automatically detect and present similar or “peer group” searches already conducted that are similar or closely relate to the user query

**AMI Knowledge Discovery v5.0's** default interface presents this information to the user, enabling them to reformulate or re-specify their request by simply clicking on the terms presented. For example:



1/47 - **Financial Instruments**  
 Web-based presentation, May 2005, "Financial Instruments - Hedges" ... Overview (PDF) for entities that use few financial instruments ...  
 Search Relevance: ★★★★★  
 Modification Date: Mon 05 Februray 2007  
[http://www.acsbcanada.org/index.cfm/ci\\_id/1580/la\\_id/1.htm](http://www.acsbcanada.org/index.cfm/ci_id/1580/la_id/1.htm)  
 New Yahoo Web Search English Similar documents

2/47 - IFRS: Classification and measurement of financial assets and derivative financial instruments for ba...  
 The definition of financial assets therefore includes, cash, deposits at central ... financial instruments issued by the entity that meet the definition of an equity ...  
 Search Relevance: ★★★★★  
 Modification Date: Thu 01 Februray 2007  
<http://www.pwc.com/Extweb/service.nsf/docid/E60E8316656E97F9802571490054B2EA> (...)  
 New Yahoo Web Search English Similar documents

3/47 - **Financial Instruments**  
 ... fair value because of the short maturity of these financial instruments. ... investments in highly rated financial instruments and financial institutions and ...  
 Search Relevance: ★★★★★  
 Modification Date: Mon 05 Februray 2007  
[http://www.bitmedia.com/var96/fin/fin\\_10.html](http://www.bitmedia.com/var96/fin/fin_10.html)  
 New Yahoo Web Search English Similar documents

4/47 - Note 28: Derivatives and Other Financial Instruments  
 Exchange rate and interest rate risks are managed with a variety of ... investment securities, separate accounts and derivative financial instruments. ...  
 Search Relevance: ★★★★★

**Filter by Site**

- www.fxstreet.com (34%)
- isostandards.blogspot.com... (4%)
- americaninfidelthinks.blog... (2%)
- guambatstew.blogspot.com (2%)
- iipm-editorials.spaces.li... (2%)
- www.asahi.com (2%)
- biz.yahoo.com (2%)

**General Concept**

- Financial Instruments (51%)
- interest rate (6%)
- Oil pushes (6%)
- agricultural products (4%)
- cold weather in U.S (4%)
- Oil bounced (4%)
- real estate (4%)

**Similar Queries (4)**

- MIFID OR Markets in Financial Instruments Directive
- Markets in Financial Instruments Directive
- market in financial instruments directive
- Financial Instruments Directive

The option "similar documents" shown above is an illustration of "Search by Example" detailed in paragraph 3.2.4. When the user chooses this option, AMI:

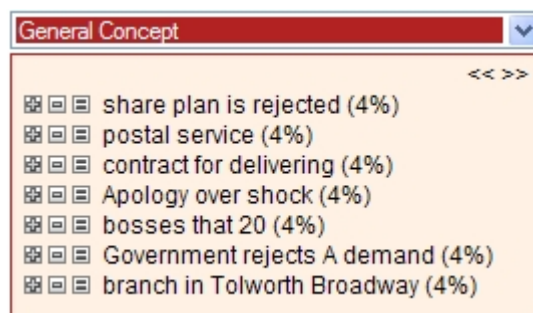
- reads the entire contents of documents on the source.
- calculates the most representative sentences.
- submits to the same sources a request built using these representative sentences (i.e. the most important noun phrases), and applies the same rules.
- provides a new set of answers of which the most relevant is, in fact, the document used as the example.

Like all functions requiring a particular user interface this one can be either **AMI Knowledge Discovery v5.0's** default setting as shown or customised.

### 5.3.4 Categorisation and Clusterisation

Documents can belong to preset categories memorised in the indexed metadata. AMI can convey these values to the user interface which can then present the documents in a tree structure corresponding to a classification plan.

Categorisation of the results, also called clusterisation, is also possible via the selection of the "general concepts" automatically generated from the set of results.



**General Concept**

- share plan is rejected (4%)
- postal service (4%)
- contract for delivering (4%)
- Apology over shock (4%)
- bosses that 20 (4%)
- Government rejects A demand (4%)
- branch in Tolworth Broadway (4%)

### 5.3.5 Retrieval of Document Information

The set of results contains an address or an identifier allowing the retrieval of the document itself. The application implementing **AMI Knowledge Discovery v5.0** offers the user the means of choosing the documents to be consulted and the ability to open them within the browser or a dedicated application.

AMI's native user interface gives access to documents via http protocol.

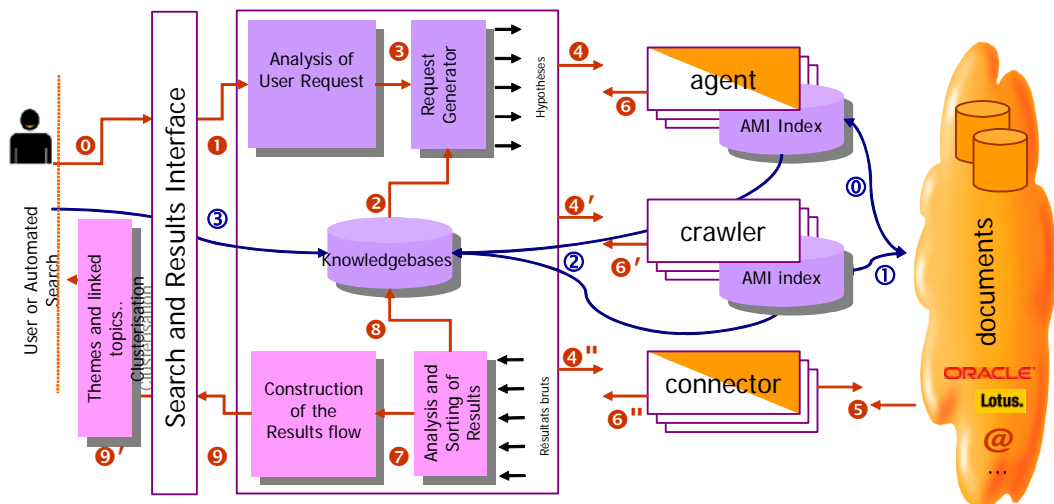
#### Downloading Document content

**AMI Knowledge Discovery v5.0** presents documents where they are, at the source of origin. Users of the AMI module AMI Publish, part of the Enterprise Intelligence suite, can also download the contents to store in an AMI managed classification plan.

## 6 Architecture

### 6.1 Overview

The diagram below shows the principal stages of the respective search and indexing operations.



## 1. Indexation and Knowledgebase Management (Dataflow shown in Blue)

- 0) Indexing of a source by an agent (specific program);
- 1) Indexing of a source by the crawler (AMI indexer);
- 2) Automatic population of the Knowledgebase by vocabulary contained within the index.
- 3) Population of the Knowledgebase by the system user (e.g. synonyms, importation of dictionaries...).

## 2. Management of Requests (Dataflow shown in Red)

- 0) Submission of the user request:
- 1) Analysis of a request by AMI;
- 2) Reading and cross-referencing of the Knowledgebase content to enrich the request;
- 3) Creation of a request in the language of each source (AMI or Non-AMI)
- 4) Parallel submission of requests;
- 5) Interrogation of a third party engine's and retrieval of the results if necessary;
- 6) Selection of the results by individual respective sources;
- 7) Cross-referencing of the results and calculation of their relevance;
- 8) Training: insertion of new vocabulary into the Knowledgebase resulting from the "flow" of results;
- 9) supply of the result to the end-user with detection of linked topics, close/similar requests and clusterisation of the results.

## 6.2 Services

**AMI Knowledge Discovery v5.0** provides, in http form, the following services:

<u>Federated Search</u>	Submission of a request to a meta-engine.
<u>AMI Request</u>	Submission of a request to the AMI engine.
<u>Synchronisation</u>	Modification of an index at the time of insertion or removal or modification of a document within a corpus of documents.
<u>Syntax Extraction</u>	Analysis of a document and the extraction of its principal noun phrases. This allows the "highlighting" of the essential elements of a text, Similar Search, or the automatic creation of summaries.
<u>Learning</u>	This service is used to memorise terms in the Knowledgebase.

Similar Requests This service highlights previous similar search terms. The history of the requests can be generated automatically or be initialised against the contents of an unspecified file if required.

These services all are all available and utilised via the standard **AMI Knowledge Discovery v5.0**. interface. They can also be used separately in any application built or “made-to-order” around the engine.

### 6.3 Compatibility

**AMI Knowledge Discovery v5.0** provides full compatibility with preceding versions of the product. In particular, the indexes and knowledgebase’s can be preserved at the time of an upgrade as well as the configuration of applications using the generic connector protocol to access external sources of information.

Upgrades of earlier product versions can be planned to fully take advantage of all new functionality.

### 6.4 Supported Environments

Full details regarding AMI’s technical requirements and supported environments are given in the Document “AMI\_Technical\_Requirements” which is available via the download section of our website or on request.

## 7 Glossary

This chapter provides a summary definition of the principal components of the product **AMI Knowledge Discovery v5.0**.

### Knowledgebase

Internal AMI Database containing vocabulary and synonyms and able to import Thesaurus data. Used to enrich search requests automatically and to better describe the search subject. The Knowledgebase is fed both by the user and automatically by dedicated learning algorithms.

### Connector / Agent

Program connecting to a source of information either to question it in real time (connector) or to extract specific elements from the documents which are indexed by AMI (agent).

### **Clusterisation**

Action consisting of gathering, “on-the-fly”, answers to a request from within heterogeneous subsets relating to the same topic.

### **Crawler**

Active element of the indexer who’s action indexes http sources and filing systems. The crawler can follow html links.

### **Site Descriptor**

Set of directives associated with a source allowing the indexer to optimise the process of accessing only certain parts of the source and to index only selected elements.

### **Expression**

See “Term” below.

### **Filter**

Program which reads the contents of a document to extract the “to be taken into account” information (included text, metadata) or to carry out a specific process on certain documents. New Filters can be added to the existing Filters.

### **Generation of Hypothesis**

Action consisting of enriching a user request by referencing closely matched vocabulary and correlated information contained within the Knowledgebase.

### **Indexer**

Program creating an Index. Sources of information that do not have a native mechanism of interrogation must be indexed before searching can be performed.

### **Boolean Language**

Structured interrogation language allowing the implementation of Boolean algebra based on the industry standard AND, OR and NOT terms

### **Real Language**

Use of natural language to form search queries without the use of specific operators or particular syntax. AMI then seeks the documents containing “as much as possible” information relating to the terms of the request.

### **Preferential Pages**

Documents presented first against certain pre-defined search requests independent of their relevance.

### **Post Ranking**

Action of cross-referencing the results coming from various sources and of their sorting by decreasing relevance.

## Request

Commonly referred to as “Search”, the Request represents the terms sought in the searched for documents. Requests can be in either Boolean or real language.

## Signature

Representation calculated by AMI used to determine the essential textual contents of the document, i.e. that about which it “really speaks”. Stored in an index, the signature is used in the calculation of the relevance of a document compared to a request.

## Synchroniser

Service suggested by the indexer to update the contents of an index at the time of insertion, removal or modification of a document within a corpus.

## Term, expression

The simplest search engines index documents based on the words they contain but can neither calculate their importance nor the meaning they can have especially when cross-referenced. AMI, of course, works on words as well but more importantly on the meaning and unique significance that groups of words have. In this document, we use the word “term” indifferently to indicate the simple words or groups of words that have meaning.

## 8 Documentation and Contacts

**AMI Knowledge Discovery v5.0** is delivered with full documentation in English or French available either as a PDF file or via HTML . The documentation consists of;

- Reference Guide (320 pages)
- Programming Guide (50 pages)

This documentation provides the Administrator of **AMI Knowledge Discovery v5.0** with the information required to install the product and set up administration routines and initial connection to news sources. The documentation describes all parameters and capabilities of the software from an operational as well as technical point of view.

For any further information required relating to **AMI Knowledge Discovery v5.0** please contact us directly;

Contact details can be found via our website [www.amisw.com](http://www.amisw.com)

.oOo.