



OKKAM – Enabling a Web Of Entities

Grant Agreement No. 215032

D7.1: First OKKAM empowered tool suite

Document Number	D7.1
Document Title	First OKKAM empowered tool suite
Version	3
Status	Final
Work Package	WP7
Deliverable Type	Report
Contractual Date of Delivery	23/03/2009
Actual Date of Delivery	
Responsible Unit	
Contributors	ExpertSystem, UNITN
Keyword List	
Dissemination level	PU

Change History

Version	Date	Status	Author (Company)	Description
1	09/02/2009	Draft	Francesco Danza (Expert System)	First draft, document skeleton
2	13/02/2009	Draft	Angela Fogarolli (UNITN)	Input UNITN
4	14/02/2009	Draft	Daniele Cordioli (Expert System)	Input Expert System
5	23/03/2009	Final	Francesco Danza (Expert System)	Approved

Executive Summary

In this document we provide a guide to the first OKKAM empowered tool suite.

The first part of this document describes the first OKKAM empowered tool architecture. The objectives are to design and develop OKKAM architecture for supporting the creation of new *OKKAMized* content. In general, this will be achieved by designing and implementing plug-ins for popular content creation tools, like office applications, XML and ontology editors, etc.

In the second part we focus on every single OKKAM empowered tool developed during this first period.

Table of Contents

1. INTRODUCTION.....	6
1.1. WHAT IS OKKAM?.....	6
1.2. THE SCOPE OF THIS DOCUMENT	6
2. OKKAM EMPOWERED TOOL: ARCHITECTURE.....	7
2.1. 3-TIER ARCHITECTURE	7
2.2. THE OKKAMIZERS PIPELINE.....	9
3. FIRST OKKAM EMPOWERED TOOL SUITE.....	12
3.1. FOAF-O-MATIC.....	12
3.1.1. <i>Environment</i>	12
3.1.2. <i>Features</i>	12
3.1.3. <i>Technology</i>	12
3.2. OKKAM4P - PROTÉGÉ.....	12
3.2.1. <i>Environment</i>	12
3.2.2. <i>Features</i>	13
3.2.3. <i>Technology</i>	13
3.3. OKKAM TOOLBAR FOR FIREFOX.....	13
3.3.1. <i>Environment</i>	13
3.3.2. <i>Features</i>	13
3.3.3. <i>Technology</i>	13
3.3.4. <i>Installation</i>	13
3.4. OKKAM TOOLBAR FOR IE.....	14
3.4.1. <i>Environment</i>	14
3.4.2. <i>Features</i>	14
3.4.3. <i>Technology</i>	14
3.4.4. <i>Installation</i>	14
3.5. OKKAM4MS FOR MICROSOFT WORD	14
3.5.1. <i>Environment</i>	14
3.5.2. <i>Features</i>	15
3.5.3. <i>Technology</i>	15
3.5.4. <i>Installation</i>	15
3.6. OKKAM4MS FOR MICROSOFT OUTLOOK.....	15
3.6.1. <i>Environment</i>	15
3.6.2. <i>Features</i>	16
3.6.3. <i>Technology</i>	16
3.6.4. <i>Installation</i>	16
4. DOWNLOAD AND SCREENSHOTS	17
4.1. FOAF-O-MATIC.....	17
4.2. OKKAM4P – PROTÉGÉ.....	18
4.3. OKKAM TOOLBAR FOR FIREFOX.....	19
4.4. OKKAM TOOLBAR FOR INTERNET EXPLORER	20
4.5. OKKAM4MS FOR MICROSOFT OUTLOOK.....	20
4.6. OKKAM4MS FOR MICROSOFT WORD	21
5. CONCLUSIONS AND NEXT STEP	23
6. REFERENCES	24
7. APPENDIX – INFORMATION ON THE COGITO TECHNOLOGY	25
7.1. SEMANTIC ANALYSIS	25
7.2. COGITO® SEMANTIC TECHNOLOGY.....	25
7.2.1. <i>Parser</i>	27
7.2.2. <i>Lexicon</i>	27
7.3. COGITO® DISCOVER.....	29

Glossary

OKKAM	IST 7th Framework Research Project
ENS	Entity Name System (developed by the OKKAM project)
API	Application Programming Interface
URL	Uniform Resource Locator
XML	Extensible Markup Language
RDF	Resource Description Framework
RDFa	Resource Description Framework in attributes
WSDL	Web Service Description Language
XUL	XML User Interface Language
Sensigrafo [®]	a semantic network is a representation of linguistic knowledge and world knowledge. It is an oriented graph consisting in tags representing concepts, and arcs representing the conceptual relationship between the concepts
COGITO [®]	Cogito is the pioneer semantic software developed by Expert System. It understands the meaning of words, like a person does when reading. That's why it's a unique technology.
NLP	Natural Language Processing is the extraction process of semantic information from human being expressions (spoken or written) by a PC. This process is particularly complex due to the natural difficulty of the language caused by ambiguity.
REST	is a style of software architecture for distributed hypermedia systems such as the World Wide Web.
OKKAMization	is the process of recognizing the relevant entities in a piece of data (relational, XML, RDF/OWL, plain text, etc.) and creating an internal annotation which include its global, stable identifier.
OKKAMizer	is a tool which allows the OKKAMization of a specific type of content OKKAM-empowered tool: it is an application (e.g. word processor, ontology editor, blog platform, email client, etc.) which is extended (e.g. through a plug-in) to interact with the ENS for bringing benefits to users based on the occurrence of identifiers in the content.

1. Introduction

1.1. What is OKKAM?

OKKAM is a Large-Scale Integrating Project funded by the European Commission under the 7th Framework Program (FP7), running until **June 2010**.

The **OKKAM** project aims at *enabling* the **Web of Entities**, namely a virtual space where any collection of data and information about any type of entities (e.g. people, locations, organizations, events, products, ...) published on the Web can be integrated into a single virtual, decentralized, open knowledge base (like the Web did for hypertexts.)

OKKAM will contribute to this vision by supporting the convergence towards the use of a *single and globally unique identifier* for any entity which is named on the Web. The intuition of the project is that the concrete realization of the Web of Entities requires us to enable tools and practices for cutting to the root the proliferation of unnecessary new identifiers for naming the entities which already have a public identifier (the *OKKAM's razor*). Therefore, **OKKAM** will make available to content creators, editors and developers a global infrastructure and a collection of new tools and plug-ins which support them, to easily find public identifiers for the entities named in their contents/services, use them for creating annotations, and build new network-based services which make essential use of these identifiers in an open environment (like the Web or large Intranets).

To realize this vision, **OKKAM** proposes the following roadmap:

- Providing a scalable and sustainable infrastructure, called the **Entity Name System (ENS)**, for making the systematic reuse of global and unique entity identifiers not only possible, but easy and straightforward. The ENS will be a distributed service which permanently stores identifiers for entities and provides a collection of core services (e.g. entity matching, ID mapping and resolution) needed to support their pervasive reuse;
- bootstrapping and enabling the fast growth of Web of Entities by fostering the creation of **OKKAMized content** (i.e. content where entities are named or annotated with OKKAM IDs) in **OKKAM-empowered applications** (i.e. applications which can interact with the ENS for getting and reusing identifiers);
- showcasing the benefits of enabling the Web of Entities and, more generally, of an entity-oriented approach to content and knowledge management by building **relevant applications** on top of the new infrastructure in three important areas: *information retrieval and semantic search*, *content authoring* (more specifically, in *scientific publishing* and *news production*) and *organizational knowledge management*.

More up-to-date information can be found on the OKKAM webpage: <http://www.okkam.org>.

1.2. The scope of this document

This deliverable gives an overview of the first OKKAM Empowered tools suite. First we describe the architecture and the APIs which are available for applications to consume the services of the ENS. The second part focuses on every OKKAM Empowered tool developed for this first suite.

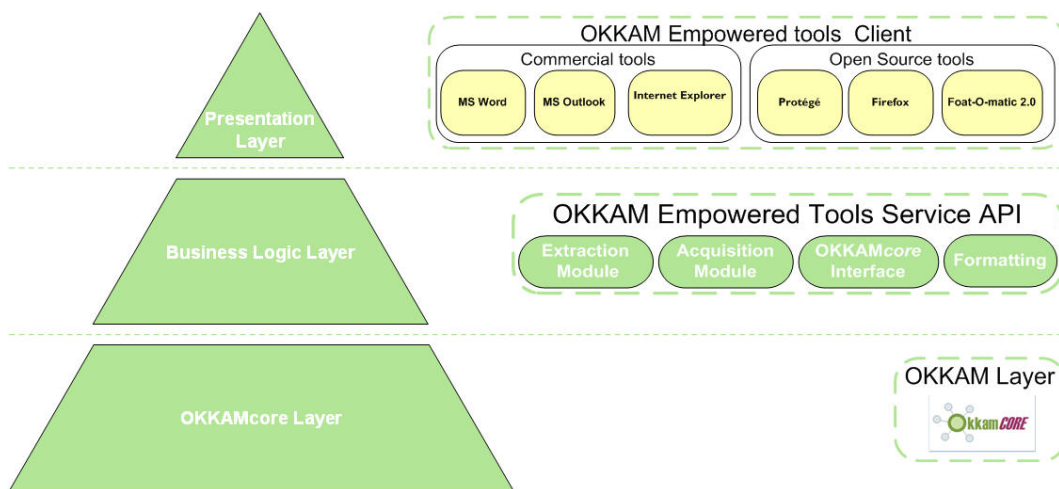
2. OKKAM Empowered tool: Architecture

The OKKAM empowered tools architecture is a typical three tier architecture. This is a client-server architecture in which the user interfaces, business logic, computer data storage and data access are developed and maintained as independent modules, on separate platforms.

In the next paragraph we describe the architecture and the workflow of process.

2.1. 3-tier Architecture

As we can see in the following picture, the system is based on a classic three tier architecture:



- **OKKAMcore Layer**

This layer contains the ENS services, which are available via Web Services. The ENS stores unique identifiers and offers services to applications in order to retrieve them. The information is then passed back to the logic layer for processing, and then eventually back to the user. A description of the Web Services API is publicly available at:

- <http://www.okkam.org/apis/web-service-api>, which contains further examples as well.
- WSDL URL: <http://api.okkam.org/okkam-core/services/WebServices?wsdl>
- SERVICE ADDRESS: <http://api.okkam.org/okkam-core/services/WebServices>

- **Business Logic Layer**

This layer contains the Application services API. In particular, it coordinates the application, processes commands, and makes logical decisions and evaluations. The Application Services API automatically extracts entities from documents and Web resources; it also moves and processes data between the two surrounding layers.

The Application Services API allows the following operations starting from a document:

- Recognizing the named entities like People, Companies, Location, and other
- Recognizing the OKKAM entities
- Generating the OKKAM core queries extracting the named entities
- Enriching the document with the OKKAMid
- Enriching the document with the RDFa

WSDL URL:

<http://host.expertsystem.it/OkkamWebService/services/OkkamWebService?wsdl>

- **Presentation Layer**

This is the topmost level of the application. The presentation tier displays information related to such services for each tool, in commercial and open source environments. The first OKKAM empowered tool suite contains plug-ins for the following tools:

- Foaf-O-matic
- Okkam4P - Protégé
- OKKAM toolbar for Firefox
- OKKAM toolbar for Internet Explorer
- OKKAM4MS for Microsoft Word
- OKKAM4MS for Microsoft Outlook

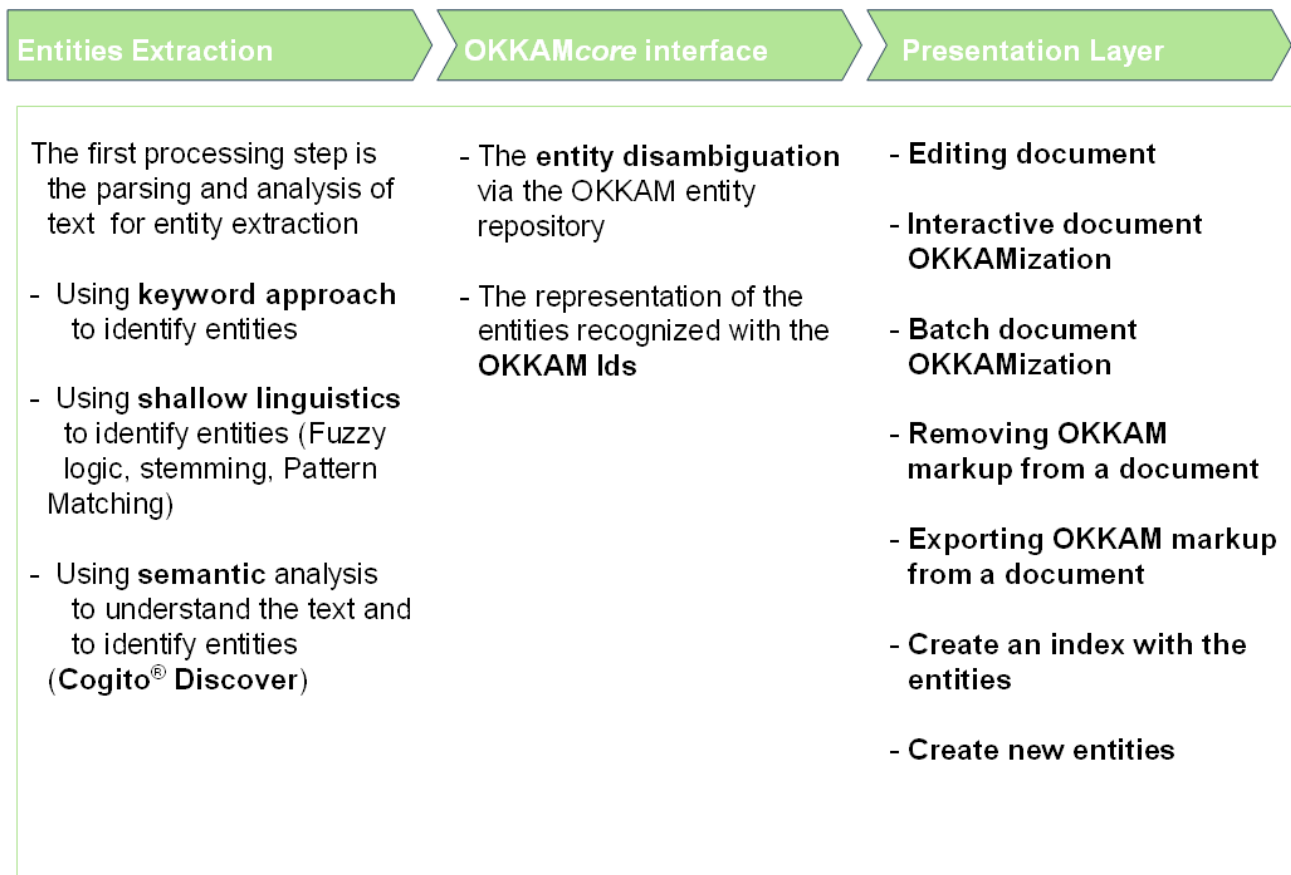
Using this approach we are able to create a set of OKKAM Empowered tools based on a shared platform. This means that:

- The solution allows integrating different tools in the architecture in a non-invasive way, because the connection is related only to the presentation layer. Changes, updates and new features involve only the presentation layer, not the logic layer.
- Using RESTful web service (simple web service implemented using HTTP and the principles of REST) or SOAP methodology the solution can be integrated with other different platforms.
- The system is able to manage different kind of documents: pdf, txt, html, xml, office.
- Using this architecture it is possible to provide integration capability based on standard technology like Java Message Service, Web Service, HTTP interface.

2.2. The OKKAMizers Pipeline

This paragraph focuses on the different steps of the OKKAMization process. In particular, we will analyze how the different tier moves and processes data, in order to better understand the OKKAM Empowered tool functionality.

The following picture describes the three main steps of this workflow:



- **Entities extraction**

It is clear that every tool analyzes different kinds of information. Depending on the situation, the gathering of data (ACQUISITION) can be carried out in various ways. After acquisition, the first processing step is the parsing and analysis of a text for entities extraction. In particular, the system contains three different approaches:

- Keyword based
- Shallow linguistic
- Semantic Analysis

The main module is the semantic analysis based on COGITO[®] Discover semantic capabilities, which provides Semantic Analysis for texts in English and Italian. Once acquired, data are cached and immediately analyzed using the semantic approach. Following the NPL, the whole text is disambiguated, where disambiguation is the process defining the meaning of one or more words in a given text, when words have distinct meanings. Disambiguation is solved through adequate algorithms. This process is based on the Sensigrafo[®], a semantic network that contains a representation of linguistic knowledge and world knowledge. It is an oriented graph consisting in tags representing concepts, and arcs representing the conceptual relationship between concepts.

See the end of the document for more information about the semantic approach.

The extractor used by the system is architecture independent, the pipeline currently contains not only COGITO[®] Discover, but also:

- the Sap extractor, in order to recognize Sap products
- Protein's extractor based on Keyword Approach

- **OKKAMcore Interface**

The Entity Matching stage generates an OKKAM ID for entities, where this is required. This process is based on the following steps:

- **Querying the system**
A query is created for each identified named entity (e.g. people, location). It contains the main features useful to better identify the potential entity in a unique way. The query is sent to the OKKAM Engine.
- **Query processing**
The OKKAM Engine analyzes the query, searches the repository and returns, if present, the correct OKKAM ID for the named entity. If the named entity is not in the repository, another task starts, creating a new persistent web identifier for entities which don't have one yet.
- **The correct OKKAM ID identification**
The OKKAM ID resulting from matching is a value ready to be included as a new field in the document.

The use of this module is the basis for building a Web of Entities, where information about the same entity is consolidated in such a way that aggregation, integration and mash up become easier and faster.

- **Presentation Layer**

At this stage, the system is able to perform different operations depending on the type of tool that is used. The main features are:

- Editing a document
- Interactive document OKKAMization
- Batch document OKKAMization
- Removing OKKAM mark up from a document
- Exporting OKKAM mark up from a document
- Creating an entities index
- Creating a new entity
- Searching for entities in the repository or in a document

In the next paragraph we will analyze the different tools developed.

3. First OKKAM Empowered tool suite

The first OKKAM Empowered tool suite contains the tools described in the following paragraphs.

3.1. Foaf-O-matic

3.1.1. Environment

FOAF is a simple technology that makes it easier to share and use information about people and their activities, to transfer information between Web sites, and to automatically extend, merge and re-use it online.

FOAF is an interface for creating and updating FOAF profiles, extended by the OKKAM infrastructure, for issuing "friends" with globally unique identifiers. All these actions are performed with just a few mouse clicks.

3.1.2. Features

- graphic interface to ease the creation of FOAF RDF files
- easily adding friends to your profile with OKKAM and Sindice
- creating your own OKKAM ID on the fly!
- uploading and updating already defined FOAF profiles
- personalizing a user's description

3.1.3. Technology

This Rich Internet Application is created using:

- IceFaces
- Sun JSF RI
- Facelet

RDF handling is supported by Jena

3.2. Okkam4P - Protégé

3.2.1. Environment

Protégé is a free, open source ontology editor and knowledge-base framework. The Protégé platform supports two main ways of modeling ontologies via the Protégé-Frames and Protégé-OWL editors. Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema.

This plug-in allows unique identification of entities inside text documents created with Protégé.

3.2.2. Features

- directly integrated in the Protégé interface
- easily add OKKAM IDs to instances
- detailed adaptability of search properties

3.2.3. Technology

- Protégé plug-in capability

3.3. OKKAM toolbar for Firefox

3.3.1. Environment

Mozilla Firefox is a free and open source web browser descended from the Mozilla Application Suite and managed by Mozilla Corporation. The Firefox OKKAM toolbar is a plug-in to OKKAM-enable the Firefox browser.

3.3.2. Features

- highlights OKKAM entities in an annotated website
- search entities directly from your toolbar
- easily search e.g. Sindice by clicking on an entity
- all entities and their attributes of the current page at a glance
- one-click creation of a new entity!

3.3.3. Technology

- Javascript and XUL integrates OKKAM functionalities directly in your browser

3.3.4. Installation

- Download the plug-in
- Open the file okkambar1_1.xpi (File -> Open...) and follow the Firefox add-ons installation process.
- Restart Firefox.
- Open the toolbar customization menu: View -> Toolbar -> Customize...
- Select the OKKAM toolbar and drag it to your main window toolbars.

3.4. OKKAM toolbar for IE

3.4.1. Environment

Windows Internet Explorer (formerly Microsoft Internet Explorer; abbreviated to MSIE or, more commonly, IE), is a graphical web browsers developed by Microsoft. The IE OKKAM toolbar is a plug-in to OKKAM-enable the Internet Explorer browser.

3.4.2. Features

- OKKAMize the web site you are on
- search entities
- one-click creation of a new entity!

3.4.3. Technology

- IE7Pro, an add-on for Internet Explorer which adds a lot of features and extras
- Javascript

3.4.4. Installation

- Download the plug-in
- Install IE7Pro
- Add the plug-in to your browser
- Allow the use of plug-ins in IE7Pro preference tab

3.5. OKKAM4MS for Microsoft Word

3.5.1. Environment

Microsoft Word® is a word processing software included within the Microsoft Office suite. Its first release dates back to the early eighties, the last one was in 2007. Microsoft Word is a powerful program which allows the user to create and share professional documents by using a set of writing tools and a number of utilities such as the contextual spelling checker, a thesaurus, the live word count, etc.

This tool is called Okkam4MsW, a MS Word plug-in for the Globally Unique Identification of Individuals in MS Word. This plug-in allows unique identification of entities inside text documents created with MS Word. The great circulation of MS Word as a word processor is the main reason for the choice of implementing Okkam4MSW. NLP and semantic technologies are used to detect entities in the text and extract contextual information enabling the matching decision within the OKKAM entity repository.

The plug-in can be accessed by means of a MS Word-integrated toolbar.

3.5.2. Features

OKKAMization

- Interactive document OKKAMization
Users can analyze part of a text just selecting it in advance
- Batch document OKKAMization
Users can analyze a text in background mode
- Interaction with OKKAM
Users can select the correct OKKAM entity

Reporting

- Create automatically an Index inside the document
- Export all entities in other formats
- Remove the OKKAMization

3.5.3. Technology

- natural language processing and semantic technologies are used for entities detection

3.5.4. Installation

- Download the plug-in setup
- Follow the step by step instructions

3.6. OKKAM4MS for Microsoft Outlook

3.6.1. Environment

Microsoft Outlook® is a personal information manager (PIM), part of the Microsoft Office suite, even though it is mainly known as an email manager. This application has a few tools:

- a contact manager;
- a calendar;
- a task manager
- a to-do bar.
- the Email manager

This tool is called Okkam4MsW, a MS Outlook plug-in for the Globally Unique Identification of Individuals in MS Outlook. This plug-in allows unique identification of entities inside an email created with MS Word.

NLP and semantic technologies are used to detect entities in the text and extract contextual information enabling the matching decision within the OKKAM entity repository.

The plug-in can be accessed by means of a MS Outlook-integrated toolbar.

3.6.2. Features

OKKAMization

- Batch document OKKAMization
Users can analyze a text in background mode
- Interaction with OKKAM
Users can select the correct OKKAM entity

Reporting

- Create automatically an Index inside the document
- Export all entities in other formats

3.6.3. Technology

- natural language processing and semantic technologies are used for entities detection

3.6.4. Installation

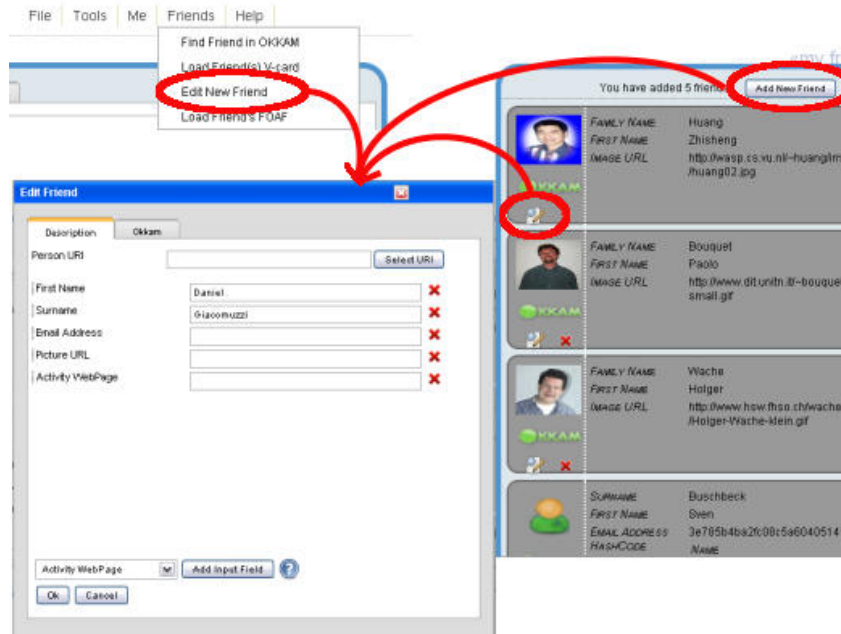
- Download the plug-in setup
- Follow the step by step instructions

4. Download and Screenshots

The first OKKAM Empowered tool suite is available at this web site:

<http://www.okkam.org/test-tubes>

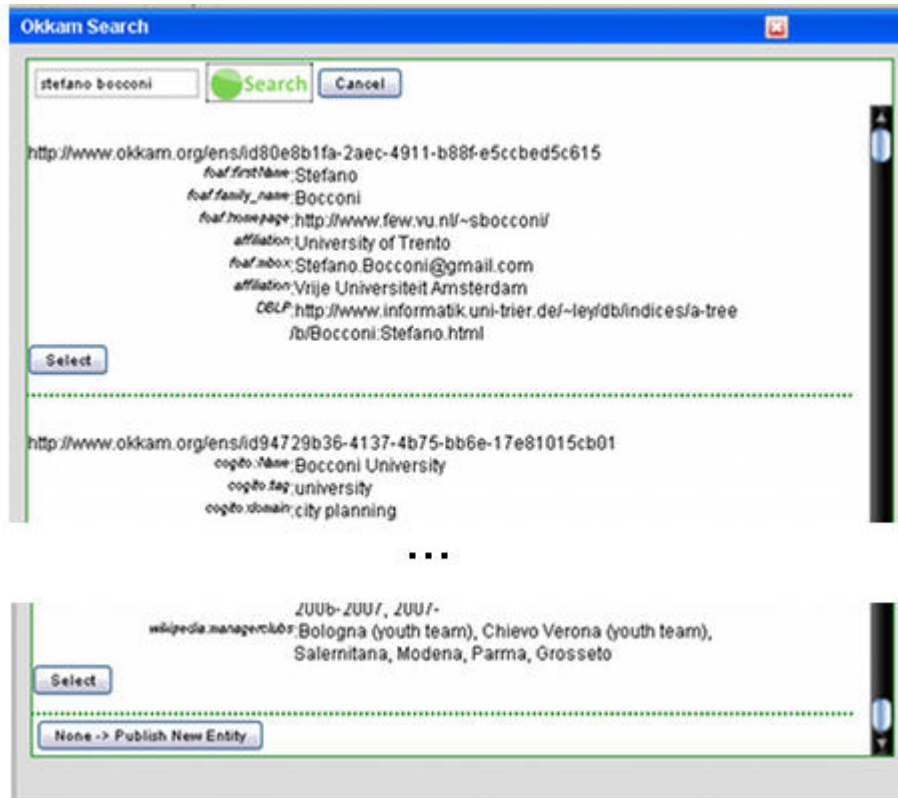
4.1. Foaf-O-matic



4-1 Foaf-O-Matic: Edit new friend

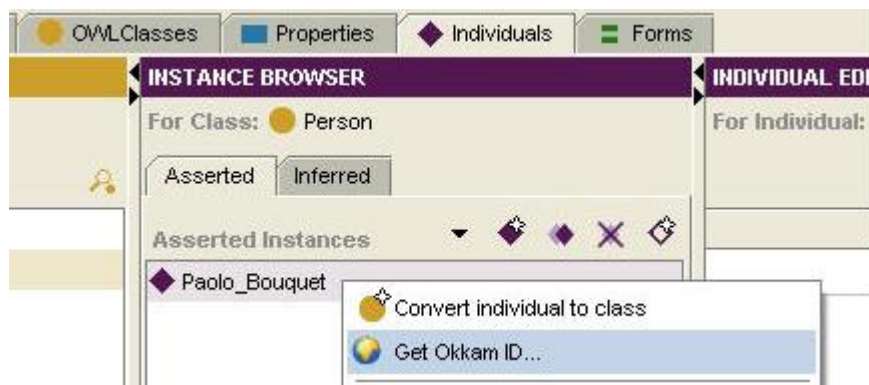


4-2 Foaf-O-Matic: interface



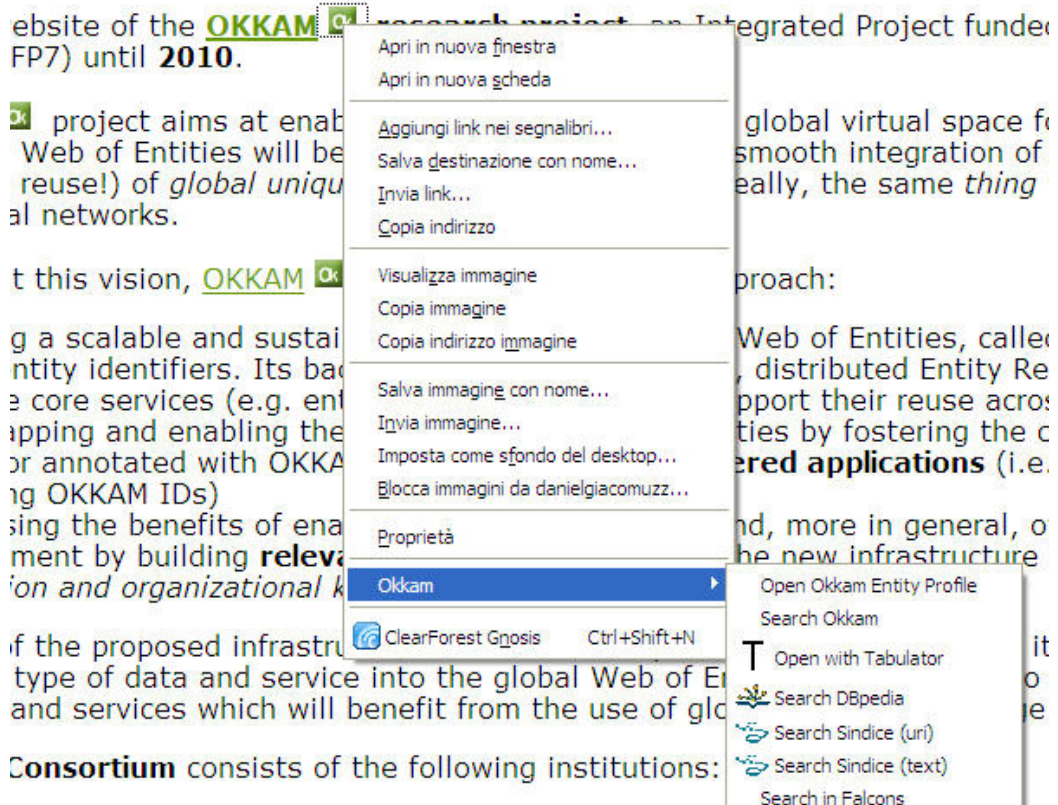
4-3 Foaf-O-Matic: Search in OKKAM

4.2. Okkam4P – Protégé

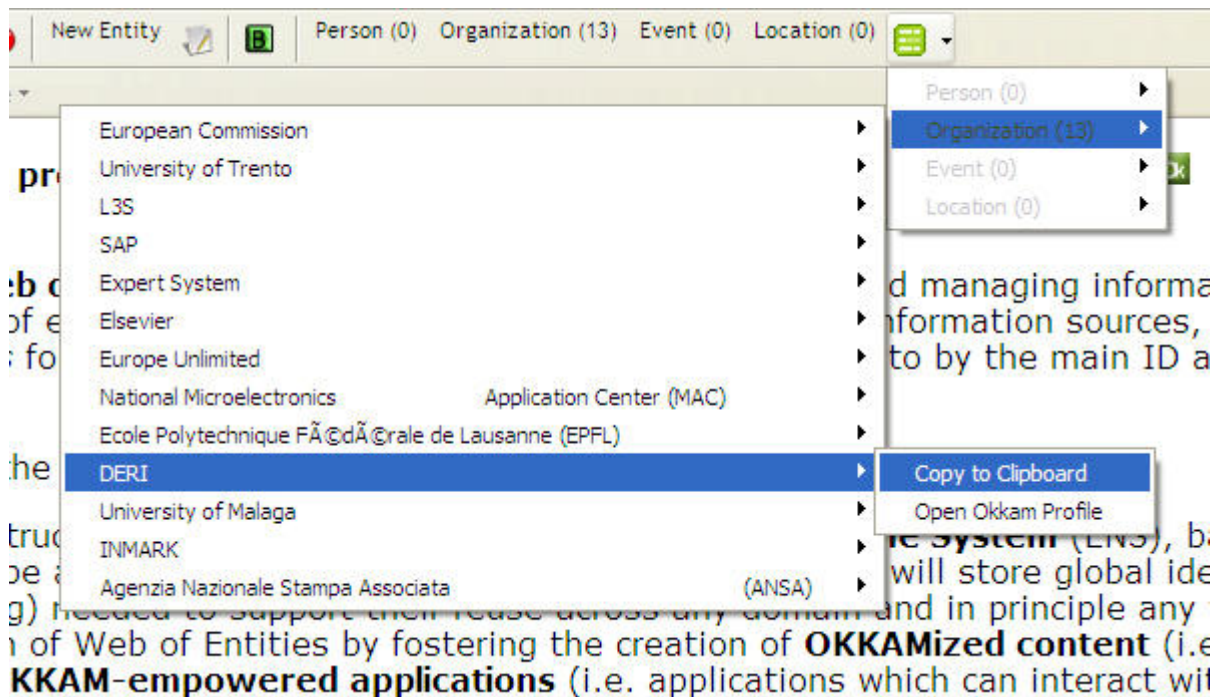


4-4 OKKAM4P: Generic interface

4.3. OKKAM toolbar for Firefox

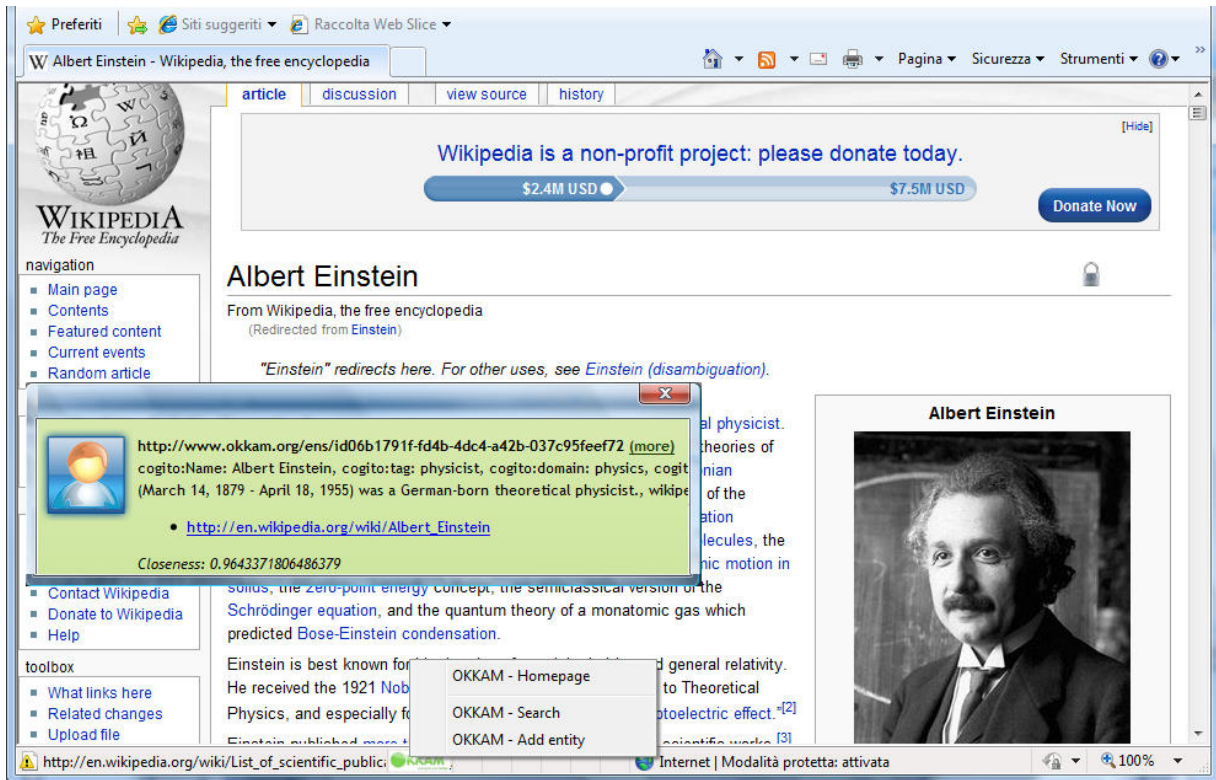


4-5 OKKAM4Firefox: Menu



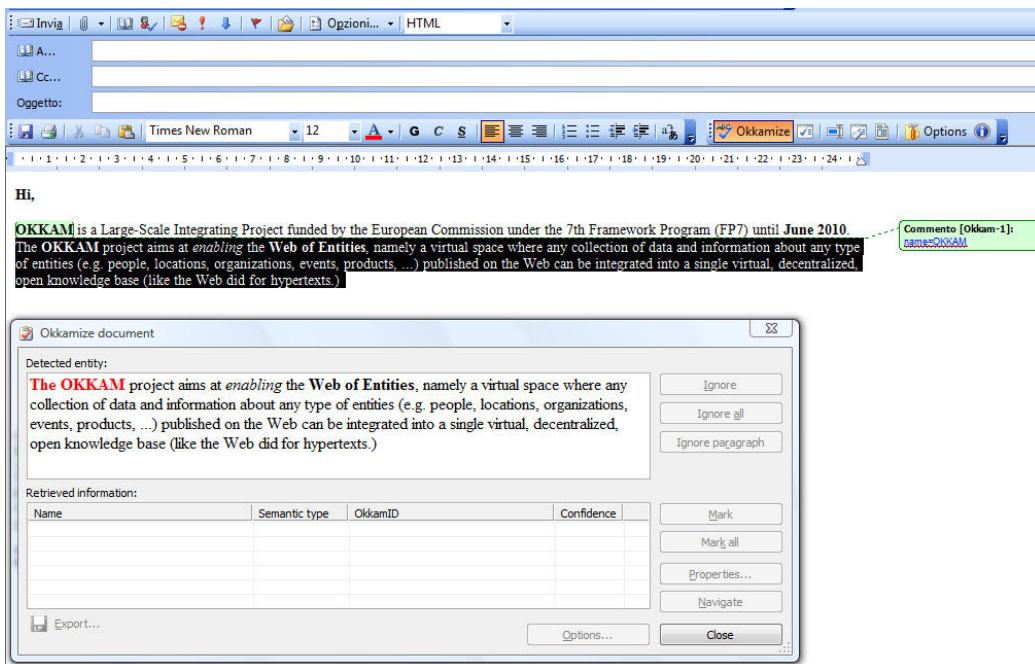
4-6 OKKAM4Firefox: Features

4.4. OKKAM toolbar for Internet Explorer



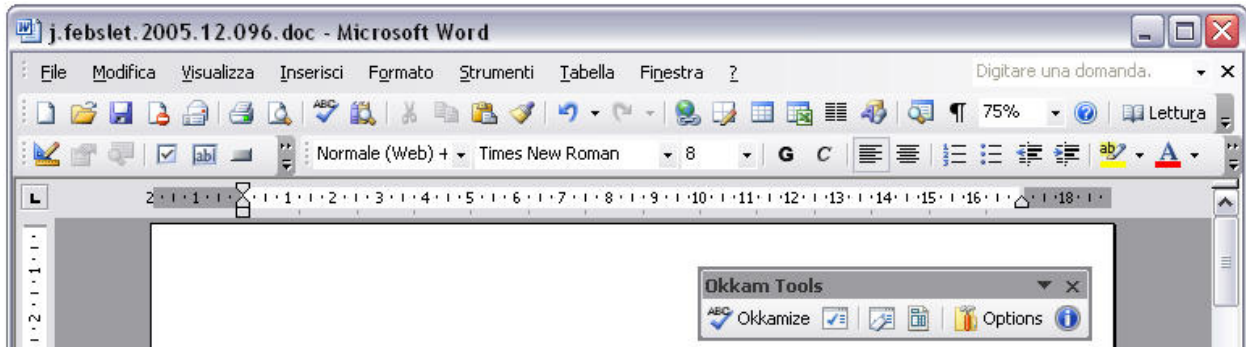
4-7 OKKAM4IE: Menu

4.5. OKKAM4MS for Microsoft Outlook



4-8 OKKAM4MS Outlook: Create a new email with OKKAM

4.6. OKKAM4MS for Microsoft Word



4-9 OKKAM4MS Word: Okkam toolbar menu

FEBS Letters 580 (2006) 940–947



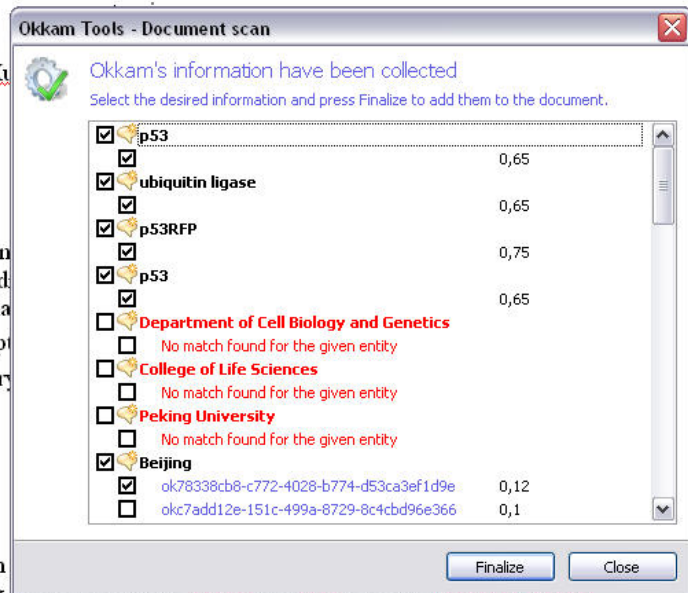
The p53-inducible E3 ubiquitin ligase p53RFP induces p53-dependent

Jun Huang^a, Liang-Guo Xu

^aDepartment of Cell Biology and Gen
100871, China ^bNational Jewish Med
Sciences, Wuha

Received 26 July 2005; revised 20 Sep
18 January

Abstract Recently, it has been shown
finger (IBR)–RING domain-containing



4-10 OKKAM4MS Word: Batch Okkamization

FEBS Letters 580 (2006) 940–947



The p53-inducible E3 ubiquitin ligase p53RFP induces p53-dependent apoptosis |

Jun Huang^a, Liang-Guo Xu^b, Ting Liu^a, Zhonghe Zhai^a, Hong-Bing Shu^{c,*}

Commento [Dkkam-1]:
name=uniprotkb:P0Z340

Commento [Dkkam-2]:
name=uniprotkb:Q66SL2

Commento [Dkkam-3]:
name=uniprotkb:O7Z419

Commento [Dkkam-4]:
name=uniprotkb:P0Z340

^aDepartment of Cell Biology and Genetics, College of Life Sciences, Peking University, Beijing 100871, China ^bNational Jewish Medical and Research Center, C080206, USA ^cCollege of Life Sciences, Wuhan University, Wuhan 430072, China

Received 26 July 2005; revised 20 September 2005; accepted 20 September 2005 Available online 18 January 2006 Edited by Varda Rotter

Commento [Dkkam-5]:
name=Beijing

Commento [Dkkam-6]:
name=China

Commento [Dkkam-7]:
name=China

4-11 OKKAM4MS Word: OKKAMid inside the document

5. Conclusions and next step

This deliverable contains a guide to the public OKKAM Empowered tool developed for this first suite. The guide reflects the functionality of the APIs as of March 2009 (the contractual date of delivery of this document). As the software evolves, we will consider revising the individual chapters at a later stage of the project.

Next step will be the development of the Second OKKAM Empowered tool suite. Activities will include:

- Working on what spins around the web, from blogging to creating online contents to spread information
- Fine-tuning of the work done, both in terms of architecture and single tools
- Analysis and creation of new tools that may spread in the open source world

6. References

- [1] <http://www.icefaces.org/main/home/>
- [2] <http://java.sun.com/javaee/javaserverfaces/>
- [3] <https://facelets.dev.java.net/>
- [4] <http://jena.sourceforge.net/>
- [5] <http://www.ie7pro.com/>

7. Appendix – Information on the COGITO Technology

7.1. Semantic Analysis

COGITO® is a software system that “understands” the language in way that is very similar to the way people do. It collects all the structural and lexical text aspects in order to understand the meaning.

COGITO® is the result of hundreds of man years of research and development., from the research and extraction, to the analysis, classification and transformation of unstructured information,

COGITO® is the most advanced technology available on the market because it overcomes the more traditional approaches to the automatic processing of natural language by leveraging semantic comprehension.

COGITO® is powered by Sensigrafo®, a comprehensive semantic network available in different languages enabling disambiguation of terms, the secret behind Semantic Intelligence.

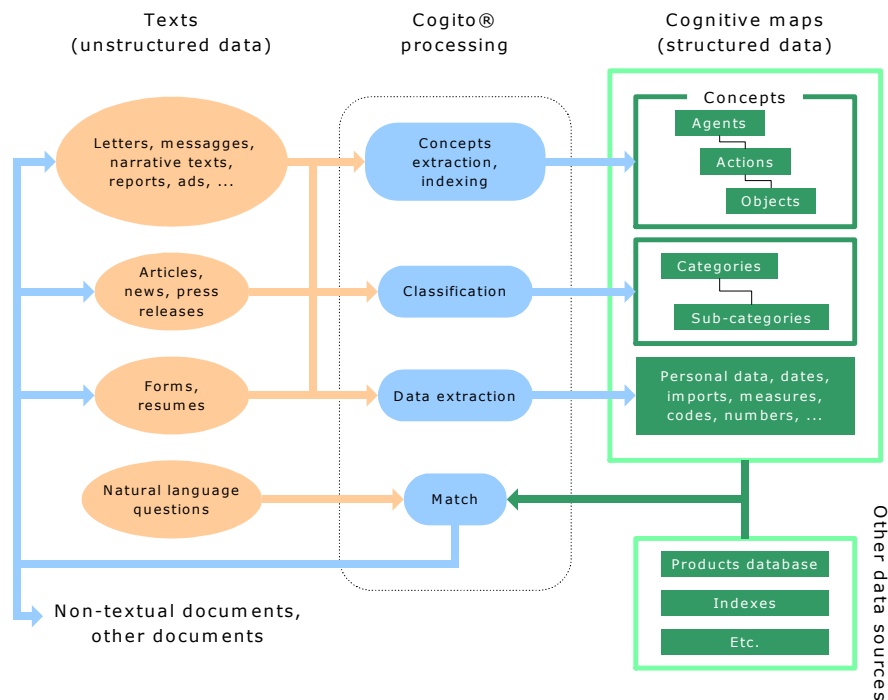
- more than 3,500,000 connections (regarding contests, constructions, subjects and domains, idioms, phraseologies...)
- more than 600,000 semantic concepts
- more than 400,000 hyponymy and hyperonymy connections
- more than 55,000 hyperonymy and troponymy connections
- more than 370,000 connections for the corpus and tens of thousands of links among subjects, objects, meronyms, etc.

7.2. COGITO® Semantic Technology

COGITO® represents our solution. It is the result achieved after 10 years of work in the development of state-of-the-art linguistic technologies.

COGITO® is a software system that “understands” the language in way that is very similar to the way people do. It collects all the structural and lexical text aspects in order to understand the meaning.

COGITO® processing result is a cognitive and conceptual map, i.e. a structured representation of qualifying aspects of incoming unstructured data. The output structuring allows the automatic processing of the most relevant elements of the text.



Rival systems have a superficial processing level. For example, they completely ignore the so-called stop words, i.e. words that do not have a proper meaning (like articles and prepositions), when indeed they are fundamental in forming the meaning of sentences.

On the other hand, COGITO®, with its linguistic analysis capabilities and semantic knowledge, takes into account every word; in the same way people do when they read, listen to and write.

In order to better understand this feature, consider these sentences:

- (a) I've been flying my bird of prey.
- (b) I've been flying my bird as prey.

The meaning of these two sentences is completely different, because of the prepositions. In the sentence (a) the words “bird” and “prey” refer to a raptorial bird - a type of bird, such as an eagle, falcon or hawk, which hunts other birds and small animals - while in the sentence (b) “bird” and “prey” refer to the use of a bird as a decoy, or lure, to facilitate the capture of another animal.

This difference is considered to be clear for a person, but this doesn't happen with current text management programs: from their point of view the two sentences are the same.

COGITO® can detect the difference and distinguish between the two concepts, allowing a conceptual search.

COGITO® has an open and scalable infrastructure, which can be easily customized in order to manage any type of contents: technical documentation, encyclopedic material, financial information, newspaper articles and so on.

The power and the level of integration between its components makes COGITO® a revolutionary technology and very unique in the marketplace. Its main components are:

- parser: it carries out morphological, grammatical and syntactical analysis of the sentence
- lexicon: it recognizes words and all their meanings
- memory: it keeps trace of previous analysis outcomes
- knowledge: a representation of real world knowledge
- content representation: text contents in the form of a conceptual map

7.2.1. Parser

The first step to take in order to understand the meaning of a sentence is to determine the grammar role of each word. For example, in these sentences:

- (a) He aged 40 years.
- (b) The wine is aged.

The word “aged” appears with two different grammar types: in sentence (a) the word is an adjective, while in sentence (b) it is a verb.

According to traditional systems the two words are the same, while the linguistic technologies assign different meaning to them.

Recognizing a word independently of its written form is equally important; nouns and verbs have several forms:

- (a) Marcello Mastroianni was the most popular Italian actor abroad.
- (b) Today it is difficult that no more young actresses play a protagonist role.

In the sentences above, two forms (“actor” and “actresses”) expressing the same concept are used. COGITO® individuates gender - masculine/feminine - and number - singular/plural – to recognize both words as forms of “acting person” associating all of them to their common meanings correctly, instead of individuating different words as other systems do.

COGITO® parser manages the grammar characteristics of the sentence completely and optimally. It carries out the logical analysis, providing a solid base for contents processing.

7.2.2. Lexicon

The information of all the possible meanings of words is fundamental in order to process text contents with high precision.

Being unable to detect different meaning brings a misleading understanding of the phrase. Consider this example:

- (a) The car driver was injured.
- (b) I used the long driver.
- (c) The driver was installed in the computer.

The word “driver” is ambiguous, because its meaning depends on context. In order to carry out word meaning disambiguation, COGITO® looks up its lexicon to find all possible meanings. These lexicons are semantic networks.

Semantic networks are not plain dictionaries, but resources that have been optimized for programming use, where word forms are knots linked to each other by multiple links denoting semantic or lexical relations. For example, the knots “secret agent” and “spy” are linked by a semantic relation named “synonymy” (they have similar meaning), while “angel” and “devil” are linked by “antonymy” (they indicate opposite concepts).

Something impossible for all those systems considering words as simple character strings, not concepts.

Sensigrafo® is an Expert System innovative semantic network. At a high level it has similarities with other existing semantic networks (like WordNet for English language from Princeton University), in that it captures the structure of a language through the definition of concepts and relationships between concepts. However, Sensigrafo® is uniquely designed from the start, implemented and optimized for automatic language processing.

Concepts (like nouns, verbs, adjectives and adverbs) are defined by “synsets”, which are lists of cognitive synonyms, interlinked through semantic and lexical relations.

Some definitions may be useful:

Lemma – A pairing of a particular orthographic form with some form of symbolic meaning representation.

Semantic relations – The relation between meanings of lemmas, such as homonymy, homophones, homographs, synonymy, hyponymy, hypernymy and polysemy.

Syncon – A syncon is a list of lemmas with the same meaning. It represents the semantic relation of synonymy between lemmas.

Sensigrafo® embodies some unique features that allow a superior semantic analysis and disambiguation of text. No identified deep text analysis solution today has these features and is able to achieve similar levels of performance. Among the unique features of Sensigrafo® are:

- Expanded definition of lemmas and semantic relations
- Categories of attributes
- Domains of usage
- Differentiated sets of attributes
- Small footprint of 50MB of memory

7.3. COGITO® Discover

The problem of information management has reached relevant proportions. The intellectual capital (know how, brands, copyrights, patents, projects) and web resources cause a constant increase in the amount of corporate data everyday. Instead of becoming a great opportunity for growth and development, this ironically becomes the main obstacle to the activities of strategy definition, analysis, control, and project management. This happens because quantity and quality often do not necessarily go hand in hand, but above all because a large part of the data is unstructured: more than 80% of the information on which the companies base their business is in fact generally in text form. The management of this vast “sea” of information (reports, in-depth sector analyses, research, meeting proceedings, analyses and balance sheets, articles, press releases, etc.), filed in different places and formats is really complicated. Selecting useful material can be difficult. Often nothing is found, documents already present are duplicated to the detriment of the quality of the results. In addition, the risk of selecting wrong information while ignoring important details or, even worse, not knowing all the data required to make the best possible decision, can result in missing goals and wasting time and money.

COGITO® Discover is capable of extracting relevant information and knowledge from the mass of data, thus making the entire corporate knowledge wealth always available: expertise, best practices, internal and external information, regardless of it being available in structured or unstructured form. The system enables the user to select the data, individuating it in an autonomous way and extracting it automatically from the available documents generally stored in different formats, or available in more than one database, already structured or semi-structured.

Unlike other tools for the automated extraction of data, COGITO® Discover relies on linguistic and semantic technologies to analyze and comprehend the meaning of texts. The ETL process (Extraction, Transformation, Loading) carried out with the linguistic intelligence of COGITO® Discover is, in fact, focused on the normalization of contents. The use of semantics in the management of content permits the tracing back to a univocal form (or “normalizing”, actually) the most simple data such as prices, dates, quantities, measures, etc.,. It also extracts and standardizes non-numerical information such as acronyms, abbreviations, names, and alternative forms to express the same concept.

The combination of the module for the extraction of concepts and of that for the normalization, guarantees a correct disambiguation of the terms (that is, the attribution of the right meaning to the words of the sentence as related to the context) and at the same time an automated selection of the data which is already normalized.

COGITO® Discover can therefore find, in any unstructured text, elements marked as interesting: names, companies or agencies, products and places, in addition to dates, prices, as well as numbers in general. Furthermore, always relying on the evolved linguistic technology of Expert System, COGITO® Discover is able to “understand” the subjects of texts, tracing and isolating automatically the relevant concepts contained in the sentences.