www.axes-project.eu

Project acronym	**AXES**
Project full title	**Access to Audiovisual Archives**
Project No	**269980** **Large-Scale Integrating project (IP)**

Deliverable D6.6

Update on final toolbox for automatic link management and structured search

September 2014

SEVENTH FRAMEWORK PROGRAMME

Objective ICT- 2009.4.1: Digital Libraries and Digital Preservation

PROJECT DELIVERABLE REPORT

Project

Grant Agreement number	*269980*
Project acronym:	*AXES*
Project title:	*Access to Audiovisual Archives*
Funding Scheme:	*Large-Scale Integrating project (IP)*
Date of latest version of Annex I against which the assessment will be made:	*24 September 2010*

Document

Deliverable number:	*D6.6*
Deliverable title:	*Update on final toolbox for automatic link management and structured search*
Contractual Date of Delivery:	*30/09/2014*
Actual Date of Delivery:	*30/09/2014*
Author (s):	*Robin Aly and Kevin McGuinness*
Reviewer (s):	*CAS*
Work package no.:	*WP6*
Work package title:	*Link Management and Structured Search*
Work package leader:	*UT*
Version/Revision:	*1.0*
Draft/Final:	*final*
Total number of pages (including cover):	*13*

CHANGE LOG

Reason for change	Issue	Revision	Date
Initial Version	UT	0.1	05/09/2014
Review by CAS implemented	UT	1.0	25/09/2014

DISCLAIMER

This document contains description of the AXES project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AXES project and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (`http://europa.eu.int/`)

AXES is a project partly funded by the European Union.

Table of Contents

SUMMARY

This document describes update on the WP6 final toolbox (milestone MS19) that were necessary for AXES-Home. More details on the nature and purpose of this software toolbox deliverable can be found in the introduction of deliverable D2.2. The main updates are in the area improved automatic query formulation, support of linking from news-feeds,

INTRODUCTION

Work package WP6 of the AXES project focuses on searching and linking information in multimedia data of large audio-visual archives. This deliverable describes updates to the final toolbox for searching and linking. We refer to the software providing this toolbox as the **Li**nk **Ma**nagement and Structured **S**earch (LIMAS) system. The overall aim of the final software toolbox is to adapt the LIMA's functionality to the requirements for the AXES-Home system. All software components of the LIMAS system have been jointly developed by UT (lead) and DCU.

The functionality described in this deliverable was implemented to satisfy requirements for the AXES-Home user interface formulated by work package WP1. We briefly describe the main requirements as they are used in the remainder of this document:

- Linking of external content: users should be able to find content related to external content, e.g. blogs or news items, in order to explore visual archives without entering queries. For example, users should be able to find relevant archive entries to a particular news item or recipe.
- Automatic search: it should be possible to automatically and transparently translate queries into search requests to several specialized search services in AXES, e.g. face search or speech search. For example, the user query 'Obama enters the White House' should involve face search for 'Obama' and visual instance search for 'White House'.
- Face track display: users should be able to see the faces of persons appearing in a video sequence. Face tracks follow a person through a video sequence and are detected as part of work package WP2.
- Display popular items: users should be presented with a list of popular search topics such that they can choose one for browsing.

The remainder of this deliverable is structured according to the relevant changes to each of the above modules. Section 2 recapitulates the architecture of the LIMAS system for the reader's reference. Section 3 describes the updates for linking external content. Section 4 explains steps taken to support auomatic search. Section 5 elaborates on the functionality to display face tracks, Section 6 describes the functionality to display popular items, and finally, Section 8 summarizes the toolbox report.

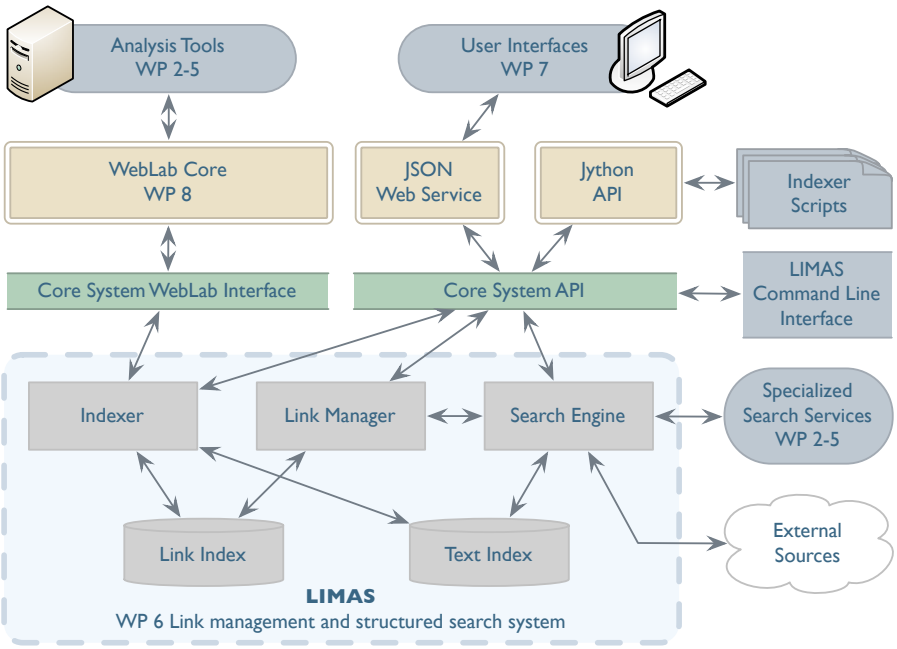

Figure 1: Overall architecture

LIMAS FRAMEWORK

In this section we recapitulate the architecture of LIMAS, shown in Figure 1. The main component of LIMAS is a Java library (limas-core.jar) that contains all core functionality of the system. This library provides a uniform interface to the three types of functionality: indexing, searching, and linking. For each functionality type, LIMAS supports several implementations that can be used in parallel. For example, one implementation of the indexing module creates inverted text indexes and another implementation creates an index for detected entities. The actually used implementations of a functionality are specified in a Python script. Here, the required Java objects are instantiated and connected to the final LIMAS instance. Limas uses Python for the configuration because it is an interpreted language, allowing configuration changes without recompilation. In comparison to the architecture specified in D6.3, the search module in the current architecture calls specialized search services from other work packages via a newly defined protocol, which is described in M6.2.

At the moment, the LIMAS is used in four different contexts:

- two web services to and from the user interface developed in WP7, i.e.:
  - a web service that uses custom defined JSON syntax for the AXES-Pro interface, and
  - a web service that uses the JSON-RPC protocol for the AXES-Research and AXES-Home interface,
- a WebLab web service, which receives output of the video analysis from work packages WP2-WP5 for a video in XML format, which is added to the LIMAS index,
- a command line tool `limas-cmd` that provides a convenient interface for performing maintenance tasks such as creating initial data structures, and,
- a jython (Java implementation of the Python scripting language) interface that provides a mechanism for scripting maintenance tasks in a convenient language.

Linking External Content

Objective In order to allow the user to link from external content, we decided to present him or her with a list of RSS feeds, which is a widely accepted standard for internet content, such as blog posts and news items. From this content, we want to extract a LIMAS query, which is used to retrieve related material from the archive.

Description We created a simple tool to periodically download RSS feeds. For our the AXES system, we selected the following RSS-feeds (with the entries downloaded between May 19th and August 26th):

RSS feed	Language	Entries
All healthy recepies	english	1280
Arseblog	english	103
Art and culture	english	1488
BBC	english	20408
BBC culture	english	81
BBC europe	english	2675
BBC technology	english	1135
CNN	english	8685
DW	english	5639
DW germany	german	1042
Economist europe	english	147
Economist tec	english	114
Economist world	english	63
Financial times europe	english	15
Financial times us	english	15
Guardian sports	english	6421
Guardian technogoly	english	2402
Nytimes educational	english	1123
Reuters	english	3411
Slashdot	english	3195
Smithsonian arts and culture	english	199
Smithsonian music and film	english	27
Taz main	german	1050
Ted.com	english	149
Times	english	32
Volkskrant	dutch	15950

The tool first remove constant text elements, so-called boilerplate text, from each feed entry before it is processed by a state-of-the-art named entity recognition and disambiguation software developed at the UT [1]. The annotated feed entries are then added to the limas index. Afterwards, LIMAS generates a query for each feed entry a query by taking the top-n rarest recognized named entities. LIMAS also creates a query for a whole RSS feed by combining the entities extracted from recent entries of this feed. These queries are then used to search for suitable archive material for this RSS feed.

Automatic Search

Objective We want to support automatic search by detecting mentioning of object categories and person names and using them in category from text and face from text search provided that are provided by WP2 and WP3.

Description To find mentions of object categories, we build a trie data structure consisting of all synonyms of a concept from the linguistic database WordNet [2] (except of person concepts). The trie allows fast identification of all object categories in a query, including those with light spelling mistakes. For person names, we index a large number of first and last names in the same trie. If LIMAS finds a first name followed by a last name, possibly separated by an initial. The query is then generated by searching all text components in the textual metadata and speech content while the found object categories and persons are searched using specialized search services.

Displaying Face Tracks

Objective We allow user interface to browse through face tracks by adding this information to the LIMAS index and allowing the user interface to retrieve this information.

Description We defined an exchange file format together with work package WP3, which is creating these tracks. Files of the format are loaded into the LIMAS index after the stage where shot information has been added. We then added the getFacetracks() method to the LIMAS interface to allow users to retrieve the face tracks for a given video.

Popular Items

Objective Because of the lack of long term user data, we limit ourselves to display few concepts from a known vocabulary. We decided to display the concepts with detection scores of a ueber classifier from LEUV because those are likely to give good results.

Description The method call getInterestingTopics() returns a list of concepts. The concepts are selected among those that were used most before for searching. We therefore hope that the detector scores of these concepts performs well.

License and Dependencies

The LIMAS software is licensed under the Apache License 2.0.

Dependencies The dependencies of LIMAS have the following licenses.

```
Apache Software License, Version 1.1
 sslext (no url defined)

BSD 3-Clause License
  Java WordNet Library (http://jwordnet.sourceforge.net/)

CDDL + GPLv2 with classpath exception
  Expression Language API 2.2 (http://uel.java.net)
  Expression Language Implementation (http://uel.java.net)
  JSP implementation (http://jsp.java.net)
  JSP implementation (https://jsp.dev.java.net/jsp-impl)
  JavaServer Pages(http://jsp.java.net)
  javax.servlet API v.3.0 (http://jcp.org/en/jsr/detail?id=315)

Common Public License Version 1.0
```

```
  JUnit (http://junit.org)

LGPL 2.1 and MPL 1.1
  Javassist (http://www.javassist.org/)

MIT License
  textrazor (https://textrazor.com)

Apache Software License - Version 2.0 and Eclipse Public License - Version 1.0
  Jetty :: Continuation (http://www.eclipse.org/jetty/jetty-continuation)
  Jetty :: Http Utility (http://www.eclipse.org/jetty/jetty-http)
  Jetty :: IO Utility (http://www.eclipse.org/jetty/jetty-io)
  Jetty :: Security (http://www.eclipse.org/jetty/jetty-security)
  Jetty :: Server Core (http://www.eclipse.org/jetty/jetty-server)
  Jetty :: Servlet Handling (http://www.eclipse.org/jetty/jetty-servlet)
  Jetty :: Utilities (http://www.eclipse.org/jetty/jetty-util)
  Jetty :: Webapp Application Support (http://www.eclipse.org/jetty/jetty-webapp)
  Jetty :: XML utilities (http://www.eclipse.org/jetty/jetty-xml)
  Jetty Orbit :: Servlet API (http://www.eclipse.org/jetty/jetty-orbit/javax.servlet)

The Apache Software License, Version 2.0
  Lucene Common Analyzers (http://lucene.apache.org/java/lucene-parent/lucene-analyzers)
  Lucene Core (http://lucene.apache.org/java/lucene-parent/lucene-core)
  MongoJack (http://mongojack.org)
  Apache Log4j (http://logging.apache.org/log4j/1.2/)
  Apache OpenNLP Maxent (http://www.apache.org/opennlp-maxent/)
  Apache OpenNLP Tools (http://www.apache.org/opennlp-tools/)
  Apache Velocity (http://velocity.apache.org/engine/devel/)
  CLI (http://jakarta.apache.org/commons/cli/)
  Commons BeanUtils (http://commons.apache.org/beanutils/)
  Commons Chain (http://jakarta.apache.org/commons/${pom.artifactId.substring(8)}/)
  Commons Collections (http://commons.apache.org/collections/)
  Commons IO (http://commons.apache.org/io/)
  Commons Lang (http://commons.apache.org/lang/)
  Data Mapper for Jackson (http://jackson.codehaus.org)
  Digester (http://jakarta.apache.org/commons/digester/)
  FindBugs-jsr305 (http://findbugs.sourceforge.net/)
  Gson (http://code.google.com/p/google-gson/)
  Guava: Google Core Libraries for Java (http://code.google.com/p/guava-libraries/guava)
  HBase (http://hbase.apache.org)
  JSON Small and Fast Parser (http://json-smart/)
  JSON-RPC 2.0 Base (http://software.dzhuvinov.com/json-rpc-2.0-base.html)
  JSON-RPC 2.0 Client (http://software.dzhuvinov.com/json-rpc-2.0-client.html)
  Jackson (http://jackson.codehaus.org)
  Jackson-annotations (http://wiki.fasterxml.com/JacksonHome)
  Jackson-core (http://wiki.fasterxml.com/JacksonHome)
  Logging (http://jakarta.apache.org/commons/logging/)
  MongoDB Java Driver (http://www.mongodb.org)
  Struts Core (http://struts.apache.org)
  Struts Taglib (http://struts.apache.org)
  Struts Tiles (http://struts.apache.org)
  Validator (http://jakarta.apache.org/commons/${pom.artifactId.substring(8)}/)
  VelocityTools (http://velocity.apache.org/tools/devel/)
  XML Commons External Components XML APIs (http://xml.apache.org/commons/#external)
```

```
  bson4jackson (http://www.michel-kraemer.com)
  jackson-databind (http://wiki.fasterxml.com/JacksonHome)
TagSoup (http://home.ccil.org/~cowan/XML/tagsoup/)

The Jython License
 Jythone Standalone 2.5.2 (http://www.jython.org/)

WTFPL
 Reflections (http://code.google.com/p/reflections/)

Provided without support or warranty
 JSON (http://www.json.org/java/index.html)

Unknown license
  JSON-RPC for java (http://code.google.com/p/jsonrpc4j/)
  Java Portlet Specification V2.0 (http://www.jcp.org/en/jsr/detail?id=286)
  antlr (no url defined)
  dom4j (http://dom4j.org)
  hadoop-core (no url defined)
  jyson (no url defined)
  limas-core (http://www.axes.eu)
  lingpipe (no url defined)
  oro (no url defined)
  persistence-api (no url defined)
  zookeeper (no url defined)
```

Conclusions

This deliverable presented updates to the final software toolbox for automatic link management and structured search as described in D6.5. The updates mainly concerned four requirements stated by WP1 for the AXES-Home interface: external linking, automatic search, display of face tracks, and search for popular items (see Section 1 for an explanation).

In summary, the developments in WP6 are on schedule. This is evidenced by both the systems developed for our fourth year of participation in TRECVid, our third year participation on the MediaEval workshop 2014, and by the provision of a working prototype for various user studies.

References

[1] M. Badieh Habib Morgan. *Named entity extraction and disambiguation for informal text: the missing link*. PhD thesis, Enschede, May 2014. URL `http://doc.utwente.nl/90645/`. SIKS Dissertation Series No. 2014-20.

[2] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.