D18.2 Deliverable

Notification of Delivery of the Publication Platform for the Results of Digitization and Documentation

Type of Deliverable: Software and Manuals

DOCUMENT IDENTIFIER	PS_WP18_Eurix_D18.2._Publication_platform_v2.3
DATE	November, 16, 2007
ABSTRACT	This is the notification of delivery of the Publication Platforms of the MAD System. This Platform will be able to manage contributions from preservation and restoration areas and will provide the software modules for accessing them with a standard browser. The whole MAD System is made up of two other platforms as discussed in the document D18.1. The two deliverable D18.1 and D18.2 will represent the complete MAD System of the Metadata Access and Delivery Area within the project
AUTHOR, COMPANY	W. Allasia, A. Damiani, M. Porzio, S. Ridolfi, F. Toscano, M. Vigilante - euriX Group
KEYWORDS	Publication, metadata access, storage
RELATED ITEMS	Annex1 – D18.2 Glossary D18.1 Documentation Platform for the MAD Factory

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0.1	2004-02-27	First Draft	Living	Confidential
1.0	2004-12-20	Working Draft	Living	Confidential
1.1	2005-01-24	Release Candidate	Living	Confidential
2.0	2005-07-31	Working Draft	Living	Confidential
2.1	2007-05-28	Release Candidate	Living	Confidential
2.2	2007-11-16	Release Candidate	Completed	Confidential
2.3	2008-02-28	Dissemination level made Public	Living	Public

Tables of Contents

Executive Summary

This deliverable will concern the description of the Publication Platforms of the MAD System. This Platform will be able to manage contributions from preservation and restoration areas and will provide the software modules for accessing them with a standard browser.

The whole MAD System will be made up of two other platforms as discussed in the document D18.1. The two deliverable D18.1 and D18.2 will represent the complete MAD System of the Metadata Access and Delivery Area within the project

1 Overview

The MAD system will be made up of two main components:

1. Documentation platform
2. Publication platform

Point 2 will be covered in this paper. Point 1 is covered in D18.1.

The Publication Platform is the component of the MAD Platform providing retrieval and browsing functionalities. In detail, it deals with instances of documents in MAD metadata format, making them available on a web representation, and it gives access to the material sources exported from the Core Platform.

The Publication platform has to deal mainly with MPEG7 and MXF formats and with EDOB format for the metadata.

The Publication Platform comprises three different main subcomponents:

- a web application, namely the user interface;
- a relational DBMS that stores information related to the available programmes;
- a text search and indexing engine (Lucene – KIM), comprising a semantic engine for processing natural language queries.

The searching interface of the Publication Platform offers several searching approaches, and the user can choose to apply for a programme or a news item, which can be filtered by programme title, broadcast date, authors, topics, and so on.

The user interface presents a video preview, currently making use of Windows Media Player. This is the only feature written specifically for Internet Explorer.

A schema of the Publication Platform is shown in Figure 1.

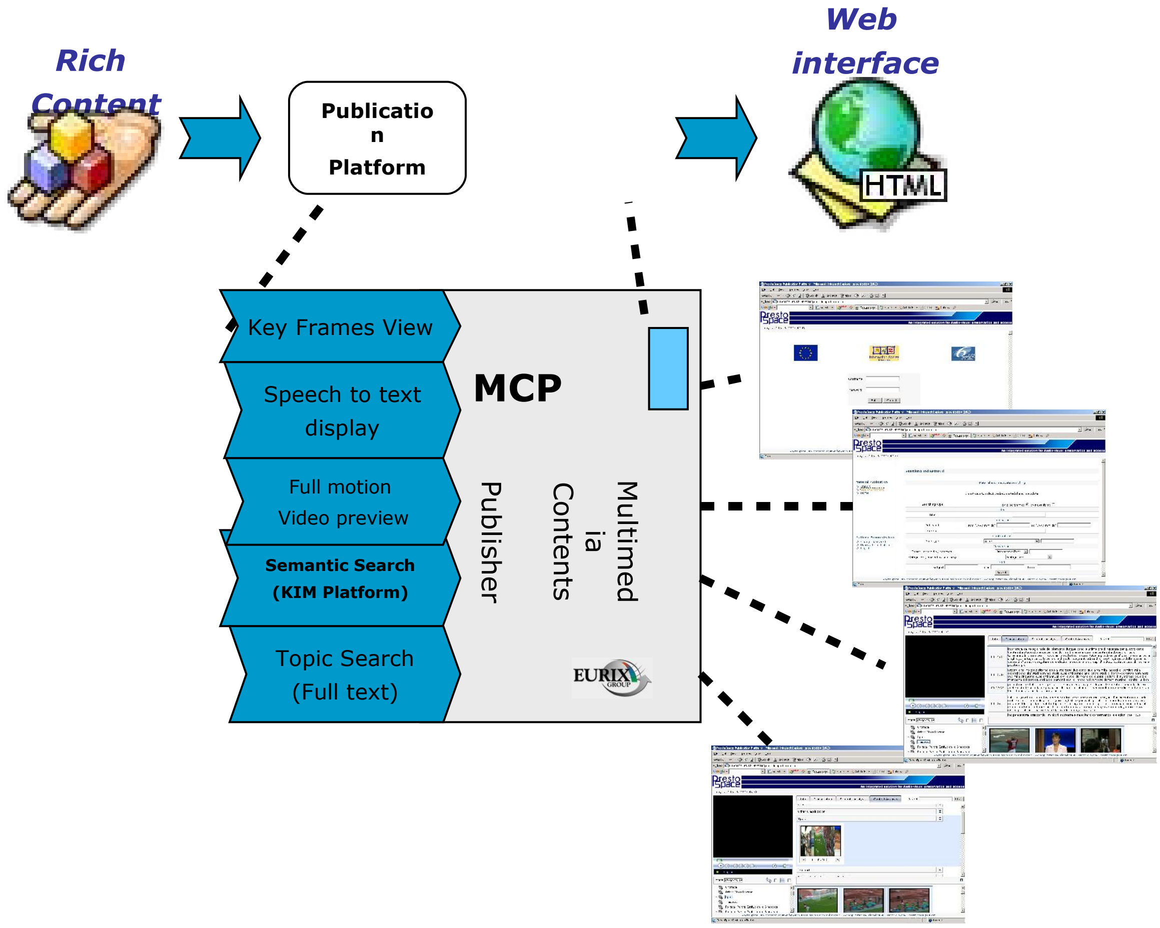


Figure 1: the Publication Platform

2 Functionalities

The Publication Platform has to provide a common web interface for importing, searching, querying and browsing metadata and essence documents.

Concerning the import functionality, a software module importing information used by the Publication Platform is provided. As we will describe later, the Publication Platform makes use of the information of a relational database and of the KIM system. By means of this import component, tuples in the database and indexes of KIM can be obtained from the EDOB produced by the Documentation Platform. It could be useful to extend this import module in order to import in the Publication Platform also information provided by another kind of document (rather than an EDOB), for instance containing Dublin Core or P-META metadata.

The main functionalities of the Publication are related to Search & Retrieve.

Search and retrieve functionality allows users to browse programmes and news items, after searching and selecting on data defined into the metadata files. Searching can be processed on programme content too by KIM Platform support.

When an user selects this functionality, a form to acquire searching criteria is presented. The user can choose between executing a programmes selection or a news items selection. Criteria are divided into 4 sections: identification, title, publication, contribution, classification and programme content (KIM section). For every section, user can define data to perform a query on database. Selected programmes or selected news items are showed into a table, where it's possible to modify list sorting by clicking on columns titles. Once the user selects searched programmes or news item, the application shows the browsing window, where user can navigate through metadata and essence.

The publication platform has to allow functionality at authenticated users. To do this, the platform is supported by Jakarta Tomcat authentication capability. To set up login permissions it has to modify tomcat-users.xml configuration file in Tomcat configuration (cfg) directory.

If a user does not do anything for almost 30 minutes, the session expires and the user is requested to log in again.

PART A: Features of the Publication Platform

3 The MAD Platform

Metadata can be defined as “Data about data”, that is to say those information that describes, or supplements, the main (or central) data. Concerning the broadcast archives scenario, this entails finding which information schemes are needed in order to make archive users able to retrieve audiovisual items with effective levels of accuracy.

The MAD (Metadata Access and Delivery) Platform is the component of the PrestoSpace project having the following objectives:

1. extracting metadata from audiovisual items;
2. offering suitable mechanisms for retrieving and accessing audiovisual contents based on metadata.

3.1 Architecture of the Publication Platform

The Publication platform will provide retrieval and browsing functionalities regarding the essence elaborated within the MAD Platform.

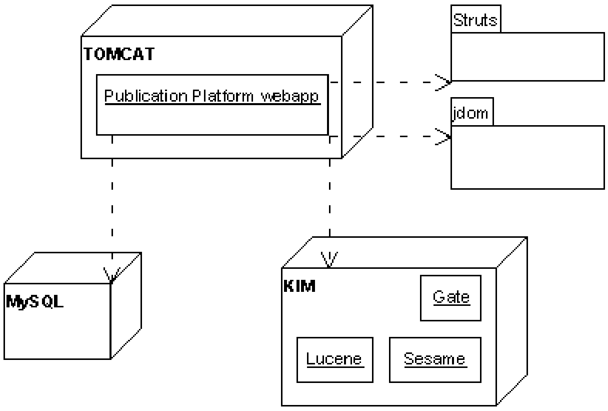


Figure 2 : The Publication Platform - architecture

The platform architecture is based on three main components: a web application to allow user interaction; a database (MySQL) to store data about available programmes and so to make easy searching and selections; the KIM platform (provided by Ontotext) to perform semantic functionality through semantic analysis of speech and full text indexing.

The Publication Platform is delivered as a web archive. Deployment is performed by posting the web archive into the servlet container of the used web server. After completed the deployment phase, it's possible to set up the platform, launching an ant build file released within the web archive.

3.2 Inputs

The Publication platform will access the data elaborated by the Documentation Platform to fill the KIM text indexing engine and the MySQL database using an export facility. Data exported from the Documentation Platform will be available as static HTML pages to be published in a web browser.

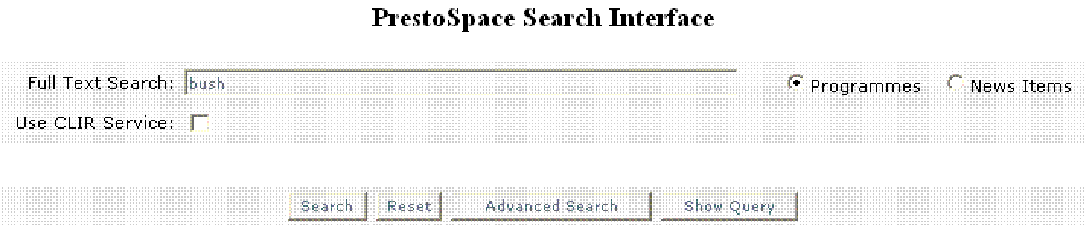
3.3 Outputs

The Publication Platform provides a web interface for searching and retrieving information produced by the Documentation Platform.

4 Web interface

The Publication Platform provides a web interface for searching and retrieving information produced by the Documentation Platform.

The entry point for queries is the form shown in Figure 3



PrestoSpace Search Interface

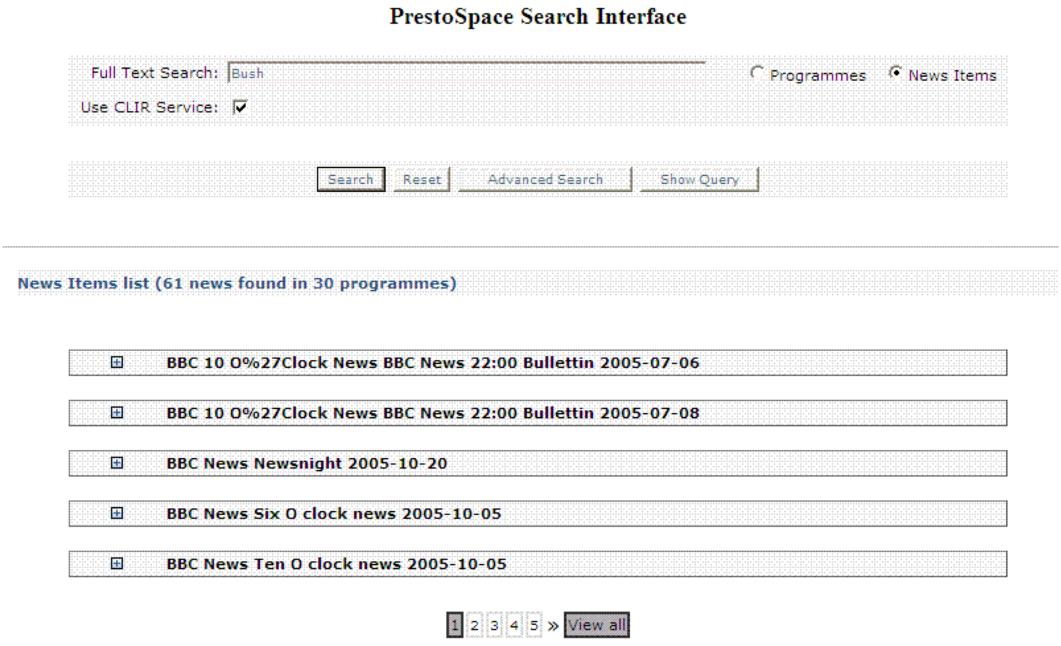
Full Text Search: bush ● Programmes ○ News Items

Use CLIR Service: □

Figure 3 : the Search Interface

Basically, the user can submit a keyword and start the search among programmes or news, searching by contribution, title, publication date, publication service, topic and named entities for semantic queries (i.e. programmes/news which contains Persons, Places, and so on).

The results of the query are then shown in a list (Figure 4) from which the user can select a document in order to browse it.



PrestoSpace Search Interface

Full Text Search: Bush ○ Programmes ● News Items

Use CLIR Service: ☑

News Items list (61 news found in 30 programmes)

⊞ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-06

⊞ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-08

⊞ BBC News Newsnight 2005-10-20

⊞ BBC News Six O clock news 2005-10-05

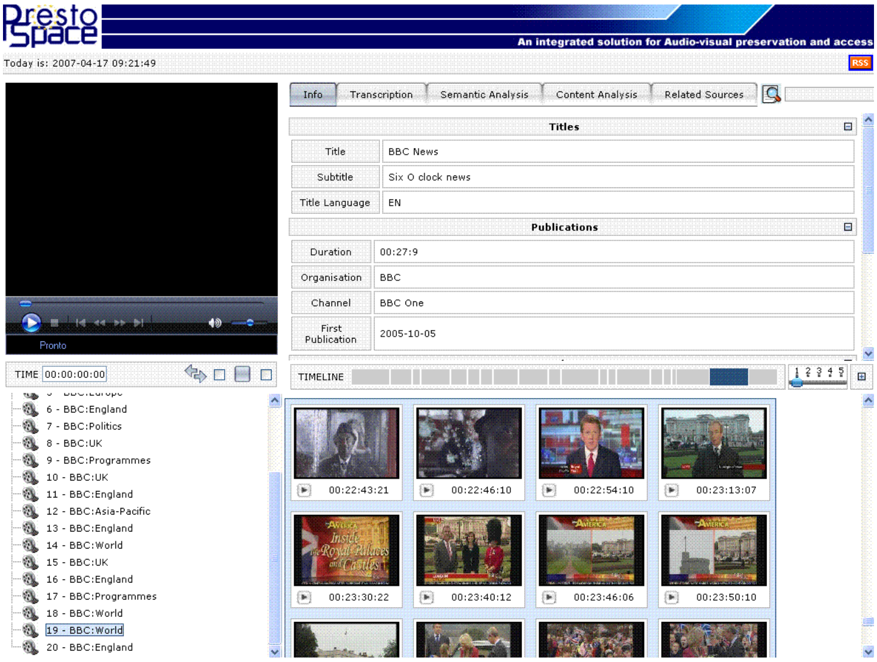
⊞ BBC News Ten O clock news 2005-10-05

Figure 4 : the list containing the results of a query

THE CLIR tool

Optionally, the user can enable a feature to translate the submitted query from the Italian language to the English language and vice versa. Source language of queries can be different from the target language (characterizing metadata).

Cross-language Information Retrieval (CLIR) is supported in the MAD publication platform by a specific server called CLIR server.



Today is: 2007-04-17 09:21:49

Titles

Title	BBC News
Subtitle	Six O clock news
Title Language	EN

Publications

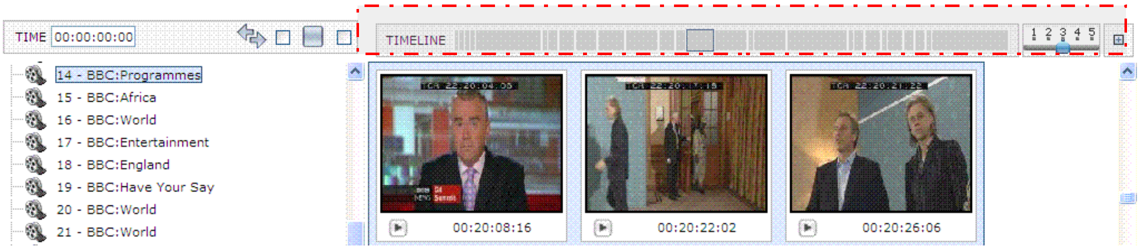
Duration	00:27:9
Organisation	BBC
Channel	BBC One
First Publication	2005-10-05

TIME 00:00:00:00

6 - BBC:England
7 - BBC:Politics
8 - BBC:UK
9 - BBC:Programmes
10 - BBC:UK
11 - BBC:England
12 - BBC:Asia-Pacific
13 - BBC:England
14 - BBC:World
15 - BBC:UK
16 - BBC:England
17 - BBC:Programmes
18 - BBC:World
19 - BBC:World
20 - BBC:England

Figure 5: the web page for browsing the selected programmes/news

In the left part of the page, there is the video section (upper), and the tree structure showing the segmentation of the programme in news. This segmentation is also shown as a timeline (Figure 6), and it is based upon a video analysis performed during the Documentation.



TIME 00:00:00:00

14 - BBC:Programmes
15 - BBC:Africa
16 - BBC:World
17 - BBC:Entertainment
18 - BBC:England
19 - BBC:Have Your Say
20 - BBC:World
21 - BBC:World

Figure 6: the timeline section

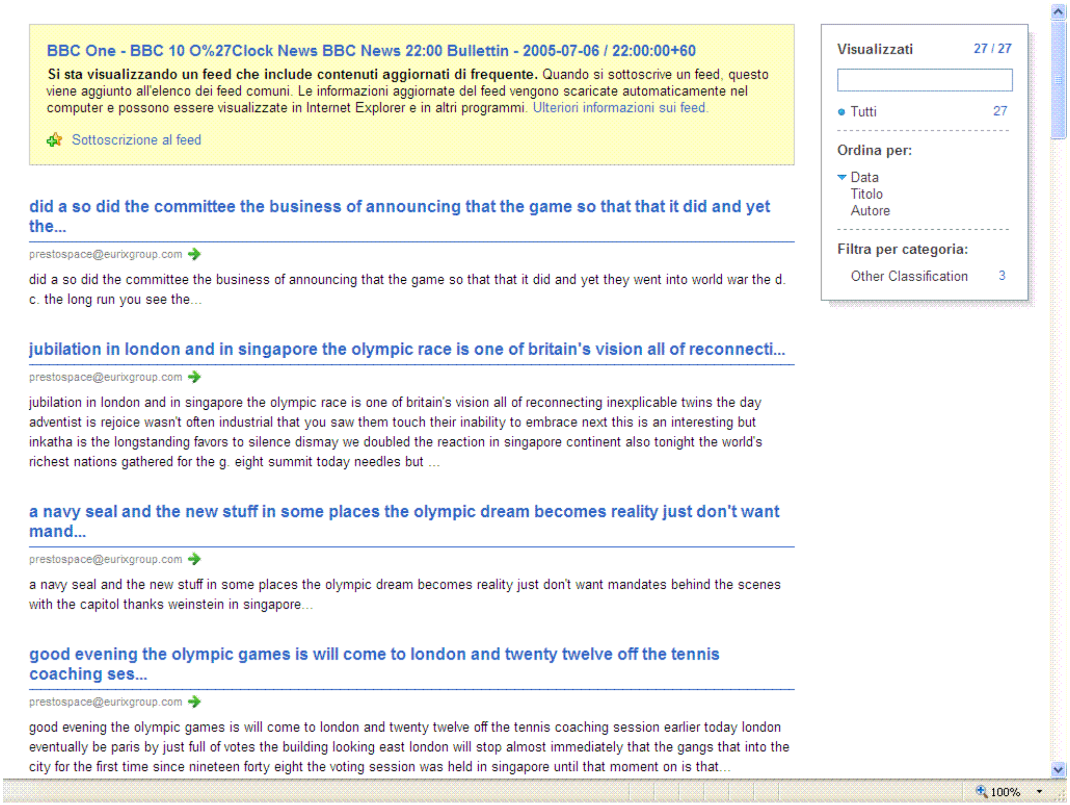
Each of the segments describes a single highlight and is related to the shots presented in the bottom of the right side of the web page.

The remaining (main) part of the page provides several tabs showing:

- Info (titles, publications, contributions and identifiers) : legacy data
- Transcription: the entire text converted from speeches (the user can do a textual search)
- Semantic analysis (using KIM facility – see section 5.1–)
- Content analysis (stripes and camera motion, if extracted during the documentation)
- Related sources (correlated news from external web sites)

RSS system

The Platform supplies the feature for exporting the programme in the RSS (Really Simple Syndication) format, and then read it with the aim of a feed reader (Figure 7).



BBC One - BBC 10 O%27Clock News BBC News 22:00 Bullettin - 2005-07-06 / 22:00:00+60

Si sta visualizzando un feed che include contenuti aggiornati di frequente. Quando si sottoscrive un feed, questo viene aggiunto all'elenco dei feed comuni. Le informazioni aggiornate del feed vengono scaricate automaticamente nel computer e possono essere visualizzate in Internet Explorer e in altri programmi. Ulteriori informazioni sui feed.

Sottoscrizione al feed

did a so did the committee the business of announcing that the game so that that it did and yet the...

prestospace@eurixgroup.com

did a so did the committee the business of announcing that the game so that that it did and yet they went into world war the d. c. the long run you see the...

jubilation in london and in singapore the olympic race is one of britain's vision all of reconnecti...

prestospace@eurixgroup.com

jubilation in london and in singapore the olympic race is one of britain's vision all of reconnecting inexplicable twins the day adventist is rejoice wasn't often industrial that you saw them touch their inability to embrace next this is an interesting but inkatha is the longstanding favors to silence dismay we doubled the reaction in singapore continent also tonight the world's richest nations gathered for the g. eight summit today needles but ...

a navy seal and the new stuff in some places the olympic dream becomes reality just don't want mand...

prestospace@eurixgroup.com

a navy seal and the new stuff in some places the olympic dream becomes reality just don't want mandates behind the scenes with the capitol thanks weinstein in singapore...

good evening the olympic games is will come to london and twenty twelve off the tennis coaching ses...

prestospace@eurixgroup.com

good evening the olympic games is will come to london and twenty twelve off the tennis coaching session earlier today london eventually be paris by just full of votes the building looking east london will stop almost immediately that the gangs that into the city for the first time since nineteen forty eight the voting session was held in singapore until that moment on is that...

Figure 7: RSS export feature

5 Implementation of the Publication Platform

5.1 Physical environment

The web application is developed on JDK 1.4.2 , using Java web technologies. It needs a web server with servlet container to run, as Jakarta Tomcat 5.5, but it's possible to use any web server compliant with Java Servlet 2.4 Specifications and JSP 2.0 Specifications. The design takes advantage of the MVC pattern to separate presentation logic and business logic. The Jakarta Struts Framework has been adopted in order to implement the controller layer, which takes into account the task of the business control flow, mapping user request with business operations of the model layer.

In order to perform searching and selections on programmes and news items, the platform is supported by a database, storing information from metadata (e.g. titles, roles, descriptions, publishing dates, services etc.). The connection between the web application and the database management system is provided by the JDBC support. So it's quite easy to change the DBMS.

The KIM Platform (provided by Ontotext) is integrated into the Publication Platform in order to provide semantic analysis capability. To give more details, the KIM Platform consists in a system based on three components: Lucene, Sesame and Gate. Together, they allow searching about semantic content of the programmes, through simple queries formulated as sentences with subject, action and target.

The Publication Platform is delivered as web archive. Deployment is performed by posting the web archive into the servlet container of the used web server. After the deployment phase is completed, it is possible to set up the platform, launching an ant build file released with in the web archive.

5.2 The Kim Platform

The KIM Platform provides a novel Knowledge and Information Management (KIM) infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content.

As a base line, KIM analyzes texts and recognizes references to entities (like persons, organizations, locations, dates). Then it tries to match the reference with a known entity, having a unique URI and description. Alternatively, a new URI and description are automatically generated. Finally, the reference in the document gets annotated with the URI of the entity. This process is called (as well as the result) semantic annotation. This sort of meta-data can be used for indexing, retrieval, visualization and automatic hyper-linking of documents.

For the purposes of semantic annotation, indexing, and retrieval of documents, KIM also uses a seed *knowledge base* (KB). The knowledge base (KB), in this context, is a body of formal knowledge about entities, representing non-ontological formal knowledge. It consists of instance data – descriptions of entities and their interrelations, i.e. for each entity, the KB contains information about the entity's type, aliases (incl. a main alias - official or well-known name), attributes, and relations. The KIM KB provides coverage of popular real-world entities of common interest, which are considered well-known and thus not explicitly introduced in the documents. Most important and used entities in the KIM KB are *geographic names and* *organizations*. The entities that represent geographical features are imported from *GNS* (*GEOnet Names Server*) and other sources. They are organized so as to represent instances of *Location* (and its subclasses) having the property *subRegionOf* as it is applied between *Continents*, *GlobalRegions*, *Countries*, and other subclasses of *Location*. Some of the subtypes of *Location*, contained in KIM KB are *Country, Province, County, CountryCapital, City, Ocean, Sea*, etc. The locations are given together with several of their aliases, including in English and French, as well as with their geographic coordinates (*Long/Lat*), the designator (*DSG*) and Unique Feature Index (*UFI*), according to *GNS*. All this provides a useful basis for cross-linguistic querying and retrieval. The entities in the KB are derived or collected from various sources like geographical and business intelligence gazetteers.

As a part of the Publication Platform, the KIM engine supplies an indexing of the EDOB's metadata.

The role of KIM in MAD is to provide a language independent representation for Named Entities as a specific metadata common to the two languages. As an example, consider that the *"White House"* is translated in other languages (e.g. in Italian the correct translation is *"Casa Bianca"*). The ontology representation for this entity is via a single id (i.e. an Uniform Resource Identifier *"URI"*), that is for its nature language independent. This realizes a systematic and consistent approach to multilingual indexing and searching.

5.3 The CLIR Server

As described in Figure 8 , the CLIR Server includes several components:

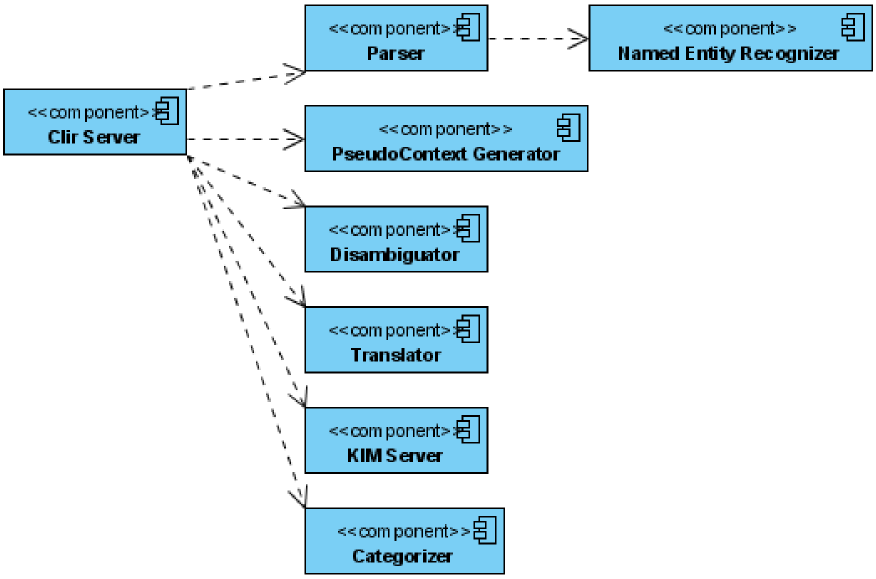


Figure 8: the CLIR server

- The *NL Parser*, to extract *Named Entities* and other nouns from the query *q,* in the source language *L*;
- *Pseudo Context Generator*, to generate for each target lexical item *t* in *q,* the most relevant terms that are topically related to *t*;
- *Sense Disambiguator*, to disambiguate all common nouns in the source language *L*;

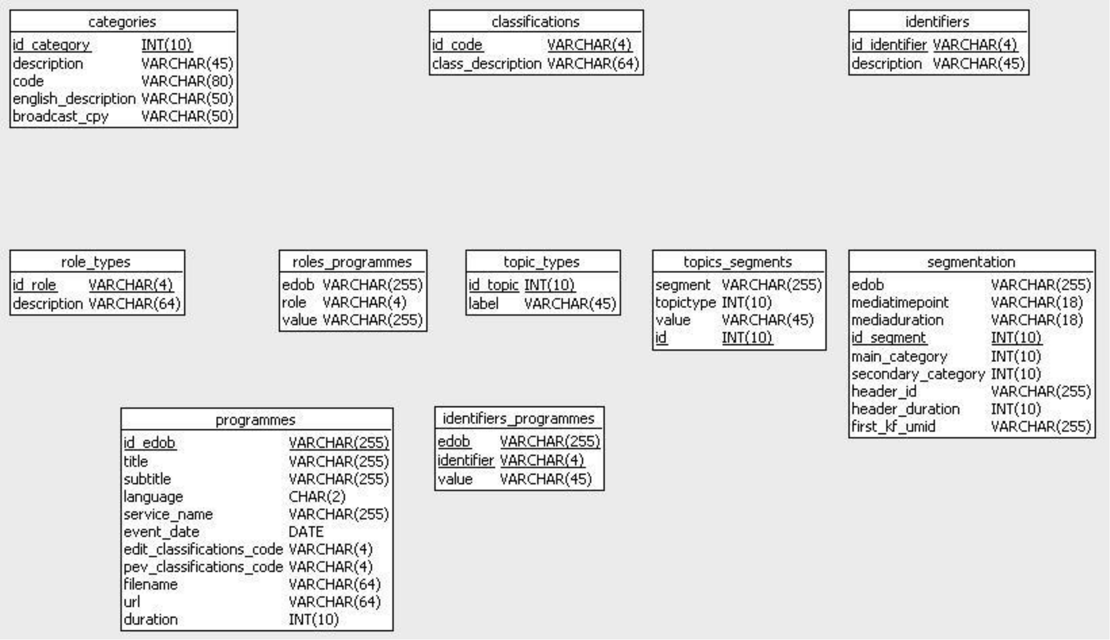
- *Translator*, to translate the disambiguated common nouns from the source language *L* to the target language *L2*;
- *Kim Server*, to annotate the ontological entries as they are found in a query *q*;
- *Text Categorizer* that classifies the query *q*.

The CLIR Server communicates with these components and manages the internal workflow. The NL parser, Text Categorization and Kim annotation processes are the same used for the Semantic Analysis.

A distinctive feature of the CLIR server is the adopted technique for Sense Disambiguation and Translation. Translation of all common nouns is required as they are very language specific and must be consistently combined to the language-independent representation of Named Entities. The sense disambiguation algorithm adopted has been presented in [1].

5.4 The MySQL database.

It provides a data set of the EDOBs published and the related METADATA.



categories	
id_category	INT(10)
description	VARCHAR(45)
code	VARCHAR(80)
english_description	VARCHAR(50)
broadcast_cpy	VARCHAR(50)

classifications	
id_code	VARCHAR(4)
class_description	VARCHAR(64)

identifiers	
id_identifier	VARCHAR(4)
description	VARCHAR(45)

role_types	
id_role	VARCHAR(4)
description	VARCHAR(64)

roles_programmes	
edob	VARCHAR(255)
role	VARCHAR(4)
value	VARCHAR(255)

topic_types	
id_topic	INT(10)
label	VARCHAR(45)

topics_segments	
segment	VARCHAR(255)
topictype	INT(10)
value	VARCHAR(45)
id	INT(10)

segmentation	
edob	VARCHAR(255)
mediatimepoint	VARCHAR(18)
mediaduration	VARCHAR(18)
id_segment	INT(10)
main_category	INT(10)
secondary_category	INT(10)
header_id	VARCHAR(255)
header_duration	INT(10)
first_kf_umid	VARCHAR(255)

programmes	
id_edob	VARCHAR(255)
title	VARCHAR(255)
subtitle	VARCHAR(255)
language	CHAR(2)
service_name	VARCHAR(255)
event_date	DATE
edit_classifications_code	VARCHAR(4)
pev_classifications_code	VARCHAR(4)
filename	VARCHAR(64)
url	VARCHAR(64)
duration	INT(10)

identifiers_programmes	
edob	VARCHAR(255)
identifier	VARCHAR(4)
value	VARCHAR(45)

Figure 9: Tables of the MySQL database used by the Publication Platform

The above relational database describes the data used for characterizing the EDOBs, such as role types, topic types, categories, and the programmes and segmentations related to the EDOB itself.

6 Notification of Delivery

The Publication Platform is delivered using a VMWare virtual machine. In this way it’s possible to install the Publication Platform simply running this virtual machine with the VMWare Player (that can be used for free.

This virtual machine is a Linux system with all the necessary installed in it (KIM, CLIR, Apache Tomcat, MySQL and so on).

When the virtual machine starts all the services are started up so when the startup of the system is completed you only have to connect to this url http://192.168.11.128:8080/PublicationPlatform using the Internet Explorer of your host pc.

That’s all, your Publication Platform is ready to be used.

This kind of delivery was used for the last demos of the Publication Platform done during some public events like IBC or some international conferences.

This is a very simple and successful way to delivery the Publication Platform because nothing should be installed or configured (you have only to download and install VMWare Player).

PART B: Utilization of the Publication Platform

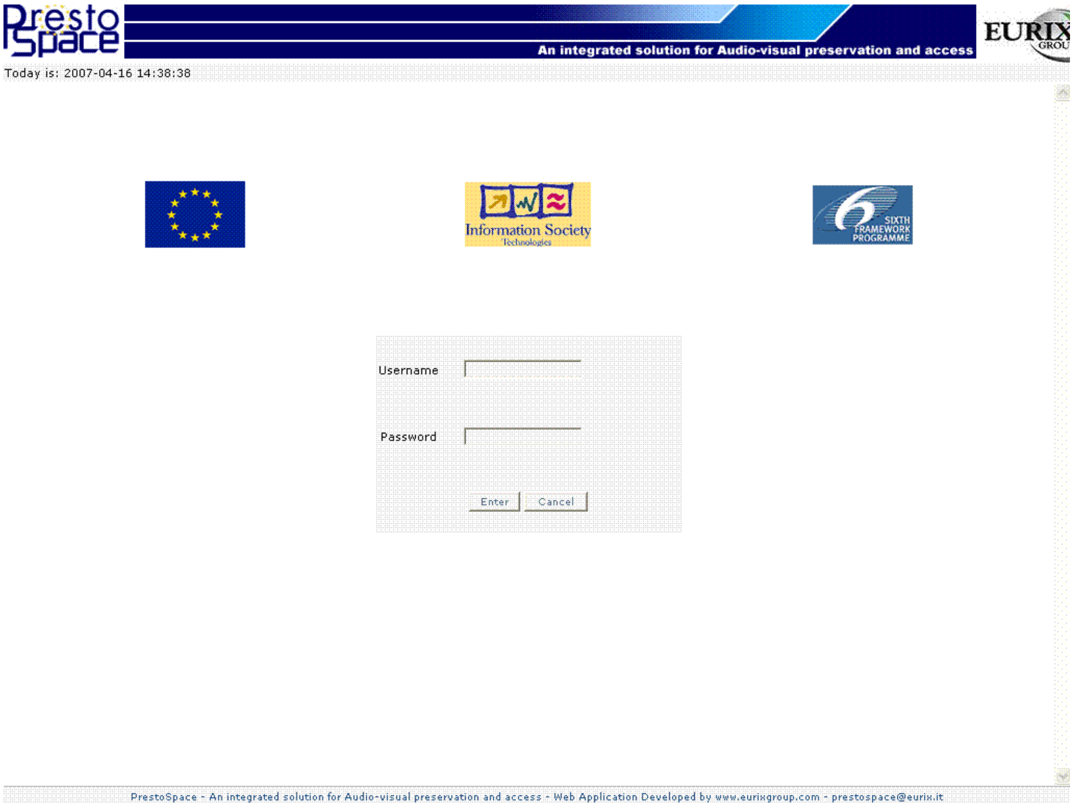
7 How to use the Publication Platform

7.1 How to access the Publication Platform

The Publication Platform is accessible via any browser at the following URL:

http://prestospace.eurix.it/PublicationPlatform

The user has to submit a valid Username and Password:



An integrated solution for Audio-visual preservation and access

Today is: 2007-04-16 14:38:38

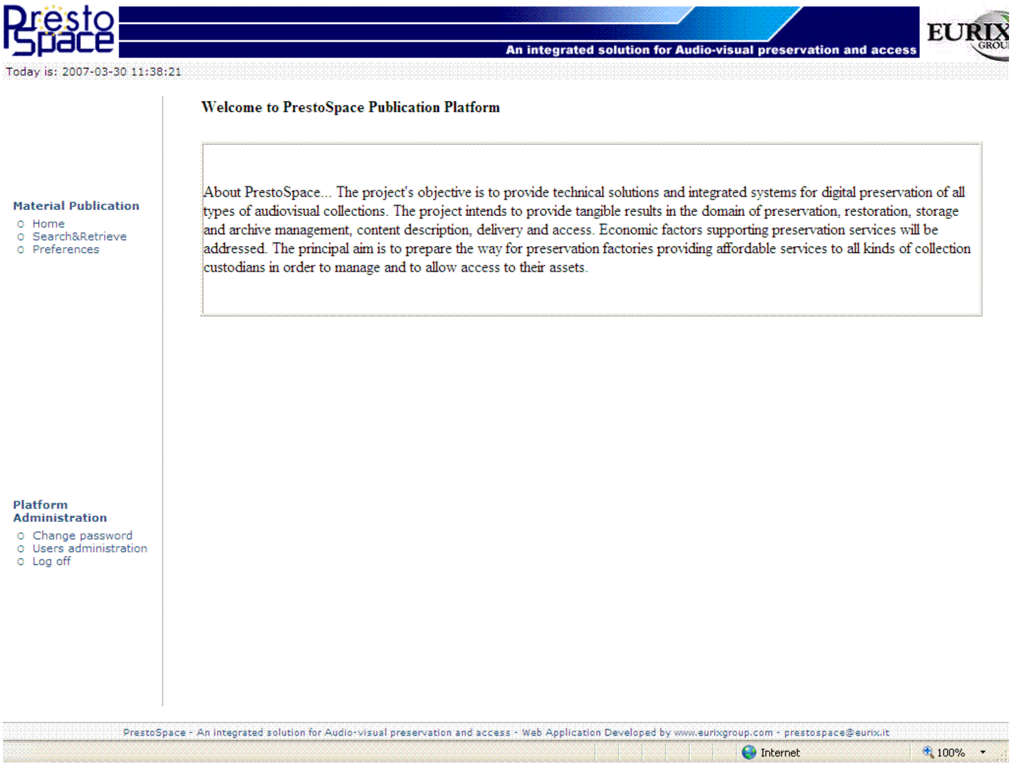
Username

Password

PrestoSpace - An integrated solution for Audio-visual preservation and access - Web Application Developed by www.eurixgroup.com - prestospace@eurix.it

Figure 10 : The Welcome Page of the Publication Platform

After that, it is possible to access the contents of the platform.



Welcome to PrestoSpace Publication Platform

About PrestoSpace... The project's objective is to provide technical solutions and integrated systems for digital preservation of all types of audiovisual collections. The project intends to provide tangible results in the domain of preservation, restoration, storage and archive management, content description, delivery and access. Economic factors supporting preservation services will be addressed. The principal aim is to prepare the way for preservation factories providing affordable services to all kinds of collection custodians in order to manage and to allow access to their assets.

PrestoSpace - An integrated solution for Audio-visual preservation and access - Web Application Developed by www.eurixgroup.com - prestospace@eurix.it

Figure 11: The Publication Platform Web Interface

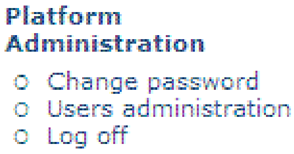
The user can access the documented programme/news or manage the user accounts.

7.2 Platform Administration

By means of the web application, a user can perform the usual administration activities, namely:

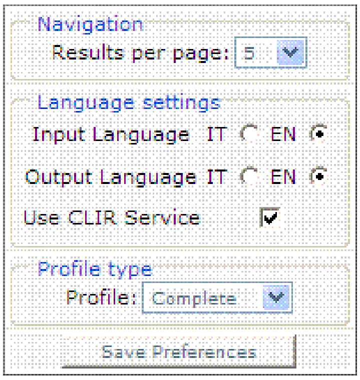
- change its own password
- access to the administration of users
- log off the system.

The web page is shown here below.



7.3 Material Publication

Preferences



Results per page: 5

Input Language IT ○ EN ●

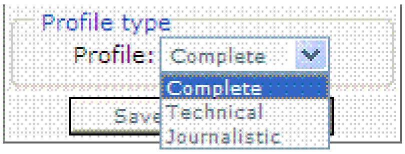
Output Language IT ○ EN ●

Use CLIR Service ☑

Profile: Complete

Figure 12 : Preferences

The user can set the number of results displayed in a page (default: 5), the languages of inputs and outputs if using the CLIR service, and the type of information. In fact, it is possible to choose among technical (only key frames, camera motions and other technical information), journalistic (only the documented editorial parts) or both.

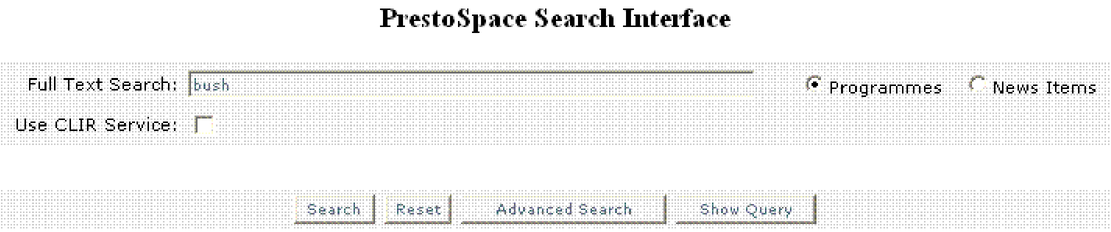


Profile: Complete

Search&Retrieve

Simple Search

The simplest task it to find a keyword by a full text search. This is equivalent to find a word within the EDOBs submitted by the Archive(s)**.**



Full Text Search: bush

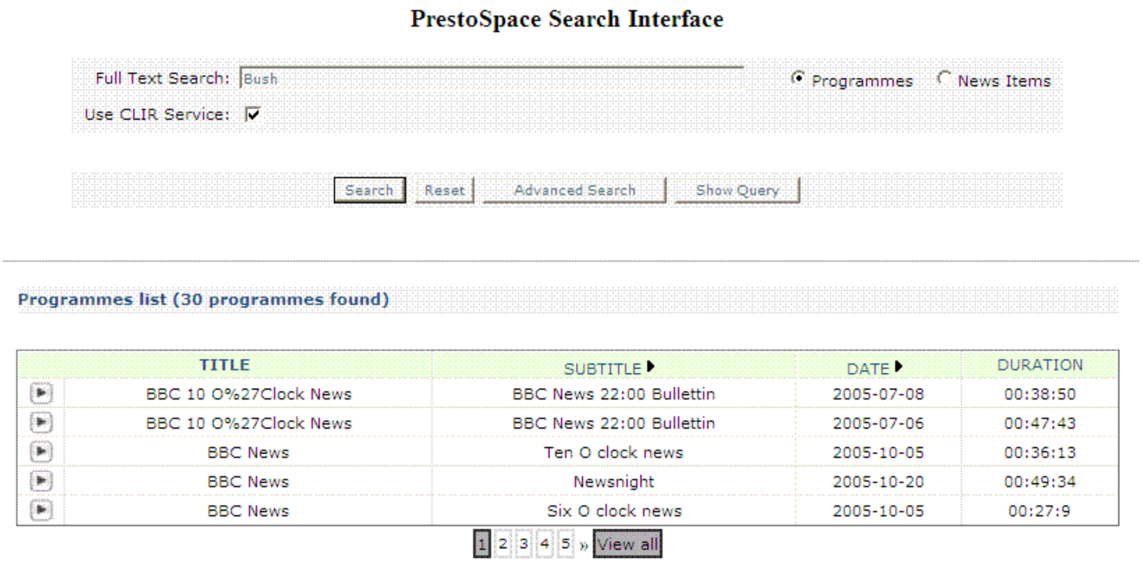
● Programmes ○ News Items

Use CLIR Service: □

Figure 13 : The Search Interface

It is possible to search among Programmes or News Items. By clicking on the “Search” button the user starts the search.

Following the preferences, the page displays the results of the query.



PrestoSpace Search Interface

Full Text Search: Bush

Use CLIR Service: ☑

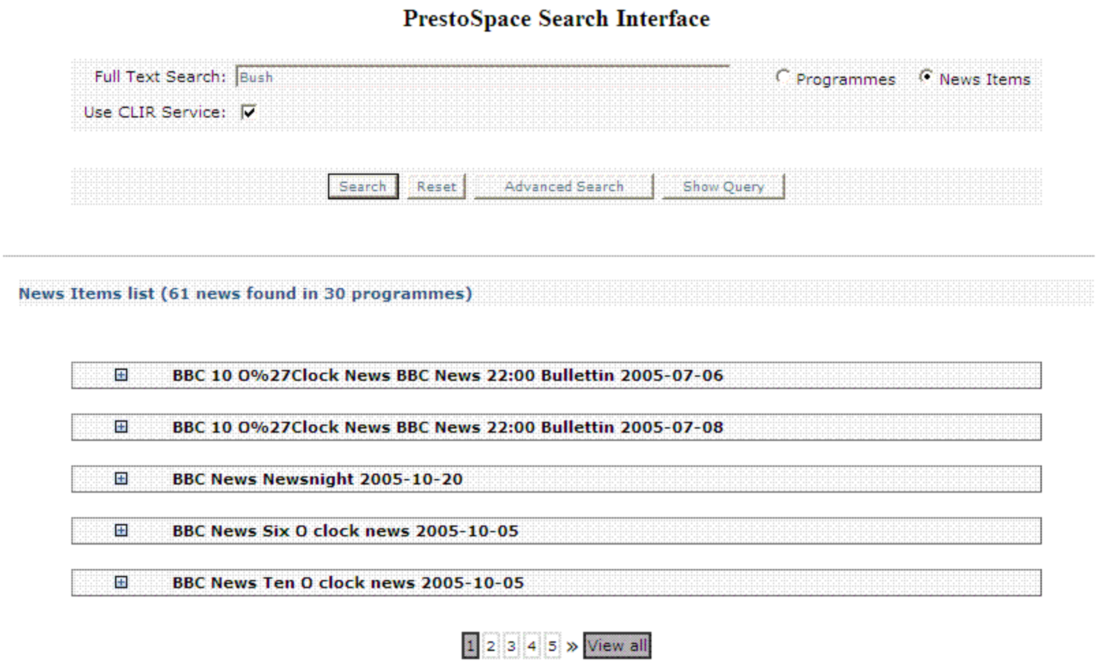
● Programmes ○ News Items

Programmes list (30 programmes found)

	TITLE	SUBTITLE ▶	DATE ▶	DURATION
▶	BBC 10 O%27Clock News	BBC News 22:00 Bullettin	2005-07-08	00:38:50
▶	BBC 10 O%27Clock News	BBC News 22:00 Bullettin	2005-07-06	00:47:43
▶	BBC News	Ten O clock news	2005-10-05	00:36:13
▶	BBC News	Newsnight	2005-10-20	00:49:34
▶	BBC News	Six O clock news	2005-10-05	00:27:9

Figure 14 : Results of the query (programmes)

Selecting a query of news items, the page displayed looks like this one:



PrestoSpace Search Interface

Full Text Search: Bush

Use CLIR Service: ☑

○ Programmes ● News Items

News Items list (61 news found in 30 programmes)

⊞ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-06

⊞ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-08

⊞ BBC News Newsnight 2005-10-20

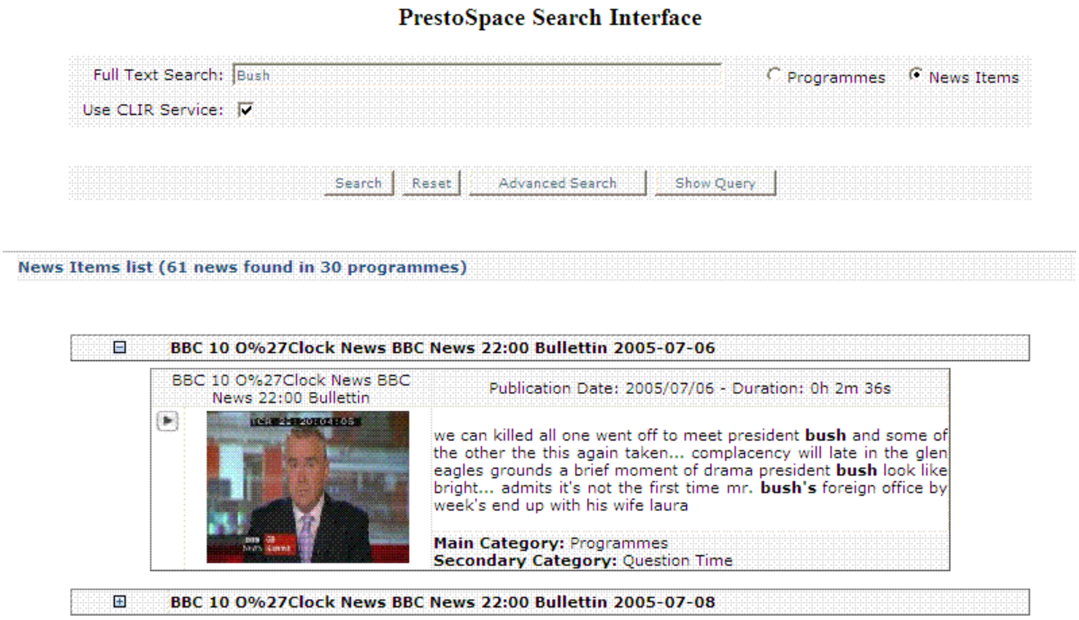
⊞ BBC News Six O clock news 2005-10-05

⊞ BBC News Ten O clock news 2005-10-05

Figure 15 : Results of the query (news items)

In the news items displayed list, it is possible to see (by clicking on the

⊞ icon) a highlight of the news in which the word is found:



PrestoSpace Search Interface

Full Text Search: Bush ○ Programmes ● News Items

Use CLIR Service: ☑

News Items list (61 news found in 30 programmes)

⊟ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-06

BBC 10 O%27Clock News BBC News 22:00 Bullettin

Publication Date: 2005/07/06 - Duration: 0h 2m 36s

we can killed all one went off to meet president **bush** and some of the other the this again taken... complacency will late in the glen eagles grounds a brief moment of drama president **bush** look like bright... admits it's not the first time mr. **bush's** foreign office by week's end up with his wife laura

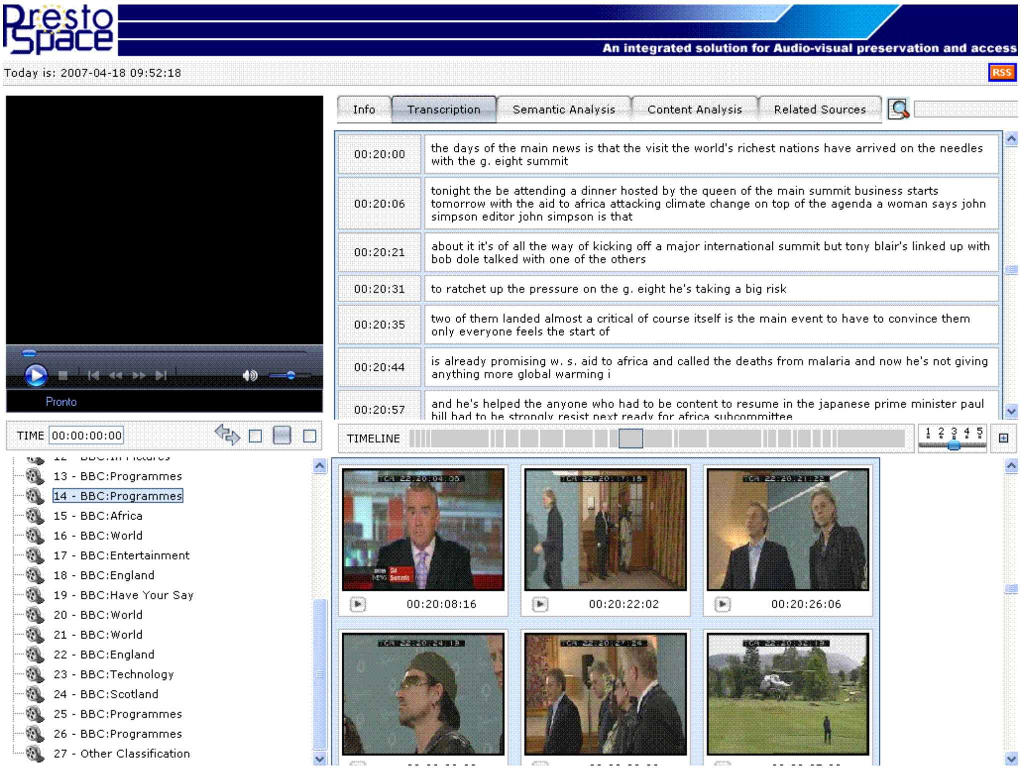
Main Category: Programmes
Secondary Category: Question Time

⊞ BBC 10 O%27Clock News BBC News 22:00 Bullettin 2005-07-08

Figure 16 : Results of the query (news items) -- expanded

Clicking the ▶ button on the right of the programme or news chosen will display a pop-up window with the streaming video of the retrieved EDOB.

Clicking on one of the retrieved items will open a new window showing the contents produced by the Documentation Platform (Figure 17).



PrestoSpace
An integrated solution for Audio-visual preservation and access

Today is: 2007-04-18 09:52:18

00:20:00	the days of the main news is that the visit the world's richest nations have arrived on the needles with the g. eight summit
00:20:06	tonight the be attending a dinner hosted by the queen of the main summit business starts tomorrow with the aid to africa attacking climate change on top of the agenda a woman says john simpson editor john simpson is that
00:20:21	about it it's of all the way of kicking off a major international summit but tony blair's linked up with bob dole talked with one of the others
00:20:31	to ratchet up the pressure on the g. eight he's taking a big risk
00:20:35	two of them landed almost a critical of course itself is the main event to have to convince them only everyone feels the start of
00:20:44	is already promising w. s. aid to africa and called the deaths from malaria and now he's not giving anything more global warming i
00:20:57	and he's helped the anyone who had to be content to resume in the japanese prime minister paul hill had to be strongly resist next ready for africa subcommittee

TIME 00:00:00:00

13 - BBC:Programmes
14 - BBC:Programmes
15 - BBC:Africa
16 - BBC:World
17 - BBC:Entertainment
18 - BBC:England
19 - BBC:Have Your Say
20 - BBC:World
21 - BBC:World
22 - BBC:England
23 - BBC:Technology
24 - BBC:Scotland
25 - BBC:Programmes
26 - BBC:Programmes
27 - Other Classification

Figure 17 : the web page for browsing the selected programme/news

In the left part of the page, there is the video section (upper), and the tree structure showing the segmentation of the programme in news. This segmentation is also shown as a timeline (Figure 6) and it is based upon a video analysis performed during the Documentation. Each of the segments describes a single highlight (and is related to the shots presented in the bottom of the right side of the web page).

Notice that the segment in which the keyword has been found (plain text or named entity) is highlighted and the related shots and transcription are displayed.

The remaining (main) part of the page provides several tabs showing:

- Info (titles, publications, contributions and identifiers) : legacy data
- Transcription: the entire text converted from speeches (the user can do a textual search)
- Semantic analysis (using KIM facility – see section 5.2–)
- Content analysis (stripes and camera motion, if extracted during the documentation)
- Related sources (correlated news from external web sites)

Info tab (legacy data)

This tab shows legacy data: Titles (title, subtitle, title language), Publications data (duration, organisation, channel, date of first publication), Contributions (like production company, news reader, editor-in-chief etc.) and Identifiers related to the source archive (programme number and archive number).

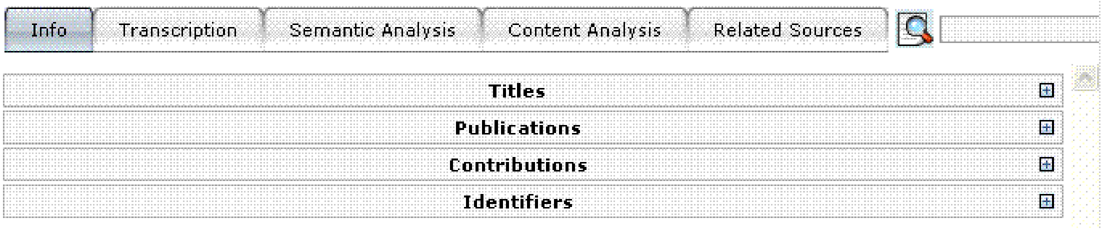
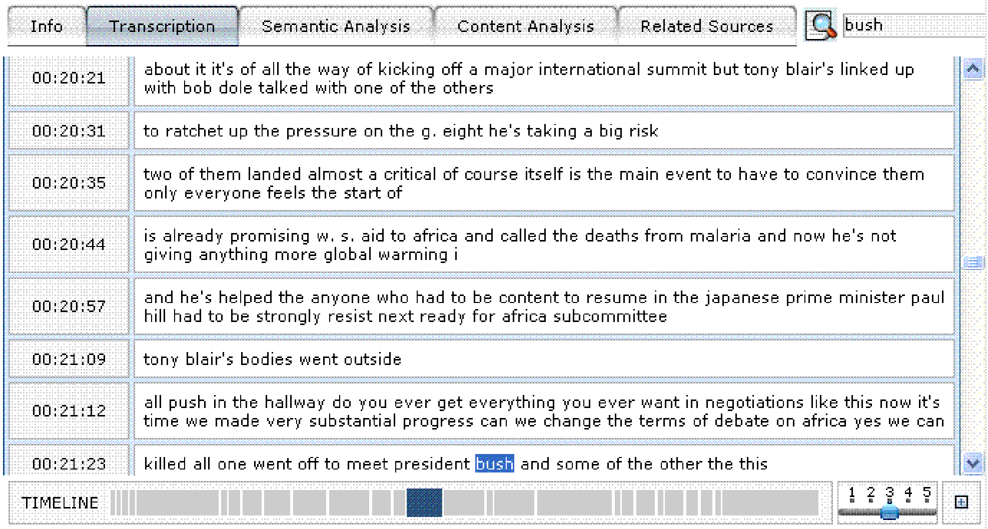


Figure 18 : legacy data – The Info tab

Transcription Tab



00:20:21	about it it's of all the way of kicking off a major international summit but tony blair's linked up with bob dole talked with one of the others
00:20:31	to ratchet up the pressure on the g. eight he's taking a big risk
00:20:35	two of them landed almost a critical of course itself is the main event to have to convince them only everyone feels the start of
00:20:44	is already promising w. s. aid to africa and called the deaths from malaria and now he's not giving anything more global warming i
00:20:57	and he's helped the anyone who had to be content to resume in the japanese prime minister paul hill had to be strongly resist next ready for africa subcommittee
00:21:09	tony blair's bodies went outside
00:21:12	all push in the hallway do you ever get everything you ever want in negotiations like this now it's time we made very substantial progress can we change the terms of debate on africa yes we can
00:21:23	killed all one went off to meet president bush and some of the other the this

Figure 19 : The Transcription tab

It shows the results of the speech-to-text analysis of the Documentation Platform. Every segment (corresponding to a silence or a change of the news reader) is labelled with the time from the starting point of the programme/news. The actual segment of the EDOB (in which the keyword submitted in the query has been found) is also shown in the timeline section.

As in all the tabs, it is possible to perform a simple plain text searching task using the input form in the upper right part of the page and clicking on the 🔍 icon.

Semantic Analysis Tab

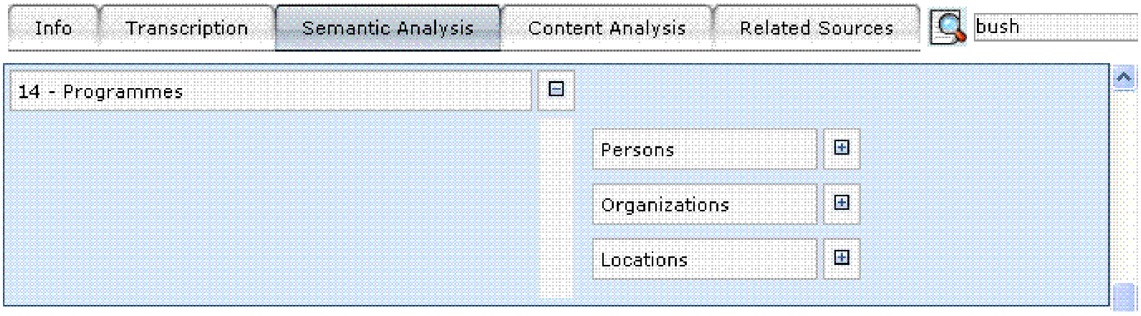


Figure 20 : Semantic Analysis Tab

Within this section the user can browse the named entities (and their categorization) founded by the semantic section of the Documentation Platform.

By clicking on the white rectangles under the categories (see Figure 21), it is possible to highlight the named entities in the transcription tab and then, clicking on them, see a pop-up window (Entity explorer – issued by KIM) with an ontological description of the entity based on the Knowledge Base of the KIM system (Figure 22).

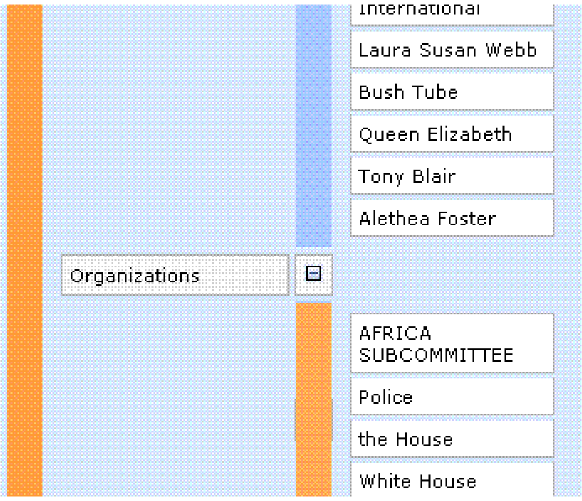
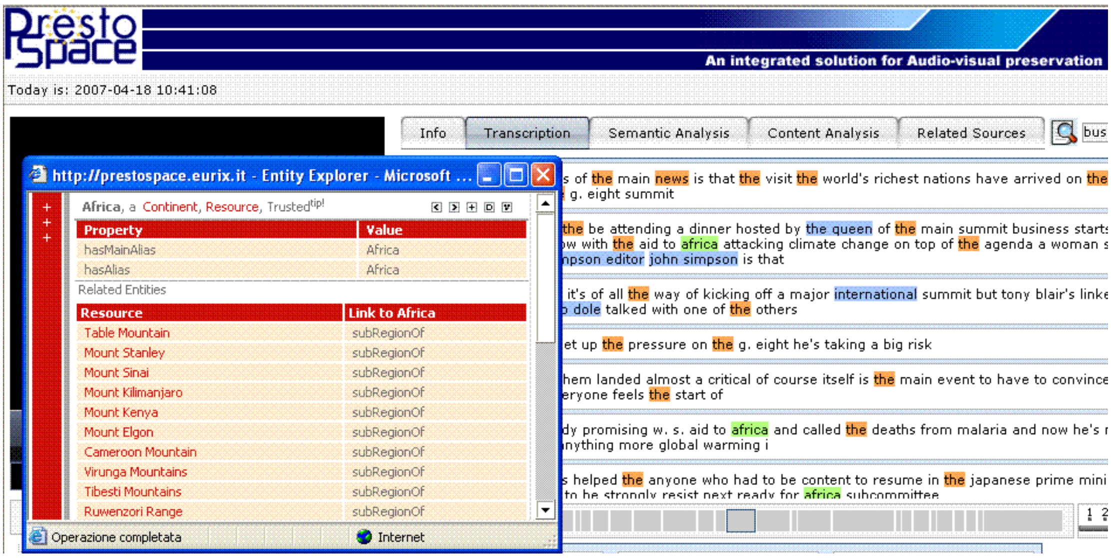


Figure 21 : named entities


PrestoSpace
An integrated solution for Audio-visual preservation
Info
Transcription
Semantic Analysis
Content Analysis
Related Sources
bus
http://prestospace.eurix.it - Entity Explorer - Microsoft ...
Operazione completata
Internet

Today is: 2007-04-18 10:41:08

Africa, a Continent, Resource, Trusted$^{tip!}$

Property	Value
hasMainAlias	Africa
hasAlias	Africa

Related Entities

Resource	Link to Africa
Table Mountain	subRegionOf
Mount Stanley	subRegionOf
Mount Sinai	subRegionOf
Mount Kilimanjaro	subRegionOf
Mount Kenya	subRegionOf
Mount Elgon	subRegionOf
Cameroon Mountain	subRegionOf
Virunga Mountains	subRegionOf
Tibesti Mountains	subRegionOf
Ruwenzori Range	subRegionOf

s of the main news is that the visit the world's richest nations have arrived on the
g. eight summit

the be attending a dinner hosted by the queen of the main summit business starts
ow with the aid to africa attacking climate change on top of the agenda a woman s
npson editor john simpson is that

it's of all the way of kicking off a major international summit but tony blair's linke
o dole talked with one of the others

et up the pressure on the g. eight he's taking a big risk

hem landed almost a critical of course itself is the main event to have to convince
eryone feels the start of

dy promising w. s. aid to africa and called the deaths from malaria and now he's r
nything more global warming i

s helped the anyone who had to be content to resume in the japanese prime mini
to be strongly resist next ready for africa subcommittee

Figure 22 : The Entity Explorer

Content Analysis Tab

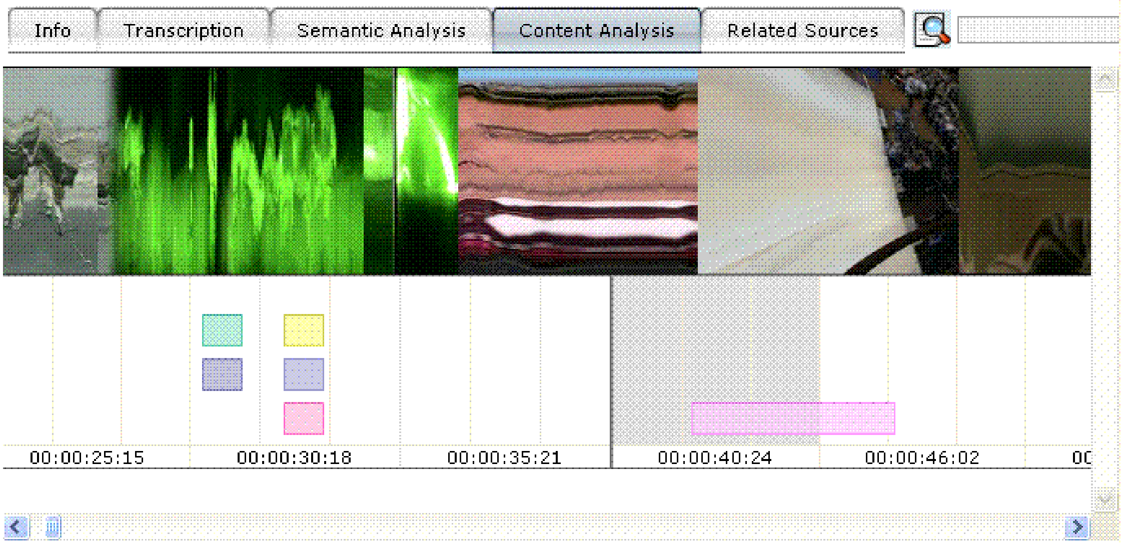


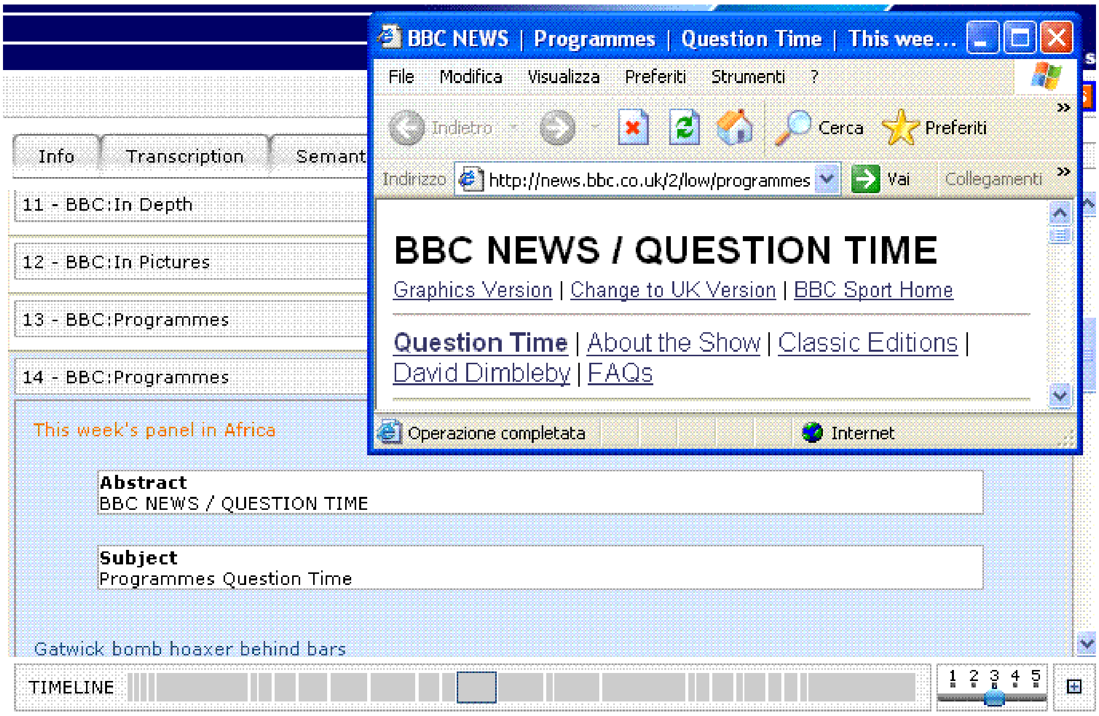
Figure 23 : Content Analysis tab

Within this section, the user can see the technical information related to the video part of the EDOB. This page shows the stripe images that represent the combination of the central column of each key frame of the shots representing the video and are useful to see changes in the camera motions, zooms, and changes at editorial level (i.e. a different editorial part).

The technical information related to the camera are displayed as coloured rectangles with an area that extends from the starting point ant until the end of the camera motion/zoom.

The displayed information are: camera pan (left – right), camera tilt (up – down) and zoom in and zoom out events.

Related Sources Tab



http://news.bbc.co.uk/2/low/programmes

11 - BBC:In Depth

12 - BBC:In Pictures

13 - BBC:Programmes

14 - BBC:Programmes

BBC NEWS / QUESTION TIME

Graphics Version | Change to UK Version | BBC Sport Home

Question Time | About the Show | Classic Editions | David Dimbleby | FAQs

This week's panel in Africa

Abstract
BBC NEWS / QUESTION TIME

Subject
Programmes Question Time

Gatwick bomb hoaxer behind bars

Figure 24 : Related Sources Tab

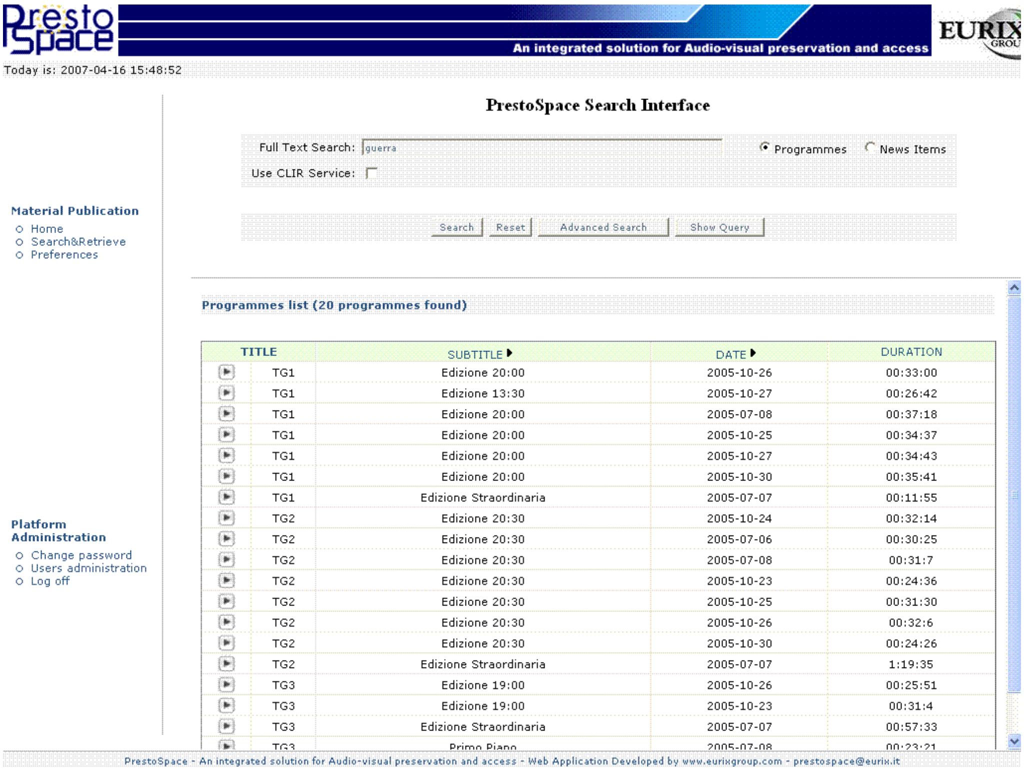
In this section the user can browse the related news founded by the Documentation Platform and see them on a new window clicking on the link in the upper left part of the panel showing the news itself.

Advanced Search / CLIR

Source language of queries can be different from the target language (characterizing metadata). For this reason, it is possible to use a Cross-language Information Retrieval (CLIR) service.

This service translates the submitted query from a language to another one (from English to Italian and vice versa). By using this tool, it is possible to extend the search to EDOBs with metadata of a language different from the one of the query.

As an example, if the user submit a simple query containing the word ‘guerra’ (the Italian word for ‘war’), it can see only the EDOBs related to audiovisual materials with an Italian origin (like news from a broadcasting Italian company i.e. TG1-RAI, Italian radiotelevision):



Today is: 2007-04-16 15:48:52

PrestoSpace Search Interface

Full Text Search: guerra ● Programmes ○ News Items

Use CLIR Service: □

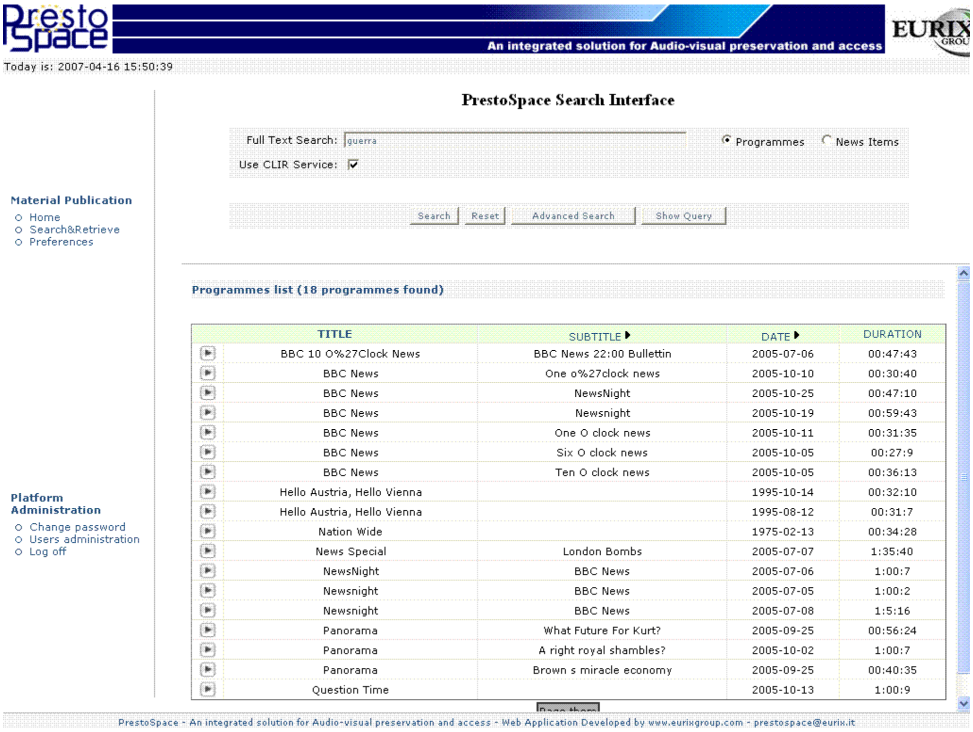
Programmes list (20 programmes found)

TITLE		SUBTITLE	DATE	DURATION
▶	TG1	Edizione 20:00	2005-10-26	00:33:00
▶	TG1	Edizione 13:30	2005-10-27	00:26:42
▶	TG1	Edizione 20:00	2005-07-08	00:37:18
▶	TG1	Edizione 20:00	2005-10-25	00:34:37
▶	TG1	Edizione 20:00	2005-10-27	00:34:43
▶	TG1	Edizione 20:00	2005-10-30	00:35:41
▶	TG1	Edizione Straordinaria	2005-07-07	00:11:55
▶	TG2	Edizione 20:30	2005-10-24	00:32:14
▶	TG2	Edizione 20:30	2005-07-06	00:30:25
▶	TG2	Edizione 20:30	2005-07-08	00:31:7
▶	TG2	Edizione 20:30	2005-10-23	00:24:36
▶	TG2	Edizione 20:30	2005-10-25	00:31:30
▶	TG2	Edizione 20:30	2005-10-26	00:32:6
▶	TG2	Edizione 20:30	2005-10-30	00:24:26
▶	TG2	Edizione Straordinaria	2005-07-07	1:19:35
▶	TG3	Edizione 19:00	2005-10-26	00:25:51
▶	TG3	Edizione 19:00	2005-10-23	00:31:4
▶	TG3	Edizione Straordinaria	2005-07-07	00:57:33
▶	TG3	Primo Piano	2005-07-08	00:23:21

PrestoSpace - An integrated solution for Audio-visual preservation and access - Web Application Developed by www.eurixgroup.com - prestospace@eurix.it

Figure 25 : results of a query without using the CLIR service – only 'Italian' EDOBs founded

Whit the CLIR service enabled, the user can extend the search and can then also find the EDOBs with English metadata:



Today is: 2007-04-16 15:50:39

PrestoSpace Search Interface

Full Text Search: guerra ● Programmes ○ News Items

Use CLIR Service: ☑

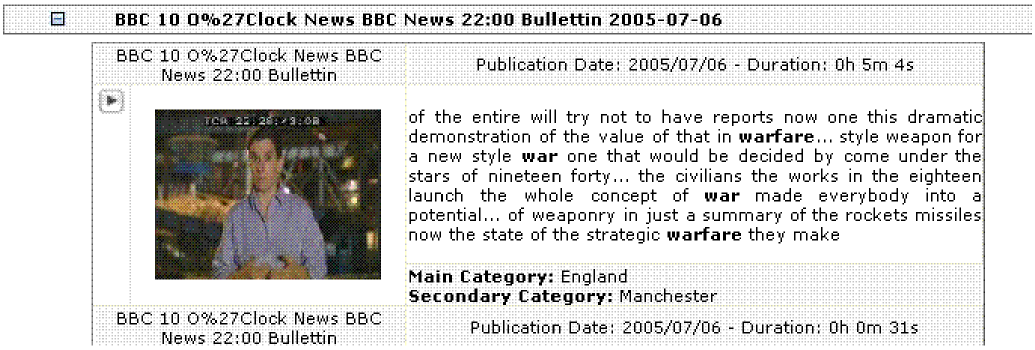
Programmes list (18 programmes found)

	TITLE	SUBTITLE	DATE	DURATION
▶	BBC 10 O%27Clock News	BBC News 22:00 Bullettin	2005-07-06	00:47:43
▶	BBC News	One o%27clock news	2005-10-10	00:30:40
▶	BBC News	NewsNight	2005-10-25	00:47:10
▶	BBC News	Newsnight	2005-10-19	00:59:43
▶	BBC News	One O clock news	2005-10-11	00:31:35
▶	BBC News	Six O clock news	2005-10-05	00:27:9
▶	BBC News	Ten O clock news	2005-10-05	00:36:13
▶	Hello Austria, Hello Vienna		1995-10-14	00:32:10
▶	Hello Austria, Hello Vienna		1995-08-12	00:31:7
▶	Nation Wide		1975-02-13	00:34:28
▶	News Special	London Bombs	2005-07-07	1:35:40
▶	NewsNight	BBC News	2005-07-06	1:00:7
▶	Newsnight	BBC News	2005-07-05	1:00:2
▶	Newsnight	BBC News	2005-07-08	1:5:16
▶	Panorama	What Future For Kurt?	2005-09-25	00:56:24
▶	Panorama	A right royal shambles?	2005-10-02	1:00:7
▶	Panorama	Brown s miracle economy	2005-09-25	00:40:35
▶	Question Time		2005-10-13	1:00:9

PrestoSpace - An integrated solution for Audio-visual preservation and access - Web Application Developed by www.eurixgroup.com - prestospace@eurix.it

Figure 26 : Results of the query using the CLIR service

The Figure 27 shows a retrieved BBC News with the 'war' and 'warfare' words founded in the transcription of the speeches.



⊟ **BBC 10 O%27Clock News BBC News 22:00 Bulletin 2005-07-06**

BBC 10 O%27Clock News BBC News 22:00 Bullettin — Publication Date: 2005/07/06 - Duration: 0h 5m 4s

of the entire will try not to have reports now one this dramatic demonstration of the value of that in **warfare**... style weapon for a new style **war** one that would be decided by come under the stars of nineteen forty... the civilians the works in the eighteen launch the whole concept of **war** made everybody into a potential... of weaponry in just a summary of the rockets missiles now the state of the strategic **warfare** they make

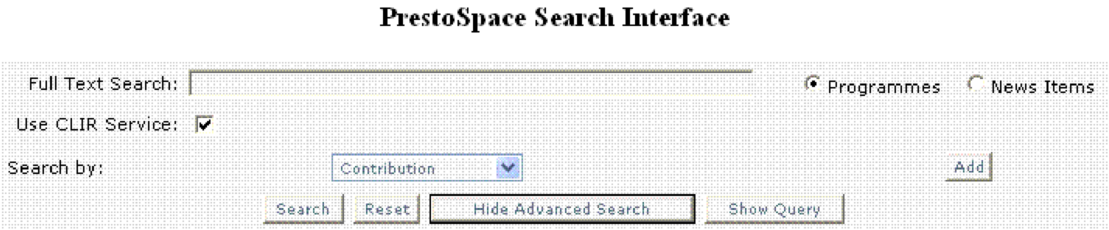
Main Category: England
Secondary Category: Manchester

BBC 10 O%27Clock News BBC News 22:00 Bullettin — Publication Date: 2005/07/06 - Duration: 0h 0m 31s

Figure 27 : Results of using the CLIR service

Advanced Search

Clicking on the Advanced Search button the user can submit queries more complicated than a simple full text searching task.

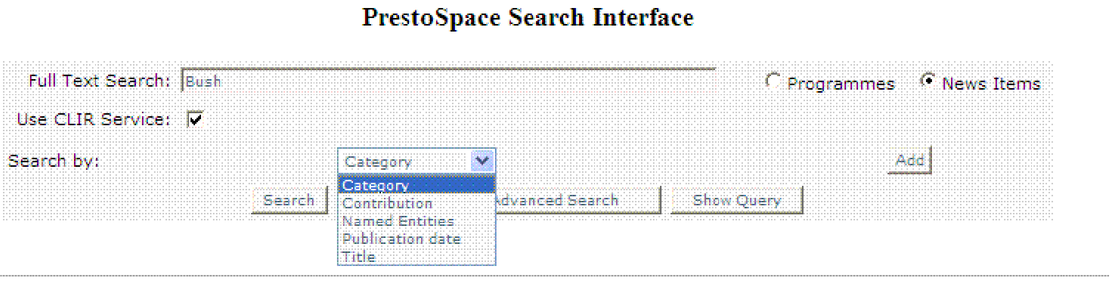


PrestoSpace Search Interface

Full Text Search: ● Programmes ○ News Items
Use CLIR Service: ☑
Search by: Contribution

Figure 28 : Advanced Search

The user can use filters searching by Category (only for News items), Contributions, Named entities, Publication Date, Publication Service (only for Programmes) and title.

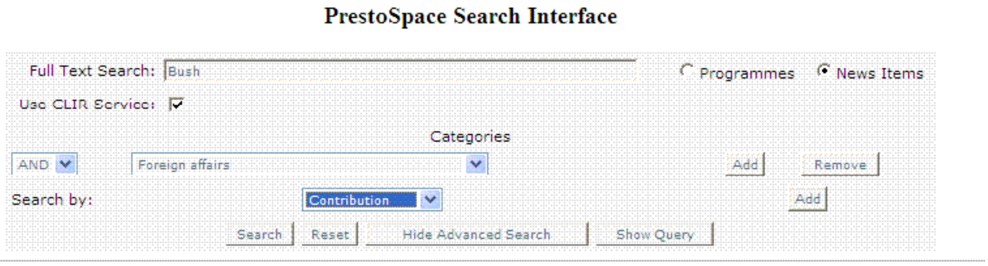


PrestoSpace Search Interface

Full Text Search: Bush ○ Programmes ● News Items
Use CLIR Service: ☑
Search by: Category

Figure 29 : Filters of the queries in the Advanced Search

Clicking on the ‘Add’ button the filter is added to the query. It is also possible to make logical operation on the filters of the same type (AND/OR):



PrestoSpace Search Interface

Full Text Search: Bush

○ Programmes ● News Items

Use CLIR Service: ☑

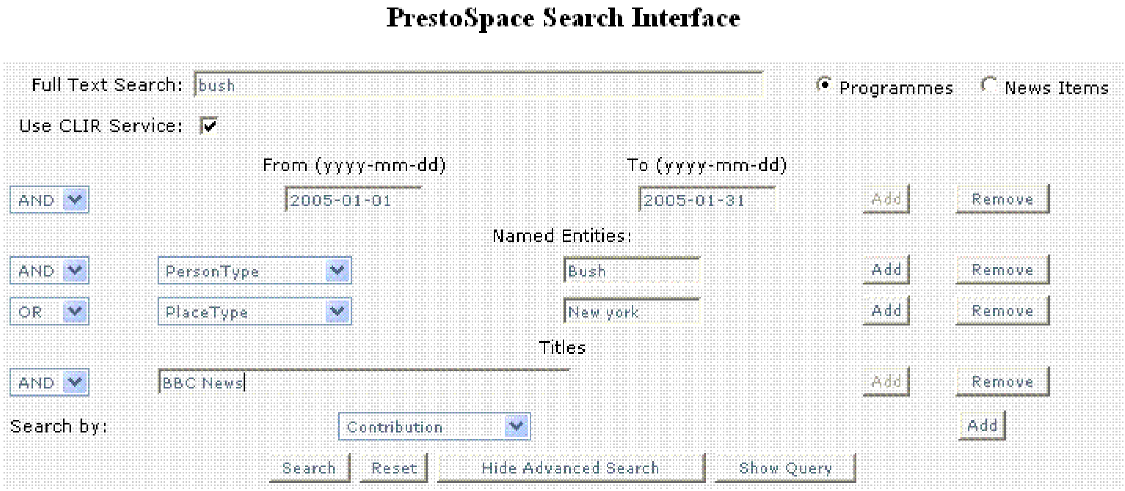
Categories

AND | Foreign affairs

Search by: Contribution

Figure 30 : Filter of the query

In the following Figure it is selected to search any BBC News that contains the word ‘bush’, published in the January of 2005, containing the Person ‘Bush’ OR the Place New York:



PrestoSpace Search Interface

Full Text Search: bush

● Programmes ○ News Items

Use CLIR Service: ☑

From (yyyy-mm-dd) | To (yyyy-mm-dd)

AND | 2005-01-01 | 2005-01-31

Named Entities:

AND | PersonType | Bush

OR | PlaceType | New york

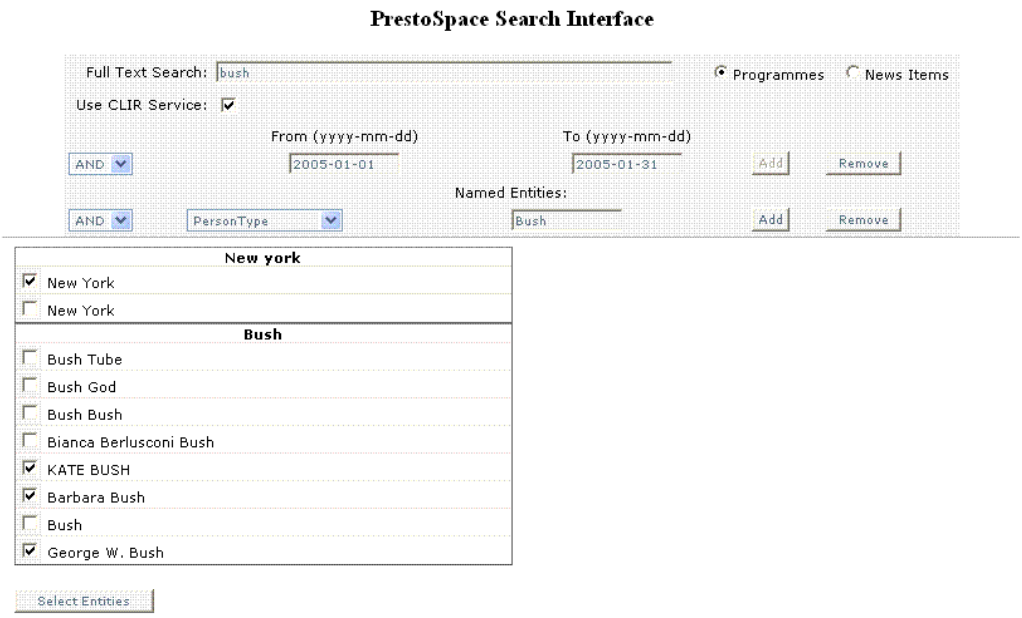
Titles

AND | BBC News

Search by: Contribution

Figure 31 : Example of an Advanced Search

Clicking on the Search button, the user can select the entities to be inserted in the query:



PrestoSpace Search Interface

Full Text Search: bush ● Programmes ○ News Items

Use CLIR Service: ☑

AND From (yyyy-mm-dd) 2005-01-01 To (yyyy-mm-dd) 2005-01-31

AND PersonType Named Entities: Bush

New york

☑ New York

□ New York

Bush

□ Bush Tube

□ Bush God

□ Bush Bush

□ Bianca Berlusconi Bush

☑ KATE BUSH

☑ Barbara Bush

□ Bush

☑ George W. Bush

Figure 32 : Named Entities of the query

PART C: Conclusions

8 Licensing

The software product described in this deliverable is a prototype in an advenced state of implementation. In order to engineer this software, an improvement of the feedback machinery is needed.

9 Bibliography

[1] R. Basili M. Cammisa, A. Gliozzo, Integrating Domain and Paradigmatic Similarity for Unsupervised Sense Tagging, Proceedings of the European Conference on Artificial Intelligence, Riva del Garda, (Italy), 2006.

[D.15.4] Content Analysis Tools.

[D.15.5] Cross-Linguistic IE Tools Analysis.

[D.15.6] Semantic Interpretation Tools.

[D.16.2] Conceptual Search.

[D.16.4] Delivery Models.

[D.18.1] The Documentation Platform for the MAD Factory

[D.18.3] The Turnkey System.

[D.19.0.1] External and Internal Models and Protocols for the PrestoSpace Factory.

[D.19.0.2] The PrestoSpace Orchestrator (PSO).

Annex 1 – D18.1 Glossary

Term	Definition
CLIR	Cross Language Information retrieval. It is a system for submitting a query using a language and retrieve results of the corresponding terms in another language, i.e: searching for 'sun' and finding also 'sole' (Italian word for 'sun').
EDOB	Editorial Object: the xml document describing the AV file and its related metadata.
MAD	Metadata Access and Delivery
MPEG-7	MPEG-7 (ISO/IEC 15938, formally named "Multimedia Content Description Interface") is a standard for describing multimedia content, independent of the encoding of the content, and allows different levels of granularity of the description. MPEG-7 has been designed to support a broad range of applications. MPEG-7 descriptions can be represented either as XML (textual format, TeM) or in a binary format (binary format, BiM).
Named entity	In the expression *named entity*, the word *named* restricts the task to those entities for which one or many rigid designators, as defined by Kripke, stands for the referent. For instance, the *automotive company created by Henry Ford in 1903* is referred to as *Ford* or *Ford Motor Company*. Rigid designators include proper names as well as certain natural kind terms like biological species and substances.
RSS	**RSS** (an acronym for Really Simple Syndication) is a family of web feed formats used to publish frequently updated digital content, such as blogs, news feeds or podcasts. RSS is analogous to a table of contents. An RSS "feed" provides a table of contents for a site's content for a certain period of time; it does not provide the content itself, but links to the content. RSS is useful because it helps aggregate lots of content into an easily accessible place.