Development of the Information and Analysis Centre (IAC) in the Verity K2 Environment

Alessandra Ussorio

Institute for the Protection and Security of the Citizen
Ispra, January 31 2007
January 31, 2007

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

European Commission
Joint Research Centre

Contact information
Giacomo Cojazzi
Address: Via Fermi, 1 (T.P. 210), 21020, Ispra (VA), Italy
E-mail: giacomo.cojazzi@jrc.it
Tel.: +39 0332 785085
Fax: +39 0332 787748

http://www.jrc.ec.europa.eu

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

A great deal of additional information on the European Union is available on the Internet.

It can be accessed through the Europa server
http://europa.eu/

JRC PUBSY REQUEST 7963
JRC 37963

EUR 22932 EN
ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

© European Communities, 2007
Reproduction is authorised provided the source is acknowledged

Printed in Italy
Table of Contents

1. Introduction ................................................................................................................... 1

2. Verity K2 ...................................................................................................................... 3
   2.1 K2 Software Architecture ..................................................................................... 3
       2.1.1 Master Administration Server ....................................................................... 3
       2.1.2 K2 Broker ......................................................................................................... 4
       2.1.3 K2 Servers ......................................................................................................... 4
       2.1.4 K2 Ticket Server ............................................................................................... 4
       2.1.5 K2 Spider ........................................................................................................... 5
       2.1.6 Collection .......................................................................................................... 5
   2.2 Knowledge Trees ..................................................................................................... 6
   2.3 Topic trees ............................................................................................................... 7
   2.4 Thesaurus ................................................................................................................ 9
   2.5 Score calculation .................................................................................................... 10
   2.6 Operators ............................................................................................................... 10
   2.7 Verity K2 manuals ................................................................................................. 11

3. Customisation of Verity K2 to the IAC ...................................................................... 13
   3.1 Requirements Specification ................................................................................... 14
   3.2 Technical description of the developments for customisation .............................. 14
       3.2.1 Collection of documents .................................................................................. 14
       3.2.2 Multilanguage .................................................................................................. 15
       3.2.3 Security ............................................................................................................ 16
       3.2.4 Collection-level security ................................................................................. 17
       3.2.5 Document-level security ................................................................................. 17
       3.2.6 Duplicates detection ........................................................................................ 18
       3.2.7 Knowledge trees .............................................................................................. 18
       3.2.8 Web interface .................................................................................................. 19
   3.3 ENRICHMENT Topic tree .................................................................................... 19

4. Verity K2-IAC user manual ........................................................................................ 23
   4.1 Procedure to upload documents ............................................................................. 23
   4.2 To open a session ................................................................................................... 23
   4.3 Searching documents ............................................................................................ 28
       4.3.1 Simple Searches .............................................................................................. 28
       4.3.2 Advanced Searches ......................................................................................... 31
       4.3.3 Searching documents through Knowledge Trees ............................................ 32
   4.4 Use of topic trees for searches ............................................................................... 34

5. Conclusions and Further Developments ..................................................................... 39

Acknowledgments ............................................................................................................. 40

References ......................................................................................................................... 41
List of Figures

Figure 2.1  Verity K2 Architecture (From Ref. [6]). 3
Figure 2.2  Collection schema (From Ref. [12]). 5
Figure 2.3  An example of Knowledge Tree. 7
Figure 2.4  Two Knowledge Trees for the same set of documents. 8
Figure 2.5  Topic tree schema (Adapted from Ref. [2]). 9
Figure 2.6  Enrichment Topic tree schema. 10
Figure 2.7  Query Expression. 11
Figure 3.1  Data flow for Collection of Documents. 17
Figure 3.2  Tree-like structure of the ENRICHMENT topic tree. 22
Figure 3.3  Partial expansion of the ENRICHMENT topic tree. 23
Figure 4.1  IAC Home page. 26
Figure 4.2  SEARCH page. 27
Figure 4.3  BROWSE page. 28
Figure 4.4  Login link in Search page. 29
Figure 4.5  Authentication window. 30
Figure 4.6  SEARCH page description-1. 31
Figure 4.7  SEARCH page description-2. 32
Figure 4.8  Advanced Search fields. 34
Figure 4.9  Search Results list. 34
Figure 4.10  Browse page - Knowledge trees list. 35
Figure 4.11  Browse page – Query in the Iran category on the “nuclear” term. 36
Figure 4.12  The use of Topic Trees for searches. 37
Figure 4.13 An example of use of Enrichment Topic Tree for searches. 38

Figure 4.14 An example of use of Enrichment Knowledge Tree for searches. 39
Document History and Approval

Document Revision
The text has been revised and cleared by G. Cojazzi
Action Leader of JRC action AMENUS
(Signed)
1. Introduction

The IPSC Institute of DG JRC started to organise an Information and Analysis Centre (IAC) with the purpose of collecting information concerning Nuclear, Chemical and Biological Non Proliferation, Disarmament and Weapons of Mass Destruction from the Open sources, analysing and validating them and then producing studies on related subjects [1].

The data base consists of Political/commercial/social information such as treaties, agreements, conventions, regulations etc... because in direct relation with the scientific/technical problems of Non Proliferation, Disarmament, and so on. So, all the needed information could be found in a unique location.

At the beginning documents were collected in a simple file system structure, but as their number was increasing, data retrieval was becoming very difficult. So the need to have a documentation management system grew up.

After an investigation of the commercial documentation management systems, Verity Portal One Application Suite © developed by Verity Incorporated was selected for compatibility with tools in use at IAEA [2].

This system has been customized according to the project needs [3].

Then a prototype Geographic Information System (GIS) interface, complementary to Verity interface, was developed to provide a graphical interface to further simplify the search of documents about nuclear sites [4]. This new interface uses the attributes stored in GIS as search keys, sending them to Verity in an appropriate format; these keywords eventually trigger predefined queries stored into Verity, called Topic Trees [5]. Verity handles the search and sends the Search Results to the GIS interface. The Verity system with GIS interface represented a first trial of interrogation of a huge number of documents through a map.

In 2005 a new version of the Verity system (Verity K2 ©) was released, with new relevant functionalities implemented, so a migration of IAC data to the new Verity K2 system was necessary.

The objective of this report is to provide a guide for users, a support for administrators and to point out the potentiality of the System for further developments.

This report contains a description of the work carried out for IAC management through migration to Verity K2 system [6-17]. In particular Section 2 gives a brief description of the Verity K2 system, Section 3 is dedicated to the project and user requirements, together with how the system was customized according to those needs. Section 4 is dedicated to the VERITY-IAC user manual and Section 5 contains some guidelines for the administrator. Conclusions and further developments are reported in Section 6.
2. Verity K2

The Verity K2 is a set of software to index, organise, search, retrieve and view documents (fig. 2.1).

2.1 K2 Software Architecture

The most general schema of K2 Software architecture, which could be implemented, is shown in fig. 2.1.

![Verity K2 Architecture](image)

Figure 2.1 Verity K2 Architecture (From Ref. [6])

2.1.1 Master Administration Server

The Master Administration Server is a centralized configuration repository that allows configuring and administering a distributed K2 system. It tracks and distributes information on services and indexes to all other Administration Servers. Each Administration Server is responsible for starting up and administering services such as K2 Broker, K2 Ticket Server, K2 Server and K2 Spider Controller that runs on a host.

The Master Administration Server:

- Centralizes and distributes system configuration information
• Starts up and shuts down K2 services
• Monitors the health of K2 services and provides alerts and restarting of services
• Manages K2 Spider indexing tasks, indexes and Profile Nets

2.1.2 K2 Broker

The Verity K2 Broker manages communications between clients and the K2 Servers. It receives client requests and sends them to available K2 Servers. A K2 Broker is needed for a distributed K2 domain. Each Verity K2 Server performs its appropriate subset of task against its indexes, and returns the results to the K2 Broker. These operations are simultaneous, whether or not they are on the same host.

2.1.3 K2 Servers

K2 Servers can run on a single host or on multiple hosts. A Verity K2 Server contains the following components [6]:

• Search Engine Service — Searches a specific set of documents indexed into Verity collections
• Viewing Service — Allows to view documents returned from searches and highlights the search terms.
• Profile Service — Classifies documents based on a set of rules
• Knowledge Tree Service — Performs navigation and searching within data
• Parametric Search Service — Performs parameter-based selections on data, using an index that is optimised for those kinds of selections
• Recommendation Engine — Uses the behaviour of users who have performed similar searches to learn from their feedback on the quality of the results to affect future rankings.

2.1.4 K2 Ticket Server

The K2 Ticket Server authenticates users to the K2 domain and allows them to search secure indexes. It consists in security modules that communicate to third-party information stores to validate logging in the K2 domain. The third-party user/information group stores that are supported include LDAP Servers, Windows Domain, and UNIX.
2.1.5 K2 Spider

To search or classify information stored in repositories, it must first be indexed. Verity K2 uses the K2 Spider and various gateways to access information and to build and update a universal index, called a collection.

2.1.6 Collection

The core element of the system is the Collection (fig. 2.2).

![Collection schema](From Ref. [12])

A Collection represents “metadata” (that means data on the data) for a set of documents and includes word indexes, an internal documents table and pointers to the original locations of the original documents.

For example a user uploads a document to the repository, the document is indexed into the collection that means that:

1. Word indexes are created
2. In the documents table, for that document are filled in custom and predefined fields with values for the document such as the title, the date and so on. This allows the user to perform searches on one or more of these fields.
3. Pointers to the original location of the original document are created.

A collection supports documents in any format, and its architecture is optimised for searching.

Some of the features supported by the collection architecture are:
- Collections can be updated continuously without affecting current searches.
- Document indexing can occur continuously.
- General housekeeping functions are performed to ensure efficient search performances.

### 2.2 Knowledge Trees

A **Knowledge Tree** is a hierarchical set of categories, which organises the documents indexed into a Verity collection so that users can enhance information access by integrating searching and browsing through that structure.

As can be seen in the figure 2.3, a Knowledge Tree allows the user to see documents organised in a structure similar to that of the file system (folder, subfolder, and so on…), but it is a logic and not physic organisation.

![Knowledge Tree Diagram](image)

**Figure 2.3** An example of Knowledge Tree.

In other words it can create different views for a set of documents, defined by the administrator on user request.
Categories, through category capture methods, organise documents into subjects of interest. This combined with the hierarchical structure of the Knowledge Tree, similar to a file system organisation, helps the user to find documents of interest.

More than one Knowledge Tree can be associated with a collection to provide different views of the same documents set and a document can be in more than one category.

The figure 2.4 shows an example of two different Knowledge Trees for the same set of documents. In the leftmost tree documents are organised for country, while in the rightmost tree they are organised by document type (e.g. Treaties, Agreements and so on).

![Figure 2.4 Two Knowledge Trees for the same set of documents.](image)

There are several techniques to define a Knowledge Tree. It can be built defining rules for document assignment to each category, reflecting the file system physical structure or reflecting the structure of a Topic Tree (see Section 2.3). When creating the Knowledge Tree reflecting the Topic Tree structure, documents in the collection are organised according to the Topic Tree categories (the sub-topics), giving a static view of the Topic.

This doesn’t produce duplicates because we have only links to the original documents and not the documents themselves.

2.3 Topic trees

Verity System uses a technology that allows users to search for “concepts” in documents, rather than individual words or phrases. Verity System treats specific words and phrases as evidence of the presence of a concept in a document, that is, it allows giving a context to words.

Search terms are encapsulated in an entity called “Topic Tree”. Each term is exploded into sub-concept and so on, leading to the construction of a tree-like structure (see fig.2.5):
Importance weights can be assigned to sub-concepts to reflect the fact that some words or other concepts are more important than others in expressing the specific concept.

A topic tree is a grouping of information related to a concept or a certain area of interest. In other words, a topic tree is a predefined query that supports very efficient searches.

The definition of the topic tree is done with usual words. Developing a topic tree consists of modeling the description with main words (those that are important for the description of the process), linked together by simple rules of proximity, exclusion, etc.

For example the sentence “enrichment of uranium” will be modeled in the following way: (enrichment)<near/5>(uranium). This means that will be searched all the occurrences of the word “enrichment” and all the occurrences of the word “uranium” at a distance of five words from each other, at most.

Once the “Topic Tree” has been built and inserted into the system by the administrator, each user can make searches with the help of it. For each document in the data base, the software will detect if the modeled sentence is present. If so, the document will be selected and ranked according to a “ranking algorithm”. So ideally a user can perform a search on certain technical process without being an expert, as the System will use the terms that describe the process in the specific context.

The Enrichment Topic Tree is based on a IAEA’s type Physical Model relating to enrichment [2].

There are a lot of similarities in structure of the TOPIC search tree and the technical process for Enrichment (see fig. 2.6).
Building a Topic Tree some points have to be considered:

- Topic tree is based on the detection of given key words. Often open-source documents use more colloquial terms than the ones used by a nuclear expert. TOPIC systems were enabled to handle synonyms easily, so once lists of synonyms have been defined in a Thesaurus, they can be used in the Topic tree through the <THESAURUS> operator. There is one standard thesaurus that is part of the K2 system. The users can also define their own thesaurus that will be added to the predefined thesaurus.

- Some key terms are not sufficiently specific to the subject of the search. So they cannot be used by themselves because they would retrieve too many irrelevant documents. To ensure that key words are being used in a relevant context, most of them must be related by a “proximity condition” that means that if the key word is found in a document, the document is retrieved only if another identified key word is found within a specified distance of it (e.g. a distance of five words or less, within the same paragraph, etc).

- Weights could be assigned to key words to reflect how much a certain term is more relevant than others (This also reflects the weights of the indicators of the technical process).

2.4 Thesaurus

A thesaurus control file contains synonym list definitions. The thesaurus is accessed during search processing when user queries include the <THESAURUS> operator.

Verity K2 contains a predefined thesaurus for the English language; it can be substituted or integrated with a custom thesaurus.
2.5 Score calculation

A query expression is a statement you enter as criteria for performing a search. It consists of words and operators see figure 2.7.

![Figure 2.7 Query Expression.](image)

When the default parser (the simple query parser) is used, a query expression can be typed using simple or explicit syntax. Simple syntax consists in words or phrases in the natural language, for example “nuclear facilities”. Explicit syntax means words and logical operators and modifiers are typed in the search field, for example “nuclear AND facilities”. The syntax used determines whether the search words entered will be stemmed (that means that also the variations of the word specified are searched), and whether the words that are found will contribute to relevance-ranked scoring.

When documents are relevance-ranked, they are listed in an order based on their relevance to the search criteria. Relevance-ranked results are presented with the most relevant documents at the top of the list.

Verity K2 uses an internal proprietary algorithm to calculate the score; the use of weights, operators and modifiers can affect the way K2 engine assign the score to documents.

When simple syntax is used to specify a query expression, the search engine implicitly interprets single words entered as if they were preceded by the MANY modifier and the STEM operator. By implicitly applying the MANY modifier, the search engine calculates each document’s score based on the word density it finds; the denser the occurrence of a word in a document, the higher the document’s score.

As a result, the search engine relevance-ranks documents according to word density as it searches for the word specified, as well as words that have the same stem. For example, “processes” and “processing” are stemmed variations of the word “process”. To search for documents containing the word “process” and its stem words, simply enter the word “process” (using simple syntax) [13].

2.6 Operators

Verity query language operators include [13]:

- **Evidence operators**: ALL, AND, ANY, OR, ACCRUE.
• **Proximity operators**: IN, NEAR, NEAR/n, PARAGRAPH, PHRASE, SENTENCE.

• **Concept operators**: SOUNDEX, STEM, THESAURUS, TYPO/N, WILDCARD, WORD.

For a complete list of basic and advanced Verity query language operators and their description refer to Ref. [13].

2.7 **Verity K2 manuals**

Documentation provided for Verity K2:

• **Verity K2 Installation and Setup Guide V5.5 [6]**:

  Contains general information on the architecture of the Verity K2 and instruction for installing Verity K2.

• **Verity K2 Getting Started Guide V5.5 [7]**:

  Defines the basic concepts of Verity K2 technology.

• **Verity K2 Dashboard Administrator Guide V5.5 [8]**:

  Describes how to manage the Verity K2 trough the K2 Dashboard.

• **Verity K2 readmin Guide V5.5 [9]**:

  This guide is for readers who want to administer their K2 system using a command-line tool, it is an alternative to the K2 Dashboard.

• **Verity K2 Recommendation Engine Guide V5.5 [10]**:

  This documentation is for K2 administrator who wants incorporate the Recommendation Engine in K2.

• **Verity Command-Line Indexing Reference V5.5 [11]**:

  This guide is for administrators who want to create and maintain Verity collections with the command-line indexers.

• **Verity Collection Reference V5.5 [12]**:

  The *Verity K2 Collection Reference Guide* describes the architecture and design of Verity collections. This book is for Verity administrators who have a Verity installation and who are familiar with the basic concepts of search applications.

• **Verity Query Language and Topic Guide V5.5 [13]**:
Shows how to construct queries with Verity query language, and how the four parsers included with Verity products parse those queries. Furthermore it defines Topic concepts.

- **Verity Intelligent Classification Guide V5.5** [14]:

  Describes Intelligent Classifier, a tool for creating, viewing, editing, and testing Verity topics, taxonomies and Knowledge Trees.

- **Verity Locale Configuration Guide V5.5** [15]:

  This guide is for administrators and developers of Verity K2 applications who need to know how to administer or develop an application that supports indexing and search in multiple languages.

- **Verity K2 ODBC Gateway Guide V5.0.1** [16]:

  This guide is for administrators who are responsible for indexing data from ODBC-compliant servers into Verity collections.

- **Verity K2 Developer documentation V5.5** [17]:

  This documentation is intended for developers who want to understand the relationship between Verity application programming interfaces (APIs) and use them to develop K2-enabled applications.

Online documentation is provided in HTML and PDF formats.
3. Customisation of Verity K2 to the IAC

The Information and Analysis Centre is of great importance because it has as primary resource the scientific and technical knowledge. So it can provide a unique service, supporting the legislative and political activities.

The IPSC Institute of JRC Ispra, which has longstanding experience in the field safeguards, has organised the Information and Analysis Centre in the field of safeguards, non proliferation etc.

The main tasks of the Information Centre were identified as follows [1]:

- Establishment of a preliminary set of information concerning Treaties, Agreements, Regulations and the various GOs and NGOs organisations.
- Establishment of a Data Base continuously updated with analysed and validated data.
- The Monitoring of the evolution in the field of NP, Verification Regimes and Disarmament.
- The construction of a Web site to provide information to Policy decision makers and the public in general.

The IAC project started in 2000 and a first implementation of the system ended in 2003 [3]. From then on, new needs and requirements have been identified.

The Verity K2 System has been chosen as documentation management system. Beyond the functionalities already present in the previous version (flexible automatic classification, advanced indexing and search techniques, documents retrieval and viewing), it implemented new very useful functionalities, allowing the satisfaction of the new specifications.

Potential Users of the System were identified as [1]:

- European Organizations (e.g. EC, Council and related services, e.g. DG-TREN, DG-RELEX)
- Governmental and other Political Bodies
- National and International Organizations (e.g. IAEA, …)
- Public (e.g. citizens, journalists, …)
- JRC (Non Proliferation scientific community)

The system offers to the users different way of access:

- Web browser
- NUMAS
Users of the first type access the system through a web interface (see Section 4), while the second types of users, i.e. the NUMAS users, access the documents repository through the GIS interface [4].

3.1 Requirements Specification

• **Specification of the Collection of documents:**

  Users who have permission to upload files into the system should have an easy procedure to do it. When searching for a document on the web, they should have the possibility to save the document directly to a common repository, and not necessarily passing through an interface.

• **Multilanguage:**

  The number of documents written in different languages is growing, so there is the need to handle each of them in a specific way. Textual information is highly related to the language, each word, sentence or paragraph represent a block of knowledge only within the context of a specific language. An information-handling system with Multilanguage support to make indexing and searches the more efficient possible is necessary.

• **Security:**

  There is a variety of type of documents and potential users of the system. Documents are treaties, agreements, technical/scientific documents, reports, news etc. some of them with a certain level of security and sensitivity. Users are mainly experts but may also be students, politicians and so on. In this context the implementation of a system with different level of security is of primary importance.

• **Duplicates detection:**

  The growing number of documents and the fact that many experts can upload documents in the system makes necessary the implementation of a functionality for duplicates detection.

3.2 Technical description of the developments for customisation

The current implementation of the IAC-K2 system has the Master Administration Server, one Administration Server, one Broker and one K2 Server installed and configured on the same server. Other Administration Servers, Brokers and K2 Server can be added, installing and configuring them later on different hosts, if needed.

3.2.1 Collection of documents

To allow an easy upload of files to the system, a folder structure, shared through the internal network, with specific permissions for each user, has been created on the
server which hosts Verity K2. Users can directly upload or modify files in the folders they have access to. The system is in charge to update Verity collections to synchronise them with the files added or modified (see fig.3.1).

To do that, indexing jobs (for K2 Spider) have been created and scheduled to run periodically. The interval of execution can be modified accordingly to specific needs. The administrator can easily create and administer these and other jobs through K2 Dashboard. The system native security model is integrated and enforced by the K2 security system, so the permission schema is respected in the collections.

Figure 3.1 Data flow for Collection of Documents.

3.2.2 Multilanguage

Verity locales are code modules and data tables that allow documents to be indexed and searched in a language-specific manner. Locales are at the core of Verity’s support for internationalization. By installing and configuring locales, an administrator can give a language-aware client application the ability to work in languages other than English.

A Verity collection can be defined to use only one of the single-language locales or the Multilanguage locale, which provides support for more than one language.

In Verity the following single-language locales are supported:

Bulgarian, Czech, Hungarian, Polish, Russian, Greek, Turkish, Bokmal, Danish, Dutch, English, Finnish, French, German, Italian, Nynorsk, Portuguese, Spanish, Swedish.
Verity also implements a **Multilanguage locale** (the *uni* locale), which by itself provides specific linguistic support for the following languages:

Bokmal, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Italian, Nynorsk, Romanian, Russian, Polish, Portuguese, Spanish, Swedish, Turkish, Bulgarian, Hebrew.

Using the Multilanguage locale the system automatically detects the language of each document and manages it correctly.

The Multilanguage locale has been configured to adapt it to the project needs [15] (as explained in Section 5).

The collections defined for the project, which contains documents in different languages, use the Multilanguage locale.

Verity provides some locales with a predefined Thesaurus, but not the Multilanguage (uni) locale. A customised Thesaurus has been implemented specifically for it, starting from the English thesaurus, adding synonyms for a set of terms in the field of Non Proliferation.

### 3.2.3 Security

Verity K2 provides a component, the K2 Ticket Server, to authenticate users to the K2 domain and to allow them to search secure indexes. It consists in security modules that communicate to third-party information stores to validate logging in the K2 domain. For this project, the third-party user/information group store is the Windows Domain.

Each time a user provides valid login information to authenticate to a security model, Verity K2 Ticket Server gives him a ticket. This ticket is stored by K2 Ticket Server until the user logs off; at that time the ticket expires.

Each time a user makes a request (search and viewing), the K2 Ticket Server gives that user access to only those collection for which they have tickets. This integrates the system native security model into the K2 system.

The K2 Ticket Server has also a persistent store to keep a list of administrative users for the system, and security information for collections and documents repositories. The repository is the place where documents are stored; the collection is the Verity K2 entity described in Section 2.1, that is the set of indexed documents available for searching.

Two different levels of security are supported:

- **Collection-level security**
- **Document-level security**
3.2.4 Collection-level security

K2 uses native authentication and the K2 Ticket Server to achieve collection-level security.

This means that, if a user authenticates to the repository native security model, if he belongs to a group which has permission to read them, he is able to search documents protected by that domain.

Collections can include documents from more than one repository, each with its native security model. So if a user provides information to authenticate to a repository security model and then tries to search documents in that collection, K2 prompts the user to authenticate to the others repositories too.

3.2.5 Document-level security

To limit the access to documents in a collection to certain users a Document-level security can be implemented.

Document-level security uses Verity Gateways to determine a user’s access rights for each document.

Verity K2 Gateways provide secure information access to let users retrieve and view documents in their native formats from repositories located anywhere. Each gateway uses the existing native security model of the specific repository. To access a document in a repository through a gateway, users must first provide credentials to the native security model.

Two collections of documents have been created:

- IAC (HTTP gateway, public)
- IACs (FileSystem gateway, secure)

To show some of the potentiality of the system, these two collections have been created on the same set of documents, located in D:\IAC, using different gateways (HTTP gateway and FileSystem_secure) and different access modes (public and secure).

The IAC collection doesn’t need authentication, while to access the IACs collection the user have to provide credentials in the login page.

Jobs have been created to index periodically new documents and to synchronise the collections. These jobs can be seen in the K2 Dashboard, attached to the collections [8].
3.2.6 Duplicates detection

To satisfy duplicates detection requirements, documents have been indexed using VSpider with the option “Detect Duplicates Files”.

This option enables checksum-based detection of duplicates when gathering documents as candidates for indexing.

By default, duplicate detection is disabled when indexing documents using the File System gateway. By enabling this option, a checksum is computed based on the CRC-32 algorithm. The checksum, combined with the document size, is used to determine if the document is a duplicate.

3.2.7 Knowledge trees

- Enrichment – The Enrichment Knowledge Tree has been built with the same structure as the Enrichment Topic Tree. This knowledge tree gives to the user the possibility to browse statically the topic tree and to view which documents are captured by which branch of the topic tree, corresponding to a phase of the Enrichment process.

- KTIAC – The KTIAC Knowledge Tree has the same structure as the physical structure of the repository in the File System. This structure gives a view on the physical structure of folders and can help experts to browse the repository and to find documents of interest.

- IACxCountry – The IACxCountry Knowledge Tree has been built to organize documents for country. For each category a rule has been defined for automatic documents classification; this means that if a document containing the name of a certain country is inserted into the system, that document is automatically assigned to the category for that country. The user can perform searches in the folder corresponding to a specific Country or on the whole set of documents.

To help the administrator in the definition of Topic trees and in the generation of Knowledge trees, Verity K2 software includes Intelligent Classifier.

Intelligent Classifier is a tool for creating, viewing, editing, and testing Verity topics, taxonomies, knowledge trees, and parametric indexes [14].

An alternative to the Intelligent Classifier, to create and manage Knowledge Trees, is the command-line utility ktmgr.

Intelligent Classifier has been used to define and generate knowledge trees; the command line utility ktmgr has been used to create batch files, to synchronise knowledge trees with changes occurred in the collection [14]. The batch files corresponding to the above described knowledge trees are named KTIAC.bat and KTEnrichment.bat.

They are in the D:\bat folder and have been scheduled to run periodically (once a day).
3.2.8 Web interface

A Web interface has been developed to access the Search and Browse templates.

3.3 ENRICHMENT Topic tree

A Topic Tree allows users to search for concepts in documents, rather than individual words or phrases. The Verity System treats specific words and phrases as evidence of the presence of a concept in a document i.e. it allows giving a context to words.

Search terms are encapsulated in this entity called “Topic Tree”.

Once the “Topic Tree” has been built and inserted into the system by the administrator, each user can take advantage of it. So a user can perform a search on certain technical terms without being an expert, as the System will use that terms in the specific context.

Moreover, the possibility to use a dictionary allows carrying out multilingual searches.

In collaboration with I. Maschio, a topic tree was modelled related to the subject area of the ENRICHMENT in the context of the nuclear fuel cycle [2].

The Enrichment Topic Tree has been migrated from Verity Portal One to Verity k2.

To enable users performing searches, the system has been configured registering the Topic Tree to the collections (see Section 5).

Users can execute the Topic Tree or each of its sub-topics.

In the figure 3.2 the tree-like structure of the ENRICHMENT topic tree is presented:

![Figure 3.2 Tree-like structure of the ENRICHMENT topic tree.](image-url)
At the root of the tree there is the main concept (topic), this concept is decomposed in sub-concepts, and so on until we reach a single word as you can see in the figure 3.3 where some nodes have been expanded:

```
Figure 3.3  Partial expansion of the ENRICHMENT topic tree.

Relevant operators used in the topic tree can be classified in two categories:

- **“Evidence operators”**:
  - `<THESAURUS>`: expands the search to include the word you enter and its synonyms. As explained in the Section 3.2, a standard Thesaurus is delivered with K2 and can be customized, adding terms for the specific field of application, for the user needs. As much synonyms as possible have been inserted into the customized Thesaurus and the `<THESAURUS>` operator has been applied to that terms.
  - `<WILDCARD>`: Matches wildcard characters in search strings.

- **“Proximity operators”**: 

```
- <NEAR/N>: Selects documents containing two or more search terms within \( N \) number of words of each other, where \( N \) is an integer between 1 and 1024. The closer the search terms are within a document, the higher the document's score.

- <SENTENCE>: Selects documents that include all of the words you specify within the same sentence.

More topic trees will be developed for all the other nuclear fuel cycle phases (mining and milling, conversion, fuel fabrication, reactors, heavy water production, reprocessing, spent fuel management).
4. Verity K2-IAC user manual

This section describes the set of operations to be performed for:

- Uploading documents
- Opening a Verity session
- Searching documents:
  - Simple searches
  - Advanced searches
  - Searching documents through Knowledge Trees
- The use of Topic Trees for searches

4.1 Procedure to upload documents

To allow an easy upload of files to the system, a folder structure, shared through the internal network, has been created on the server that hosts Verity K2. This makes sending documents to IAC very easy. So, when a user goes on the Internet and finds documents, he can directly upload files in the folders they have access to, without being obliged to open a Verity session. The system is in charge to update Verity collections to synchronise them with the files added or modified.

The system can index document of almost any format [12]. However, regarding specific formats such as images Verity K2 can index only metadata. In order to keep the metadata attached to the document, the suggestion is to convert images into pdf format, adding searchable metadata to the document properties before inserting them into the system.

The same conversion is recommended for documents in Microsoft mht format, as K2 doesn’t provide a filter for this format.

4.2 To open a session

- Open Internet Explorer and type the following address: http://safepc113/IAC

The visibility of the web site is internal to JRC but the access is restricted, by specific permissions, to authorised users

  - “Home” page (see fig. 4.1) corresponding to the “Home” link:

This is the welcome page introducing to IAC.
What is IAC?

The PSC Institute of JRC started in 2000 to organise an Information and Analysis Centre (IAC) with the purpose of collecting information concerning Nuclear, Chemical and Biological Non-Proliferation, Disarmament and Weapons of Mass Destruction from the Open sources, analysing and validating them and then producing studies on related subjects.

The data base consists of political, commercial and social information such as treaties, agreements, conventions, regulations, articles etc. which are in direct relation with the scientific/technical problems of NPF, Disarmament, and so on.

How to use it?

- **Search**: Provides a basic/advanced search page. Select from the list of currently available collections the one(s) you want to search. Performs free-text or fields based searches in those collections.

- **Browse**: Provides a number of tree-like views to browse and search documents organised into categories.

- **Upload**: To gain permissions to upload documents contact the System administrator.

Figure 4.1 IAC Home page.
o **“SEARCH” page** (see fig. 4.2) corresponding to the “Search” link:

This page allows the user to search for documents in the collections he has access to.

He can also have access to secure collections; in that case the user must provide credentials in the Authentication window clicking the login link (see section 4.2.3).

![SEARCH page](image)

Figure 4.2 SEARCH page.
“BROWSE” page (see fig. 4.3) corresponding to the “Browse” link:

This page allows the user to navigate through a tree-like structure that organises documents in “views” similar to a directories structure. The organisation is based on categorisation rules, for instance documents could be organised for country.

On the right of the Search field the combo box allows listing all the knowledge trees that can be browsed and searched in.

Once selected, the tree is shown in the left-bottom section of the page, and in the right side there is the Search Results list, where all the documents found for the specific search performed are listed according to the score.

It should be noted that the browse function is performed in the collections the user has access to.

The Browse functionalities are presented in paragraph 4.3.4

![Figure 4.3 BROWSE page.](image)

“login” (Searching Secure Collections) (see fig. 4.4):

In each of the “Search” or “Browse” pages, a link towards a login box is displayed on the left side, above the frame.
As said previously, collections can be created with different levels of security.

In the “Simple Search” page, the secure collection(s) are not visible (see fig. 4.2). To search secure collection users having access to collections other than the public ones must provide credentials in the Authentication window for logging into the domain (see fig. 4.5).

Clicking the login link, above the search box as shown in fig. 4.4.

Then provide credential and click Submit.
The collections the user has access to will be shown in the list. Then the search is performed in the same way as in section 4.3.1 and section 4.3.2.

4.3 Searching documents

4.3.1 Simple Searches

Figure 4.6 shows the “Search” page:
In the leftmost side of the picture there is the “Search” field where the user can launch searches with the parameters described below on some terms.

The following **Options** are available to the user.

1. **Search**: simple word or sequence of words. The system will find and rank those documents matching the word or sequence of words, in the example the word “nuclear” (see fig. 4.4).

2. **New Search/Refine Search**: The default is “New Search” to tell the system to perform a completely new search; if “Refine Search” is selected instead, the system allows the user to refine searches inside the results of the initial search (see fig.4.6).

3. **Results**: In this list the user can select “Verbose” or “Brief” (see fig.4.7). “Verbose” is to show, in the Results List, the summary and some metadata of each document found; “Brief” to see just the title for each document in the Results List.

4. **Max Docs Returned**: In this combo box the user can select the maximum number of documents returned in the Results List (see fig.4.7).

5. **VQL Operators**: This is the list of Verity Query Language operators. Clicking the specified operator it will be copied in the Search box, helping the user in the query definition.
6. **Choose collection to search over:** the user selects one or more collections in those to which he has access.

   The collections listed in the bottom part of the page are all the **Collections enabled or Public Access**:
   
   - **IAC (online):** The collection of IAC documents.
   - **verity_docoll (online):** The collection of Verity technical documentation.
   - **minicar (online):** An example collection
   - **usercoll (offline):** An example collection for the administrator

7. **Advanced search:** see paragraph 4.3.2

![SEARCH page description-2](image)

At the bottom of the page there is the Search Results list, where all the documents found for the specific search performed are listed according to the score.

**To search documents and display of the results:**

- Click the “Search” link to go to the Search page (see fig. 4.4).
• Then select one or more of the options described in 1 to 7 above, according to the objectives of the search

• Click “Search”. The Results list will appear at the bottom of the page, in the “Search Results” section (see fig. 4.5)

4.3.2 Advanced Searches

The parameters used for advanced searches, for example searches on metadata are displayed on the right side the “Search” page.

The specific fields available for advanced searches are: title, authors, or summary content, period of time. They can be used in addition to the fields specified in 4.3.1 for simple search, in particular regarding the keywords and collections to search in.

• Fill in one or more fields in the Advanced Search section (see fig. 4.8)

• Type the word(s) in the Search text box and click the Search button.

• The Resulting document list will appear in the bottom part of the “Search” page, i.e. in the “Search Results” section as shown fig. 4.9.

Figure 4.8 Advanced Search fields.
4.3.3 Searching documents through Knowledge Trees

This page allows the user to search for documents in a collection logically organised in a tree-like structure similar to a file system. In fact, in some cases the user prefers to organise documents according to a logic (for example for Country), which helps him navigating through the documents and finding what he is looking for (in the same way as if documents were organised in directories on his disk). The structure the user sees is just a “view” in the sense that documents are not physically stored in this way, but the system allows to show them in an organisation, according to a certain logic, that draw the user through his search. The administrator can build, for a set of documents, different views for each user, based on classification rules.

- Click the “Browse” link.
- Select a Knowledge Tree from the List of Knowledge Trees (see fig. 4.10. In the example a tree that organises documents according to the Country, “KTIAxCountry”)
• The Knowledge Tree structure will appear in the left side of the page (see fig. 4.10).

The tree is structured in Categories (in the example Countries: Algeria, Argentina, and so on).

• Browse the tree selecting a Category (in the example a Country) and on the right side of the page the list of documents belonging to that category will be displayed.

• To view a document click the link to the document and the document will be displayed.

The user can also perform a search on some criteria in a particular Category:

• Select a Category of the tree (for instance Iran) and type a query in the search text box (for instance “nuclear”) (see fig. 4.11).
Figure 4.11 Browse page – Query in the Iran category on the “nuclear” term.

- Click “Search”.

The documents list, which is displayed in the right side of the page, relates to nuclear activities in Iran. In this case (search on a criteria) documents are listed according to the score (in decreasing order).

### 4.4 Use of topic trees for searches

The figure 4.12 gives an idea about the use of topic Trees. The user enters one or more keywords in the Search textbox using (if necessary) the logical connectives AND, OR, NOT. The Verity K2 interprets the query and, if a Topic Tree is present it is triggered thus enriching substantially the number of keywords used and the logic of the query for selecting the documents of interest.
When the query is submitted the extracted documents are ranked according to their relevance. Finally each document can be displayed just by clicking on its title.

In the example shown in figure 4.13, the IAC collection has been searched typing the keyword ‘enrichment’. The system finds a Topic Tree corresponding to that keyword and triggers it. For each document a score is calculated indicating the relevance to the topic tree. Opening one document in the search result list, some of the keywords specified in the Enrichment Topic Tree can be seen highlighted.
As explained in Section 3, the Enrichment Knowledge Tree has been built with the same structure as the Enrichment Topic Tree. This knowledge tree gives to the user the possibility to browse statically the topic tree and to view which documents are captured by which branch of the topic tree, corresponding to a phase of the Enrichment process.

On the left side of figure 4.14 the Enrichment Knowledge Tree can be seen.

Clicking one of the sub-processes the user can perform searches on some terms in that category. For example clicking the ENRICHMENT_OF_UCL4 category and typing the word “pollution” in the search text box, only documents in the selected category will be searched. This allows the user to search and browse the knowledge tree corresponding to the Enrichment Topic Tree.
Figure 4.14  An example of use of Enrichment Knowledge Tree for searches.
5. Conclusions and Further Developments

The objective of the Information and Analysis Centre (IAC) is the possibility of collecting, analysing and validating information concerning Nuclear, Chemical and Biological Non Proliferation, Disarmament and Weapons of Mass Destruction, producing studies on related subjects.

Open Source provides a lot of information, regulations, treaties, technical documents and news, which cannot be lost. This implies that this growing mass of data must be easy to handle and search.

Documents can have different formats. The K2 system can index document of almost any format, e.g. Word, pdf, HTML, and so on.

This information, manually or automatically collected can contain a lot of duplicates, so duplicates detection is necessary.

Moreover, some documents begin to be sensible, so a hierarchy or access rights should be defined.

Finally K2 System can be accessed by other systems. For example, of great interest is the connection of the Verity system with the JRC GIS driven geo-data base NUMAS (NUclear MAppling System). The main improvement, which makes the searching mechanism much more powerful, is the use of Topic trees associated to nuclear sites. This will gives a context to query strings sent by NUMAS to VERITY [4].

More topic trees should be developed in the future in the context of Nuclear-Fuel Cycle, in order to increase the powerfulness of the Verity system as the management tool of IAC documents.
Acknowledgments

I wish to acknowledge everyone who gave me technical and moral support during my work, especially: my Unit Heads A. Poucet, Joao G.M. Gonçalves and W. Janssens who gave me the opportunity to increase my knowledge and experience in Knowledge Management; F. Mousty and S. Contini for their friendship and collaboration; G.G.M. Cojazzi, action leader of the action AMENUS where this activity has been carried out, for his precious supervision and professional advice in the writing of this report; G. Renda for the helpful and valuable discussions and support and F. Sevini for his collaboration and suggestions.

Particular acknowledgement goes to my friend L.V. Bril for the professional collaboration, the constant support and for believing in me.

Finally, I would like to thank my friend F. Bellezza for providing technical support.
References


Abstract

The IPSC Institute of JRC Ispra started to organise an Information and Analysis Centre (IAC) with the purpose of collecting information concerning Nuclear, Chemical and Biological Non-Proliferation, Disarmament and Weapons of Mass Destruction from the Open sources, analysing and validating them and then producing studies on related subjects. After an investigation of the commercial documentation management systems, Verity Portal One Application Suite © developed by Verity Incorporated was selected. This report contains a description of the work carried out for IAC management through Verity K2 system. The objective of this report is to provide a guide for users, a support for administrators and to point out the potentiality of the System for further developments.
The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.