

AsIsKnown

**A semantic-based knowledge flow system for
the European home textiles industry**

Work package 5: Multimedia ontology

Deliverable D18 “Reports on the tests and evaluation”

Lead participant: IPP-BAS
Nature: Report
Dissemination level: PU
Delivery date: 21 PM



This document has been produced in the context of the AsIsKnown Project. The AsIsKnown project is part of the European Community's Sixth Framework Program for research and development and is as such funded by the European Commission. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.



Context

WP 5	Multimedia ontology
Task 5.5	Testing and evaluation
Dependencies	This deliverable requires input from 5.2

Contributors: Kiril Simov (IPP-BAS) Petya Osenova (IPP-BAS)	Reviewers: Kiril Simov (IPP-BAS)
--	--

Approved by: Kiril Simov, Bulgaria as WP5 Head



Executive Summary

The deliverable reported on the first steps of testing and evaluation of the multimedia ontology and NLP annotation to support the Trend Analyser. At this point of time the work is very much connected to the testing and evaluation of the domain ontology. This is why the deliverable is connected to deliverable D11-2. The test of the multimedia ontology will be done with respect to the annotation of the multimedia documents. This task will be done within the third year. The main tests performed up to know are connected with the annotation of the textual part of the magazine articles. The annotation has impact on the search facilities provided to the Trend Analyser. Here we discuss the impact of the context of search. We have defined a hierarchy of contexts in which we could perform the search. This hierarchy includes document, paragraphs, sentences and then n-gram window or syntactic structure context. The last two could be combined with respect to each other and with respect to sentences. This is why we envisage supporting both of them.



Table of Contents

EXECUTIVE SUMMARY	4
TABLE OF CONTENTS	5
LIST OF ABBREVIATIONS	6
1 INTRODUCTION AND PROBLEM STATEMENT	7
2 TUNING OF THE SEARCH RESULTS FOR THE TREND ANALYSER.....	8
3 CONCLUSIONS AND OUTLOOK.....	11
REFERENCES	12



List of Abbreviations

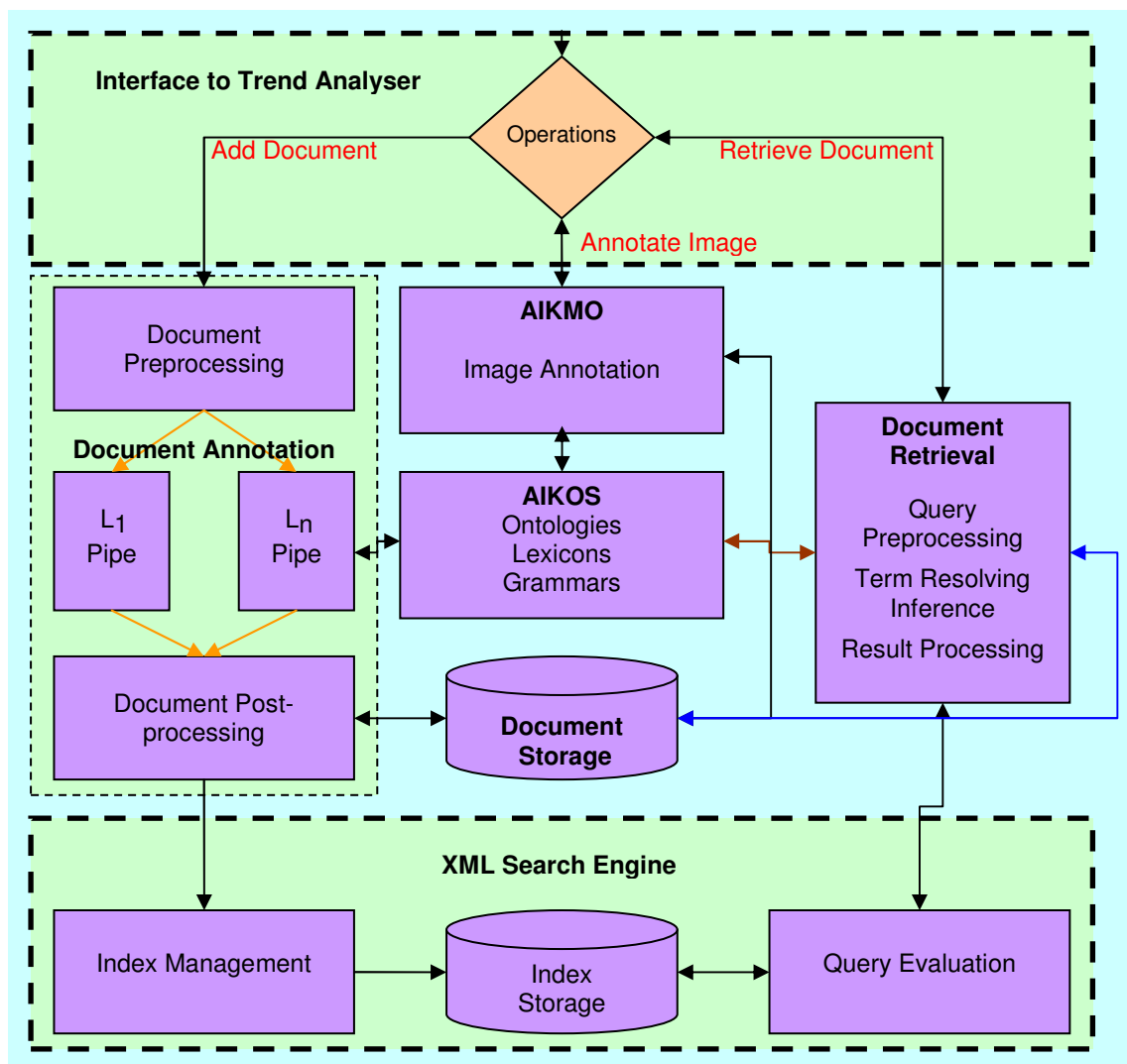
1 Introduction and Problem Statement

In this deliverable we discuss the test and evaluation of the multimedia ontology and the NLP tools developed within WP5. With respect to the ontology the evaluation is planned for the period of the third year of the project as much as the annotation of the pictures within the multimedia documents will be done in this period. The annotation will be done in two ways. First the pictures will be described in natural language with respect to objects depicted within them. On the basis of this description we will prepare a list of objects that are necessary to be annotated. The second step will be to ask the annotators to annotate the objects using the ontology and the annotation tool. Then we will compare the two annotations – the first created on the basis on the textual description of the picture and the second created by the annotation tool. Afterwards we will examine the differences and what the source is of these differences – the multimedia ontology or the annotation tool.

The evaluation and the test of the NLP module has already started. As much as it is used for the semantic annotation of the textual part of the multimedia documents with concepts from the domain ontology, we have presented most of the work done in this respect in the deliverable D11-2: "Reports on the tests and evaluation of common sense ontology" Second version. For that reason here we will report only on very specific WP5 tests and corresponding changes.

2 Tuning of the search results for the Trend Analyser

As it was described in Deliverable D17: "Software tool to design multimedia ontology" the main services which we have to provide are: Add Document, Retrieve Document and Annotate Image. The architecture is depicted in the following picture:



All the tasks are important with respect to the needs of the Trend Analyser. As it was mentioned in the Introduction above, the work done up to now with respect to annotation is reported in Deliverable D11-2, because the main changes followed from these tests and evaluation on the annotation of the textual part of the documents are related to changes in the domain ontology. The tests and the evaluation of the image annotation will be done during the third year of the project. The tests we present here are with respect to the service "Retrieve Document". The general task of this service is to provide access to semantic annotation of the multimedia document within a context. For example, a query could be formulated as: "Give me all the concepts which co-occur with the concept C within a context of type R." The XML Search Engine searches all the contexts of type R, extracts their content. The content is processed in order to deliver it to Trend Analyser. The query in principle could be more complicated than just search for one concept. The initial search

engine supported three types of contexts: documents, paragraphs and sentences. The search engine is implemented on the basis of Lucene (<http://lucene.apache.org/>) - a free (open source) system for full text search, supporting also field based search. The syntactic annotation is done with the help of Tokio parser (<http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/chunkparser/>) and the semantic annotation is done with the help of CLaRK System [1].

The experiments we have done together with the developers of the Trend Analyser have shown that the three types of context are too wide in order to support the tasks for Trend Analyser. This is so, because Trend Analyser is looking for co-occurrences which support relationships between the query concepts and other concepts within the context.

In order to define smaller context we applied two strategies: n-gram window approach and syntactic structure approach.

The n-gram window approach is when the context is reduced to two words before and two words after the query concepts. The good side of such a small context is that the co-occurrences of concepts within it definitely will capture important relationships between the concepts. The negative effect is that in many cases the window around query concepts does not contain concepts, because the words within the window are not terms for concepts in the ontology. This problem is partially solved with the introduction of new concepts to the ontology to cover a bigger domain (see Deliverable D11-2). The other solution we applied is the meaningful words as well as concepts to be returned. The significance of the words is determined on the basis of stopword list.

The syntactic structure approach requires in parallel to semantic annotation we also to have syntactic annotation. We have defined an annotation scheme in which the syntactic structure is represented in a flat manner. This means that we encode the structure as related points in the content of the sentences instead of XML tree. For example, the structure of the sentence "the man saw the boy with the book" has the following tree as a structure (we give only some of the phrases here):

```

<s>
  <np>
    <w>the</w>
    <w>man</w>
  </np>
  <vp>
    <w>saw</w>
    <np>
      <w>the</w>
      <w>boy</w>
      <w>with</w>
      <np>
        <w>the</w>
        <w>book</w>
      </np>
    </np>
  </vp>
</s>

```

We represent this structure in the following way:

```

<s>
  <beg type="np" id="id01"/>
  <w>the</w>
  <w>man</w>
  <end idref="id01"/>
  <beg type="vp" id="id02"/>
  <w>saw</w>
  <beg type="np" id="id03"/>
  <w>the</w>
  <w>boy</w>
  <w>with</w>
  <beg type="np" id="id04"/>
  <w>the</w>
  <w>book</w>
  <end idref="id03"/>
  <end idref="id04"/>
  <end idref="id02"/>
</s>

```

In this way the elements <beg> and <end> could be ignored when the concept annotation grammars are applied and in the same time they could be used for definition of different smaller than sentence contexts. For example, if we search for a concept C within an NP, we proceed in the following way: first, we find the occurrence of C in a sentence; then search the first <beg> element before C in the sentence; after finding this element we search for the closing element <end> for the same NP. In this way we select the smallest NP containing the concept. There are cases when the concept is not in an appropriate phrase, and then we could return the sentence as a context, or to apply n-gram window context. Thus, the good side of this solution is that the context could be defined depending on the structure. For example, we could combine n-gram window context with syntactic structure: if the query concept is in an appropriate context then it is taken as a context; if it is not in an appropriate context, then n-gram window is taken.

At the beginning of the third year we will test these new types of context and if necessary they will be tuned to the needs of the Trend Analyser.

3 Conclusions and Outlook

In this deliverable we report the first steps of testing the NLP annotation for supporting the Trend Analyser. Here we describe the definition of context of search. The other aspects of the tests are reported within deliverable D11-2.

References

- [1] Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, Atanas Kiryakov. *CLaRK - an XML Based System for Corpora Development*. UCREL Technical Paper number 13. Special issue. Proceedings of the Corpus Linguistics 2001 conference, edited by Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja. ISBN 1 86220 107 2. Lancaster University (UK), 29 March - 2 April 2001.