



INSEMTIVES
FP7-ICT-2007-3
Contract no.: 231181
www.insemtives.eu

INSEMTIVES

Deliverable 2.3.1

Requirements for Information Retrieval Methods over Semantic Content

Editor:	Borislav Popov, Ontotext
Deliverable nature:	R
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	30.09.2009
Actual delivery date:	30.09.2009
Version:	1.0
Total number of pages:	22
Keywords:	semantic search, multi-paradigm retrieval, semantic information retrieval, semantic annotation

Abstract

The focus of this deliverable is to provide requirements for the information retrieval (IR) methods over semantic content that need to be developed and adopted within the INSEMTIVES project. These requirements, although drawn from the analysis of the show cases, are heavily based on previous industrial experience of semantic IR and focused on serving a wider range of retrieval tasks over semantic content than the ones covered by the project. We briefly describe the different types of data involved in the semantic annotation and retrieval tasks. Based on these we draw the possible retrieval approaches as well as combinations of them covering more than one data modality. We also define the retrieval needs from the point of view of the different types of metadata involved - on the level of the entire information resource; a part of it (like an image segment, or a span of text); and annotation, or correspondence between the source content and some formally structured data set (like a knowledge base) or a conceptual model (like an ontology). In these requirements we include retrieval methods needed by the end show case applications, the underlying tools, and also from lower level tools that might be used for maintenance and engineering of the captured knowledge.

Disclaimer

This document contains material, which is the copyright of certain INSEMTIVES consortium parties, and may not be reproduced or copied without permission.

All INSEMTIVES consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the INSEMTIVES consortium as a whole, nor a certain party of the INSEMTIVES consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Impressum

INSEMTIVES – Incentives for Semantics

INSEMTIVES

WP2: Models and Methods for the Creation and Usage of Lightweight, Structured Knowledge

Report on Information Retrieval (IR) methods for semantic content - requirements and specification

[Editor: Borislav Popov, Ontotext]

[Work-package leader: Ilya Zaihrayeu, University of Trento]

[Estimation of PM spent on the Deliverable: 6 PM]

Copyright notice

©2009-2012 Participants in project INSEMTIVES

Executive Summary

The current document describes the requirements for semantic information retrieval in the context of the INSEMTIVES project and beyond. The description goes through analysis of the various sets of information resources targeted in the showcases, the requirements for the annotation types and expressivity and goes to exemplifying the retrieval needs through a set of requirements. Finally, the requirements are summarized from the point of view of types of retrieval results, retrieval methods and two approaches to retrieval need definitions: query language syntax and API-based.

The important things to mention here are that the retrieval covers:

- RESULTS OF VARIOUS DATA TYPES , like:
 - web service descriptions
 - textual documents
 - multimedia resources
 - entity descriptions and
 - annotations
- RETRIEVAL METHODS COVERING:
 - structured conceptual queries
 - annotation-centric queries
 - semantic search over annotations
 - similarity-based queries between artifacts of different data modalities (like texts and structured entity descriptions)
 - traversal queries over the topology of the data-scapes of information resources (like linked pages in a CMS)
 - full-text search queries involving conjunction, disjunction, grouping, priority and even proximity and query term weight boosting
 - semantics enabled full-text search
 - co-location of annotations in a given context, like entire information resource or a part of it
 - provenance-related queries
 - and hybrid or multi-paradigm retrieval combining all enlisted retrieval methods into a single retrieval need definition
- QUERY DEFINITION :
 - query language syntax
 - API for defining each of the retrieval methods and their harmonization into multi-paradigm queries

Additionally, we have analyzed the correspondence between the annotation models defined in D2.1.1 and D2.1.2, and the retrieval requirements defined here, specifically focusing on the structural complexity and provenance aspects of the annotations.

As a conclusion, it is important to emphasize that this set of requirements involves ambitious methods for retrieval across different data modalities and combining into multi-paradigm query definitions to serve the needs of a semantic annotation platform.

List of Authors

Company	Author
University of Innsbruck	Tobias Bürger
University of Trento	Ilya Zaihrayeu
University of Trento	Uladzimir Kharkevich
Ontotext	Borislav Popov

List of Figures

1	Annotation highlighting	12
2	Aliases of an Entity	14
3	Structured Query Pattern	14
4	SeRQL for people working for Telefonica	15
5	Organisations in the context of Telefonica	15
6	Annotations Highlighted in Text	17
7	Virtual Worlds Annotation	18

Contents

Executive Summary	3
List of Authors	4
Abbreviations	7
Definitions	8
1 Introduction	9
Motivation	9
Purpose	9
Scope	9
2 Semantic Information Retrieval Requirements	10
Contexts	10
Retrieval Scenarios	11
Requirements Summary	19
3 Correspondence of Retrieval Requirements and Annotation Models	21
4 Conclusions and Outlook	22

Abbreviations

IR Information Retrieval

OWL Web Ontology Language

RDF Resource Description Framework

URI Unified Resource Identifier

CMS Content Management System

FTS full-text search

SPARQL an RDF query language

SeRQL another RDF query language

VSM Vector-Space Model

TF.IDF Term Frequency multiplied by Inverse Document Frequency

Definitions

Annotation The term *annotation* is used both as a noun denoting a piece of additional information and as a verb referring to the process of creating this additional information.

Annotation Model An *annotation model* defines the actual form in which the annotation, that is, additional information is expressed, and how it is linked to the original content being annotated.

Ontology an ontology is considered an explicit specification of a conceptual model

PageRank linkage topology based resource importance ranking algorithm and score, highly visible in Google's implementation of scoring of query results.

HITS algorithm and score similar to PageRank, but focussed on evaluating the local importance of a resource within a given context defined by a query or its set of results and not the entire resource space.

1 Introduction

Motivation

Although the focus of the project is to elaborate on and develop further incentives for semantic annotation (or enrichment) of content, and thus the emphasis is mainly on the motivation of the stakeholders and the annotation models, in all such environments retrieval, be it obvious and user-facing or under the hood, plays a major role. The retrieval methods are providing the powerful selection mechanisms over content and metadata to abstract what data is need for manipulation or presentation purposes. Beyond the context of the project, the semantic IR methods will play an increasingly significant role, as besides the information overload becoming evident in every walk of life, content is increasingly being enriched more and more with metadata of various types. Using this metadata and the corresponding conceptual models and background knowledge has the potential to dramatically improve the exploratory experiences building on traditional IR and content topology (methods like PageRank and HITS).

Purpose

To define the requirements for semantic IR in the context of the project, but also incorporate industrial and research experience of the partners and insights on future usage of the technology.

Scope

This deliverable elaborates on the topics of semantic information retrieval in the context of semantically annotated content and existing background knowledge. It may touch some content specific types of search (as derivatives of full-text search (FTS)) but is primarily dealing with retrieval techniques with respect to semantic annotations and the corresponding background knowledge.

2 Semantic Information Retrieval Requirements

The requirements of retrieval methods needed in the context of the INSEMTIVES platform need to be understood in the context of the types of information involved in a semantic annotation application. As seen from the requirements gathered from the showcases on annotation models (INSEMTIVES D2.1.1) and through internal meetings with the partners responsible for the showcases and the toolkit (WP4), it is seen that the actual applications of these retrieval methods will involve various levels of expressivity of the annotations and the background knowledge, as well as content of various modalities. It is important to understand what is the appropriate level of meta-data modeling in the platform in order to serve these varying needs. We can identify three different types of data involved:

- content or primary information resources, including:
 - textual documents of various kinds
 - multimedia content, images, videos, and interactive multimedia like games
 - web service descriptions
- background knowledge, including:
 - thesauri, term lists or glossaries
 - conceptual models or ontologies (usually domain specific, but with open-domain aspects as well)
 - knowledge bases (e.g. with descriptions of company products or services; user profiles)
- annotations, linking the primary information resources or their parts, to the background knowledge, involving various levels of semantics, ranging from free-style tags with lexical representation, to direct links to entities or facts in the knowledge base.

Another perspective to look at the retrieval methods is from the type of their result. Generally we can distinguish:

- content-centric, in which the actual need is focused on the retrieval of the original information resource
- knowledge-centric, in which the needs is focused on retrieving a part of the structured entity and fact descriptions

We will start with describing several examples and further analyze the retrieval requirements for each of these types of data and also elaborate on hybrid retrieval paradigms involving more than one of these modalities.

Contexts

Here we will describe several retrieval examples in the context of our show cases, as well as based on our previous experience of semantic annotation environments. We will use the TID (Telefonica) show case as a primary context and only highlight the content-specific retrieval methods for the virtual worlds and web services use cases.

In the following pages, we will show retrieval scenarios in order to exemplify the retrieval requirements for the platform. The scenarios will be mainly based on interlinked content artifacts - usually textual and unstructured. Additionally, we will provide examples of multimedia retrieval to emphasize the added complexity in the annotation definitions and hence - in the definition of the information need during retrieval.

- CONTEXT: SEMANTICALLY ANNOTATED TEXTUAL DOCUMENTS

An example of this context is the Telefonica use case, in which we have a content management system (CMS) with a lot of textual content, its own topology of in- and out-bound links and some metadata. In the future the metadata layer over the content will be enriched with semantic annotations supported by a knowledge base according to a Telefonica specific conceptual model. The origin of these annotations might be both the direct contribution of users of the portal and purely automatic means. To give more shape to the example, let us focus on the semantic annotation of specific terminology and associated activities in the company, or internal projects. In this case we will have:

- texts, or parts of them associated with a term or activity through an annotation
- some representation of the original information resource (or page from the CMS)
- knowledge base with dynamic contents describing the documents, annotations to them and the actual terminology or activity within the company (i.e. the domain entities and facts)

- CONTEXT: SEMANTICALLY ANNOTATED MULTIMEDIA

Our show case with Pepper'S Ghost Productions (PGP) goes under the name of VIRTUAL WORLDS and will be an example of multimedia annotation covering various types of artifacts, like:

- IMAGES
- VIDEOS
- 3D VIRTUAL ENVIRONMENTS
- GAMES

This scenario will need the same retrieval capabilities as for the text data modality, but further complicated with content-specific characteristics of the annotations.

- CONTEXT: ANNOTATED WEB SERVICES

The nature of web service descriptions is such, that they come in structured form and are not of the same type as the unstructured content streams presented so far, although in their descriptions there is plenty of room for ambiguity as well.

Retrieval Scenarios

- GETTING THE ANNOTATIONS OF A PAGE (OR ENTITIES MENTIONED)

This retrieval scenario takes a document as a starting reference point and retrieves all annotations or linked entities as a result. There is the option to define different type of sorting techniques based on some defined rank or meta-data attributes having order (like dates or age). This retrieval can also be used for annotation highlighting and navigation-enabling front-ends like on the following diagram (Figure 1)).

- GETTING THE ENTITIES OR DOCUMENTS FROM ALL OUT- OR IN-BOUND LINKED PAGES WITHIN A GIVEN HOP DISTANCE

An extension of the previous scenario, is the case in which we are interested in obtaining entities or annotations associated with more than one resource. In the case of having interlinked pages, as in a CMS

Vodafone and Telefonica in network deal
 Mon Mar 23, 2009 11:56am GMT
 By Georgina Prodhan and Sarah Morris

LONDON/MADRID (Reuters) - Vodafone and Telefonica

agreed to share network infrastructure in four European countries to meet a surge in demand for mobile broadband, while saving hundreds of millions of pounds in costs.

The deal announced on Monday, the biggest of its kind to cover multiple countries, is a sign of the urgency to save money and the success of flat-rate data packages in stimulating demand as well as a more relaxed attitude towards equipment ownership.

The deal covers Germany, Spain, Ireland and Britain and may be extended to the Czech Republic, the two companies said. They will share sites, equipment including masts, and power supply, but will keep their own radio equipment and vendors.

"It's a real transformational deal," Michel Combes, chief executive of Vodafone Europe, told journalists on a conference call. Matthew Key, his counterpart at Telefonica, said: "The current economic situation was a catalyst."

Ireland, Spain and Britain were the first of Europe's developed economies to feel the impact of the economic downturn that has since spread to most of the world.

Telefonica's sales in its home country of Spain, grew more slowly than any of its other major markets last year, while Vodafone cut its outlook twice in 2008, citing challenging conditions in Europe.

Telecoms carriers are also increasingly deciding that they do not necessarily need to own and maintain their own network any longer to provide the best service.? Continued...

Figure 1: Annotation highlighting

or the Web, we can traverse the topology of this linkage and obtain related pages and their annotations.

- GETTING ENTITIES MENTIONED IN A SET OF DOCUMENTS

A more general example of this is to have any set of documents coming as a result of a previous retrieval method and obtain the mentioned entities. Besides possible sorting and orders, these entities can also be grouped by their entity type at the appropriate granularity level in the conceptual model. Like being interested in people, without any distinction if they are men or women, although this information might be present as more specific entity types.

- GETTING THE ANNOTATIONS RANKED BY FREQUENCY OF OCCURRENCE

When dealing with more entities, it is useful to have ranking schemes, emphasizing more important entities in the result set - the traditional being - most frequently mentioned in the set, with a possible extension to a TF.IDF like calculation, where the mentions of the entity in the set will be normalized with respect to all its mentions in the entire corpus of documents. Thus, one can also measure *keyness* or *significance* of the entity annotation for the given information resource. This is usually a computationally intensive task based on accumulative calculations based on the links between a resource and an annotation and further to an actual entity instance in the knowledge base. It is important to have the flexibility to change the ranking function in order to suit various and unforeseen needs.

- SEMANTIC SEARCH OVER TAG/ATTRIBUTE/RELATION ANNOTATIONS

In this retrieval scenario, entities, concepts, and relations from background knowledge are used in order to retrieve resources with semantically similar tag/attribute/relation annotations. For instance, if the user is searching for pictures of animals all the pictures tagged as dogs and cats should also be retrieved. As another example let us consider a situation when the user is searching for all the restaurants located in Trento. In this case, all the restaurants with relation annotation <located in, Trento> should be found. Moreover, the restaurants with semantically related relation annotations, e.g., <placed in, Trento>, might also be relevant to the user needs and therefore should also be retrieved.

- GETTING DOCUMENTS MATCHING AN FTS QUERY

As the content of these scenarios is un- or semi-structured text, naturally, humans interacting with the end-products based on the platform will expect keyword-based retrieval. Despite the fact that the system

itself and its retrieval are heavily based on structured data, the expectation is that an extensive set of FTS and Boolean operators will be supported including:

- conjunction
- disjunction
- grouping
- priority
- document meta-data field restrictions and even
- query term boosts
- proximity search

Based on the language or other characteristics of the content there might appear the need for additional processing both during indexing and query execution, like special tokenizers to identify indexed terms, or morphological analysis, to reduce the dimensions of the index space and ensure term hits despite the actual word forms and inflections. All these, of course are not in the scope of the platform development, but need to influence the architecture design, and will have their instance at least in the Telefonica show case.

- GETTING DOCUMENTS BY USING SEMANTICS ENABLED FULL-TEXT SEARCH

If documents and user information needs are represented in a natural language, then the quality of results returned by FTS can be affected by the problems related to the ambiguity of natural language, namely, the problems of polysemy (the same word have multiple meanings) and synonymy (different words have the same meaning). Moreover, FTS does not take into account (complex) concepts which are semantically related to the query (complex) concepts. For instance, a document about *black dogs and white cats* can be retrieved by FTS as an answer to the query *black cat*, because both words *black* and *dog* appear in the document, while semantically the query and the document are not related. At the same time, the document about *black dogs* will not be retrieved by FTS as an answer to the query *black animal*, because word *dog* does not appear in the document. To solve the described above problems we need to extend FTS by moving from words, expressed in a natural language, to concepts, expressed in an unambiguous formal language and by extending syntactic matching of words to semantic matching of (checking subsumption between) complex concepts.

- GETTING DOCUMENTS MATCHING AN FTS QUERY AND CONTAINING REFERENCES TO ENTITIES

Combining a keyword based query with a restriction over an entity description is a very powerful hybrid search paradigm, which brings together the best of two worlds: from one side the traditional IR approach and structured queries expressing restrictions on the entity profiles or annotations. There are two approaches to this and we would need to support both.

An example would be asking for China's capital and related reports. One can specify the capital through its relationship to the country, or through its name. The former will be possible if there is information for the relationship:

COUNTRY.CHINA HASCAPITAL CITY.BEIJING

and the latter, if there are lexical aliases associated with the capital in the knowledge base, like in the following example (Figure 2):



Figure 2: Aliases of an Entity

As *Beijing* appears with various names, (e.g. *Peking*), the benefit of combining an entity description-based restriction and a FTS query are obvious: We will be able to obtain documents containing *report* and any of the names used for the Chinese capital.

- GETTING ENTITIES THROUGH A RESTRICTION ON THEIR DESCRIPTION/PROFILE

As in INSEMTIVES, the modeling of the annotations and the background knowledge will be based on RDF and OWL, it is natural the lower level interaction with the back-end to be possible through SPARQL. Although exposing full-power access this is not the appropriate abstraction level for end-users, and often for tool developers. For this purpose a set of higher level APIs need to be defined capable of defining structural queries and name restrictions.

To perceive the gap of what goes on under the hood and what the user needs to see, we will show an example query for:

PEOPLE, *having* POSITIONS *in* ORGANIZATIONS, *which name starts with* TELEFONICA.

On Figure 3 we can see such a pattern, where the graph pattern is pre-defined and the human only needs to define the aliases.

Figure 3: Structured Query Pattern

The corresponding SeRQL query with respect to a knowledge base modeled against a basic-upper level ontology (PROTON¹) looks quite scary on the next Figure 4:

¹PROTON - <http://proton.semanticweb.org/>

```

select distinct X,XMainLabel,Y,YMainLabel,Z,ZMainLabel
FROM {X} <http://proton.semanticweb.org/2006/05/protont#hasPosition> {Y},
{Y} <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> {<http://proton.semanticweb.org/2006/05/protont#JobPosition>},
{X} <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> {<http://proton.semanticweb.org/2006/05/protont#Person>},
{Y} <http://proton.semanticweb.org/2006/05/protont#withinOrganization> {Z},
{Z} <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> {<http://proton.semanticweb.org/2006/05/protont#Organization>},
{Y} <http://proton.semanticweb.org/2006/05/protont#hasMainAlias> {} <http://www.w3.org/2000/01/rdf-schema#label> { YMainLabel },
{X} <http://proton.semanticweb.org/2006/05/protont#hasMainAlias> {} <http://www.w3.org/2000/01/rdf-schema#label> { XMainLabel },
{Z} <http://proton.semanticweb.org/2006/05/protont#hasMainAlias> {} <http://www.w3.org/2000/01/rdf-schema#label> { ZMainLabel },
{Z} <http://proton.semanticweb.org/2006/05/protont#hasAlias> {} <http://www.w3.org/2000/01/rdf-schema#label> {ZLabel0}
where ZLabel0 like "Telefonica*" ignore case|

```

Figure 4: SeRQL for people working for Telefonica

- GETTING ENTITIES THAT ARE CO-LOCATED IN A CONTEXT (DOCUMENT, SECTION, PARAGRAPH, SENTENCE, TABLE) WITH ANOTHER ENTITY

An important exploratory and navigation tool is the observation of *colocation* or *co-occurrence* of entities in a context. Entities co-occurring with a given one can provide insights additional insights on the characteristics of the entity and even form a part of its profile. Such context models are widely applicable for instance-level disambiguation or identity resolution, in combination with the structured entity descriptions or term-based (as opposed to entity-based) contexts, usually of higher dimensionality. An example can be seen on Figure ??, where we can observe:

Organizations occurring together in news articles containing Telefonica sorted by popularity in this context.

Vodafone Group, PLC
 Telefonica
 Reuters
 Apple Inc.
 MarketWatch, Inc.
 Symbian Foundation
 Nokia Corp.
 AT&T Inc
 Daily Telegraph
 British Telecommunication...
 Deutsche Bank
 Motorola, Inc.
 In Motion Ltd.
 Softbank Mobile Corp.
 Parliament
 IBM
 Morgan Stanley Group
 Reuters Group PLC
 Sun Microsystems, Inc.
 Toshiba Corporation

Figure 5: Organisations in the context of Telefonica

Depending on the types of the documents, the meaningful entity and term colocation context vary. For example, news articles are usually short, and it is acceptable to take the entire textual content as the context of colocation. Larger documents, like books or patent applications, usually present a compound context. In such cases, the document structure can be used to identify separate contexts and even concentric contexts can be defined on the different levels of granularity of looking at the content.

- GETTING DOCUMENTS OR CONTAINED ENTITIES THROUGH A COMBINATION OF FTS, ENTITY PROFILE RESTRICTION AND COLOCATION WITH OTHER ENTITIES, I.E. EMPLOYING ALL RETRIEVAL PARADIGMS AT ONCE

Having data of multiple modalities and exposing retrieval methods on each of these, it is highly recommended to be able to expose hybrid, or multi-paradigm retrieval methods. These will combine the information need definition capabilities for each of the individual modalities and apply them all to select the result set of documents or entities. It is expected that there will be a query syntax capable of encompassing the hybrid retrieval need definitions. Optionally, there might be an API for defining such hybrid queries by a sequence of method calls.

Example:

Documents containing the keyword FRAUD co-located with a COMPANY
active in the PHARMACEUTICAL INDUSTRY

- GETTING DOCUMENTS FROM ENTITIES, WITHOUT AN EXPLICIT LINK/ANNOTATION BETWEEN THEM

All the methods described so far, depend on the explicit association of the original information resource with some structured meta-data through an annotation. In many cases, this will not be present and we need to define some types of similarity in order to be able to associate artifacts in our data-scape. If these artifacts are unstructured information resources, like images or text documents, then there are content specific methods for doing this, e.g. cosine measure between document vectors in a vector-space model (VSM).

It will be beneficial if we see retrieval methods which are based on some similarity measure between artifacts of different modality. For example to be able to find documents which are relevant to an entity, without the explicit link between them:

NEWS relevant to TELEFONICA

This could be handled by a sort of normalization of the different data modalities. In this case we have structured data defining the entity TELEFONICA, and unstructured text for each news article. Through processes of complicated text analysis one can extract structured data from texts, like entity mentions or relationships. However, this is not generally applicable and it is more natural to use an approach similar to LEAST COMMON MULTIPLE in *Number Theory*. In the case of the example, this means to *render* the description of the entity TELEFONICA to a textual form.

If we imagine that the description of Telefonica in the knowledge base consists of the classes it is instance of, like:

COMPANY, ORGANIZATION, AGENT, ENTITY,

and its various names, like:

TELEFONICA, TELEFONICA S.A.

One can continue rendering everything known in the knowledge base about this entity, like:

its brands:

TELEFONICA, MOVISTAR, O2, TERRA

locations where it is active, with all their names:

SPAIN, EUROPE, UK, UNITED KINGDOM, BRAZIL, CHILE, ...

and areas of activity like:

TELECOMMUNICATIONS, MOBILE SERVICES, ADSL, ...

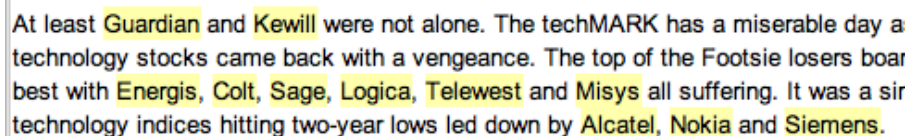
and covering in this way the entire entity profile, we will get a reasonable amount of textual content, to be used to identify similar news articles. In this case the rendered entity profile can be perceived as a vector into a vector-space model and cosine similarity and compound TF.IDF measure can be computed.

It is expected that with this approach we will be able to obtain documents relevant to an entity, even in the cases of no explicit mentions of the entity names in the text.

- ANNOTATION PATTERNS

The examples of retrieval methods described so far mostly rely on the *explicit presence of an annotation* linking a part of the content with a knowledge base entity. On the other hand the retrieval methods described rely only on the presence of the annotation *without taking into account its characteristics*. Our analysis of the applicable annotation models (INSEMTIVES D2.1.1) shows that the annotations will have various additional characteristics, the most prominent of which are related to the describing the part or section of the original information source which is being annotated.

For textual content, this is usually an indication of the offset from the start of the document. Most often this is expressed as START OFFSET and END OFFSET , selecting a particular interval in the text. Offset based retrieval, can be used directly to obtain the text covered or the contained annotations. Let us take the annotated text on the next Figure 6.



At least Guardian and Kewill were not alone. The techMARK has a miserable day a technology stocks came back with a vengeance. The top of the Footsie losers boar best with Energis, Colt, Sage, Logica, Telewest and Misys all suffering. It was a sir technology indices hitting two-year lows led down by Alcatel, Nokia and Siemens.

Figure 6: Annotations Highlighted in Text

By using offset-based query, we should be able to request for:

- text covered
- all annotations in the content window or offset interval

- annotations of a particular type, say, ORGANIZATIONS

Similarly, over multimedia content, there are specific *offset* characteristics, that might include a *geometric shape* covering a segment of the image or a video stream and also *time offset* from the start of a movie, a particular milestone in a game, or from another marker (like the beginning of the news block in a TV stream). An example can be seen on Figure 7, where a particular segment of the screen has been annotated as being the quiz section of the given virtual world from the corresponding INSEMTIVES showcase.



Figure 7: Virtual Worlds Annotation

In this case the retrieval by the offset attributes, can again return:

- the corresponding content segment
- annotations present in the segment
- actions that are associated with (a part of) the segment (like entering the quiz)

The co-positioning of annotations in streams (like textual documents) is also extremely promising, as one can define patterns of annotations, and request the matches from the set of annotated artifacts. An example is to be able to specify a query like:

TIMEINTERVAL ORGANIZATION *reported loss of* AMOUNT

and match annotated texts like:

In Q3 Vinaship reported loss of over 15 billion dong

- PROVENANCE

In all the scenarios in INSEMTIVES, as well as in our industrial applications of semantic annotation, there is the need for extensive tracking of the origins of the annotations, or the corresponding artifacts in the knowledge base. The reasons for this can be based on some attempts to evaluate:

- the authenticity of an annotation
- some temporal aspect of its creation, like time or a particular phase of a process
- the creator - an automatic process or a particular individual, and so on

Although this need for provenance is generally the same, it varies from application to application and needs to be handled in a generic way in the platform and just on higher levels of abstraction, have dedicated APIs for association with particular characteristics important for the domain, like: users, game phases, time stamps, etc.

Examples of provenance related retrieval could be:

- KIDS who annotated games related to the SOLAR SYSTEM
- Usage statistics for a web service, or who used it when, grouped by some criteria and given temporal granularity
- Most active CONTRIBUTORS of annotations to the pages from category INTERNAL PROCEDURES

Requirements Summary

Based on the retrieval scenarios described in the previous section, we have identified the requirements for the retrieval methods to be supported in the INSEMTIVES Platform. It is important to mention that, although it is natural for the platform to ignore the content-specific types of search and work only on top of meta-data, annotations and links to the background knowledge, text has become a commodity and it is present everywhere. For this reason we chose to include as a general requirement, the support of traditional information retrieval methods as FTS alongside the structured and semantic ones.

If you have read the examples, it becomes clear that we need everything which is possible with the types of data we expect to have in the context of our scenarios, which are almost orthogonal. Below we will enumerate the requirements coming from the scenarios above, but we can summarize them like this:

RETRIEVAL OF ALL TYPES OF INFORMATION ARTIFACTS, THROUGH INFORMATION NEEDS DEFINED BY USING STRUCTURED CONCEPTUAL QUERIES OVER THE BACKGROUND KNOWLEDGE, ANNOTATION PATTERNS, COLOCATION AND FULL-TEXT SEARCH.

The types of RETRIEVAL RESULTS we expect from the system are:

- documents or pages
- web service description
- multimedia resources
- entity descriptions
- annotations and their attributes
- provenance information for the annotations
- arbitrary resource sets from the background knowledge
- sets of co-occurring entities/annotations
- entities and information resources ranked and sorted by some criteria

The types of *Retrieval Methods* expected from the platform in most cases can return as result most of the retrieval results listed above, so this will not be explicitly described in the following RETRIEVAL METHODS REQUIREMENTS :

- conceptual queries over the background knowledge (domain knowledge bases and conceptual models)
- queries over the annotations, including their attributes and links to entities or classes in the knowledge base

- queries over the explicit links between annotations/entities and information resources
- semantic search over annotations
- traversal queries on the topology of the content-resources or their link-scapes
- full-text search queries supporting IR engine commodity operators like conjunction, disjunction, grouping, priority, and optionally, query term boost, and proximity search
- semantics enabled full-text search
- queries for colocation of entities in the same context (see the scenarios for elaboration on contexts)
- using document or entity as a query to obtain similar resources of another type without an explicit link, but based on another similarity measure (like the one explained in the scenarios)
- queries defining patterns over the annotations of a resource (mostly applicable to text, but meaningful to investigate time-based annotations in multimedia as well)
- queries based on provenance information (in the envisaged modeling approach it is likely that this is the same as the structured conceptual queries over the knowledge base, but for some types of provenance-related operations it is possible to use another type of database and for this reason it is meaningful to have this requirement explicitly specified)
- hybrid or multi-paradigm queries combining any of the above

Beside the requirements for the retrieval methods and their results, it is important to specify the methods for defining the information need for these. As the retrieval is over different types of data it is not trivial to combine the index-specific types of query syntax into one. It is advisable to use the query language used for the majority of the queries, in this case most likely SPARQL, and extend it with specific predicates for the definition of FTS queries, annotation patterns and colocation (although in certain annotation models these can be explicitly expressed as structures in the knowledge base).

Besides the query language syntax support, the platform needs to provide a set of APIs for defining multi-paradigm queries, hiding the complexities of the underlying models and ensuring faster toolkit and front-end development

Definitions of the RETRIEVAL NEED:

- SPARQL, and extended SPARQL
- API for defining multi-paradigm queries

3 Correspondence of Retrieval Requirements and Annotation Models

In order to evaluate the feasibility of these requirements and ensure in an early stage that they can be met in the implementation of the platform and the tools, it is necessary to look at where the elements used in the retrieval come from. We have already described the various content modalities, but the actual models for capturing annotations need more attention. Our analysis of the requirements for the annotation models and the specification of the annotation models have been covered in INSEMTIVES deliverables D2.1.1 and D2.1.2. There we elaborate on the various elements an annotation should cover in the context of the project as well as separate elements of these.

As we can distinguish between content, annotation and background knowledge -based retrieval (or a combination of these), then it is also worth of looking at the correspondence between the annotation-based retrieval and the actual annotation models in order to verify the consistency of the retrieval requirements. From the different characteristics of the annotations described in the beginning of section 2 in D2.1.2 as structural complexity, vocabulary type, provenance and versioning, and access, we need to focus only on the structural complexity and the provenance to ensure the alignment between the annotation models and the retrieval paradigms.

The structural complexity characteristic covers two aspects - the expressivity of the background knowledge being linked to, ranging from plain tags to ontologies; and the granularity of the annotation reference from the point of view of the annotated resource - if the entire resource is being annotated or a given part of it. Starting from the latter, from D2.1.2 is obvious that the granularity aspect has been elaborated upon in the definitions of annotation model objects with index 19, 20 and 21 caring for the definition of textual, image and video segments. This level of resource segment modeling is going to be sufficient for the implementation of the exemplified retrieval scenarios.

Regarding the complexity of the background knowledge used for annotation, at the requirements level, it seems that the implementation will need to cover the highest complexity model (ontology based) as a generic one and accommodate the simpler variations within it, optionally disclosing or not the modeling complexity to the client applications. If this approach is taken at the implementation stage, all the retrieval scenarios will be feasible and will be based on annotations generated with any of the annotation objects defined D2.1.2.

With respect to provenance of the annotations, the scenarios we have given here are accumulative, and are not based on a single provenance information record. On the other hand, the defined History element in the annotation models of D2.1.2 is going to provide the individual information entries that will be used for accumulative queries supporting provenance. Additionally, if the history element's attributes like "action", can find their value sets in an ontology, which will improve the flexibility of the models and lower the cost for adapting them to the use cases.

4 Conclusions and Outlook

The semantic information retrieval requirements deliverable starts with explaining the complicated environment in terms of data-scapes and requirements for annotation of various expressivity levels. We also make use of previous industrial experience with semantic annotation products to ensure the definition of requirements for the retrieval meaningful outside of the INSEMTIVES context.

We chose to present the requirements through highlighting the various retrieval scenarios accompanied by examples and placed in the context in one of the three different types of information resources to be annotated in the project show cases. Based on these we distilled the actual requirements for the retrieval of semantic content. The requirements focus on three different aspects of the retrieval:

- retrieval results
- methods for defining the retrieval need
- methods for interaction with the retrieval layer of the INSEMTIVES platform

Additionally, we have evaluated the feasibility of these retrieval methods with respect to the defined in D2.1.2 annotation models and ensured the compatibility, especially with respect to structural complexity and provenance of the annotations as an object of retrieval need definitions.

Based on these requirements we will produce the specification of the retrieval mechanisms in D2.3.2 and take it into account in the actual development of the INSEMTIVES platform in WP3.