Deliverable D18.3

Notification of Delivery of the Turnkey System

DOCUMENT IDENTIFIER	PS_WP18_EURIX_D18.3_TurnkeySys_v2.0
DATE	08/01/2008
ABSTRACT	This document is a notification of delivery for the Turnkey system. The turnkey system is a lightweight system specifically tailored for small size archives, and implements a complete MAD unit comprising the functionalities of both the Documentation Platform and the Publication Platform
KEYWORDS	metadata, web services, modular components, digitisation, metadata extraction, multimedia data access, multimedia data delivery, software integration, work flow management system
WORKPACKAGE / TASK	WP18
AUTHOR, COMPANY	W. Allasia, A. Damiani, S. Ridolfi, F. Toscano, M. Vigilante, euriX Group
NATURE	Prototype
DISSEMINATION	Public

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
1.2	2007-06-24	Release Candidate	Living	Confidential
1.3	2007-11-16	Release Candidate	Living	Confidential
1.4	2008-01-08	Final release	Closed	Confidential
2.0	2008-02-25	Document made public	Closed	Public

Contents Table

1. Introduction 3
1.1. Scope of this Document 3
1.2. Executive Summary 3
PART A: Architecture and Implementation of the Turnkey System 9
2. The MAD Platform 9
2.1. Architecture of MAD 10
3. The Turnkey System 13
3.1. The Documentation Platform 14
3.2. The Publication Platform 18
3.3. Communication between the Documentation Platform and the Publication Platform: the Export GAMP 22
3.4. Inputs, Processing, Outputs 22
3.5. The ADMIN Component 24
4. Implementation of the Turnkey System 27
4.1. Implementation of the Documentation Platform 27
4.2. Implementation of the Publication Platform 30
PART B: Utilization of the Turnkey System 32
5. How to use the Turnkey System 32
5.1. How to use the ADMIN component 33
5.2. How to use the generic GAMP 35
5.3. How to use the Publication Platform's Interface 36
PART C: Conclusions 49
6. Licensing 49
7. Bibliography 49
8. Glossary 50

1. Introduction

1.1. Scope of this Document

This deliverable is about the MAD component of the PrestoSpace factory, in particular it is focused on a software component called *turnkey system*. The turnkey system is a lightweight system specifically tailored for small size archives, and implements a *complete* MAD unit comprising the functionalities of both the Documentation Platform and the Publication Platform.

The tools here described are part of the MAD Unit. After a brief recall of the architecture of the Metadata Access and Delivery component (MAD), the turnkey system is described in detail.
This deliverable is a part of a three-piece product, which also includes the deliverables **[D18.1]** and **[D18.2]** describing the other MAD components: the Documentation Platform and the Publication Platform, respectively. As mentioned here above, the turnkey system can be seen as a unique component, containing a Documentation Platform and a Publication Platform, together with a “small” orchestrator (PSO), whose role is to coordinate the two components. Even if the functionalities of the Documentation Platform and the Publication Platform are also described in deliverables **[D18.1]** and **[D18.2]**, this document is self-contained.

This document is made up of three main parts. The first part (part A) recalls the architectures of MAD, which is implemented by the unique component called turnkey system. The second part (part B) gives detailed information on how to use the turnkey system. The third part (part C) concludes the deliverable by presenting information about legal aspects and licensing issues about the usage of the software.

1.2. Executive Summary

Recently, broadcasters have rediscovered the value of their audiovisual archives. Moreover, recent researches have shown that approaches meant to the recovery and availability of archived materials may produce consistent cost savings in the overall programme production processes.
In order to achieve this goal, it is essential to adopt *metadata*.

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 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.

MAD stands for Metadata Access and Delivery.
The MAD Platform adopts a modular and extensible architecture. It consists of two different components:

a. The *Documentation Platform*
b. The *Publication Platform*

The MAD Platform receives digitised media (audio and video files) as input: these data are processed by the Documentation Platform, which returns different materials as key frames, camera motions and metadata. These materials are then indexed and published on a web server by the Publication Platform.

Broadcasters often need to digitised audiovisual archives of very big size, requiring sophisticated mechanisms of information retrieval. These broadcasters usually have their own Publication Platform, therefore they are interested in using the Documentation Platform only. The modular structure of the MAD Platform naturally satisfies the requirements of this kind of broadcasters.

In contrast, many other potential users need to digitised their own audiovisual archives, having very small size. As an example, a Department of a University could need to digitise some e-learning lessons. As another example, a soccer fan could be interested in digitising and retrieving a set of a hundred of matches stored on VHS tapes.
This kind of users essentially have the following requirements:

- they aim to digitise and access small size archives
- they do not have their own Publication Platform, i.e. typically they need a *complete* software performing the functionalities of the MAD Platform.

In order to satisfy the requirements of this kind of users, a software component called **Turnkey System** has been developed. The Turnkey System is a lightweight system specifically tailored for small size archives. It is made up of both the Documentation and the Publication Platform with customized features, that is to say it is a fully automatic system for content enrichment and web publishing/searching. Big size archives should use subparts of the Turnkey System because they have their content management systems and web search and publication features.

The Turnkey System is represented in Figure 1.

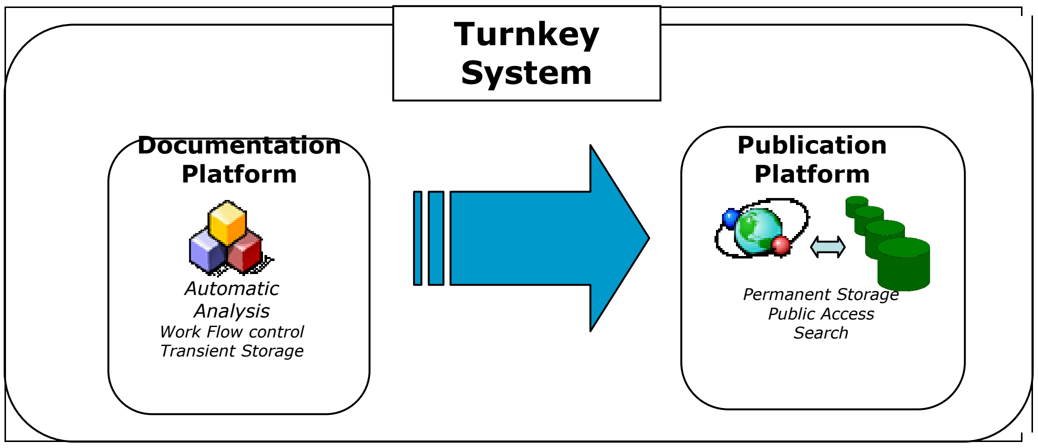


Figure 1: the Turnkey System.

The *Documentation Platform* is made up of a core component, called *Core Platform*, and a set of pluggable software processors named *GAMP*s, where GAMP stands for Generic Activity MAD Processor. A GAMP is a software component that extracts the metadata from the digitised material.

The Core Platform offers the following main services:

- Workflow management service, responsible for starting processes in the right order and for resolving dependencies between GAMPs;
- Interaction with the Essence and Metadata Storage (EMS) system, which stores the audiovisual material sources and the associated metadata;
- Interaction with the Concurrent Versioning System, tracking every change to the metadata operated by the GAMPs, built on a standard CVS engine;
- Delivery of enriched metadata and related material created by the GAMPs within the Documentation Platform.

EMS and CVS are two components of the “small” PSO of the Turnkey system. They are used in order to manage the storage of materials within the factory and track different versions of these materials, respectively.
The main features of the Documentation Platform can be represented as shown in Figure 2.

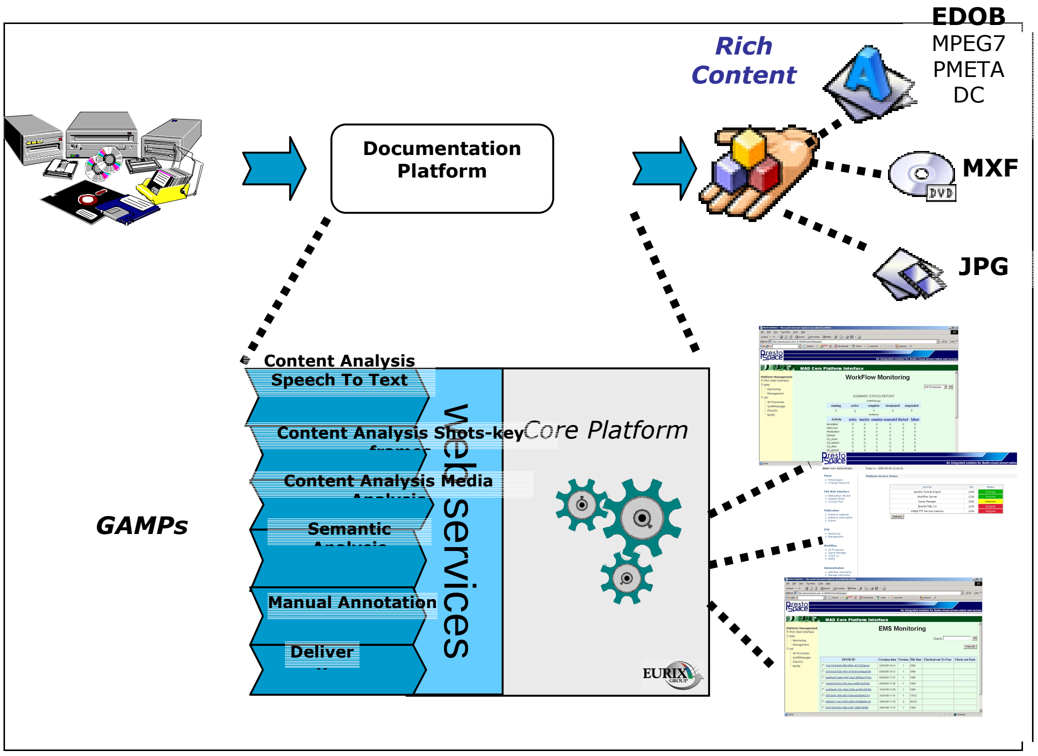


Fig. 2: the Documentation Platform

The overall services offered by the Core Platform are available through web services interfaces based on SOAP. Using web services, every GAMP polls the Core Platform asking for a job, and then submits the produced metadata and notifies the completion of its work to the Workflow Manager. By using web services, GAMPs can be implemented by using any programming language supporting SOAP and web services protocols.

The architecture of the Documentation Platform has the following peculiarities:

- it is modular, since GAMPs can interact with the Core Platform even being totally different in implementation details and functionalities;
- it is extensible, in the sense that it is easy and natural to insert a new GAMP;
- it is platform independent, since the Core Platform itself is implemented in Java, therefore portable to several operating systems;
- it is characterized by a multi-tier distribution, in the sense that every GAMP can be installed on a different physical system, provided that a network link to the Core Platform is available.

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 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 3.

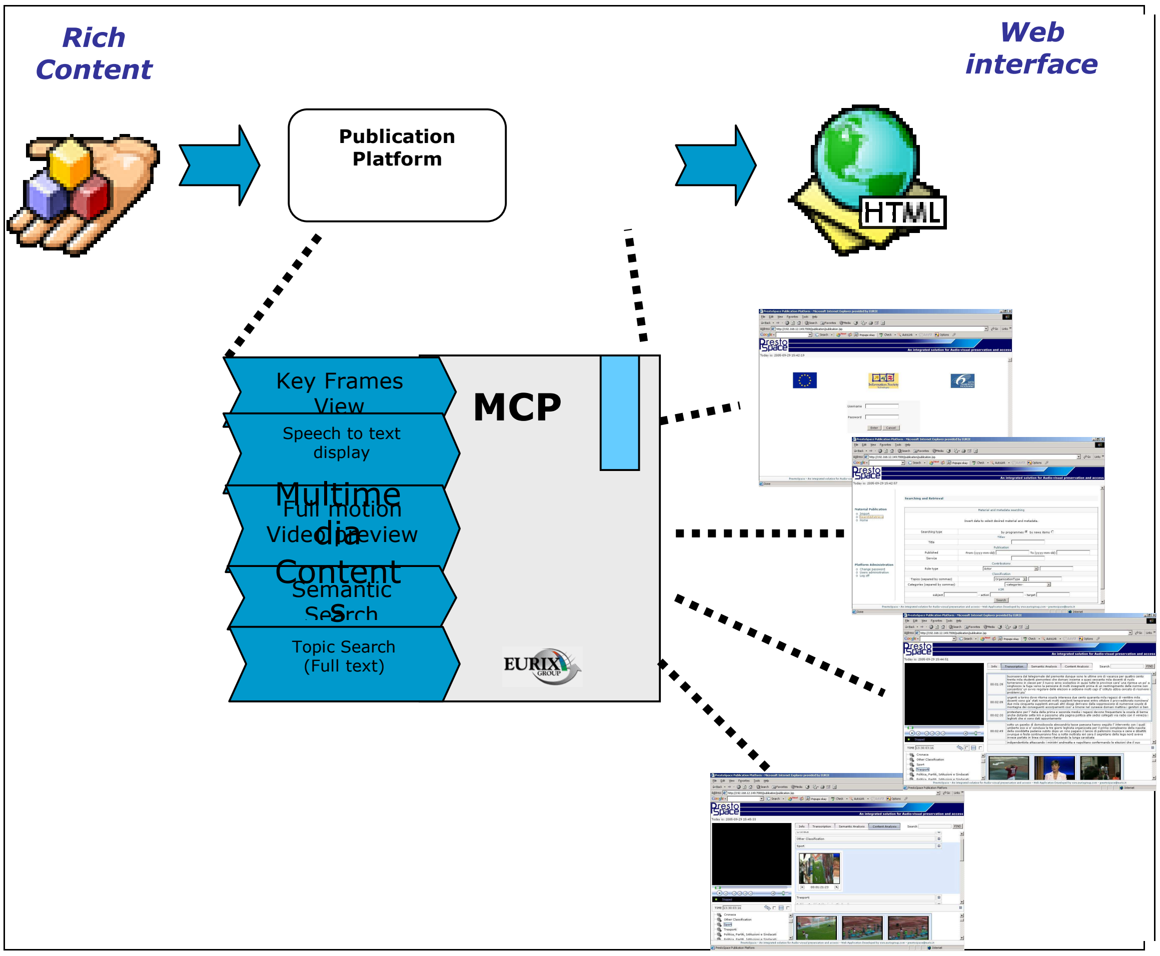


Fig. 3: the Publication Platform

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

As mentioned above, this document presents the architecture of the turnkey system, describing its components in detail. The Documentation Platform and the Publication Platform are also described in deliverables **[D18.1]** and **[D18.2**], respectively; however,

this document is self-contained, then users interested in the turnkey system can restrict their attention to this deliverable only.
The structure of the document can be summarized as follows: first, we recall the architecture of MAD, introducing the turnkey system, which is a fully implementation of the MAD functionalities (part A); second, we describe how to use the turnkey system (part B). A third part (part C), containing a glossary and some information about licences, concludes this deliverable.

PART A: Architecture and Implementation of the Turnkey System

2. The MAD Platform

In the recent years, broadcasters have rediscovered the value of their audiovisual archives. Moreover, recent researches have shown that approaches meant to the recovery and availability of archived materials may produce consistent cost savings in the overall programme production processes.
In order to achieve this goal, it is essential to adopt *metadata*.

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.

Researches within the PrestoSpace project have determined that the required information for a typical audiovisual archive exploitation processes can be partitioned in the following fundamental classes:

- *Identification information*, such as titles, credits, and programme publication information;
- *Editorial parts of information*, i.e. information about the relevant editorial sub-items of a programme, such as news items;
- *Content-related information*, such as text of speech, descriptions, and visual low level descriptive features;
- *Enrichment information*, coming from external sources related to the programme content.

The data model adopted, representing the above classes, together with a data format carrying all the entities and relations of it, consists of a single XML-based document format, resulting from the combination of MPEG-7 and P_META. More in detail, MPEG-7 has been used thanks to its powerful temporal segmentation tools and for its comprehensive set of standard audiovisual descriptors, whereas P_META has been adopted in order to capture information structures for identification, classification and publication-related features of a programme.

In Figure 4 a schematization of the adopted document format is presented. Ad hoc data structures, introduced to represent those information not supported neither by MPEG-7 nor by P_META, are emphasized. In addiction, it is worth noticing that the SMPTE UMID standard has been adopted in order to capture the unique identification of the instances of audiovisual material all throughout the platform, namely original media, digitally re-mastered media, and all the material generated by the documentation process (e.g. key frames).

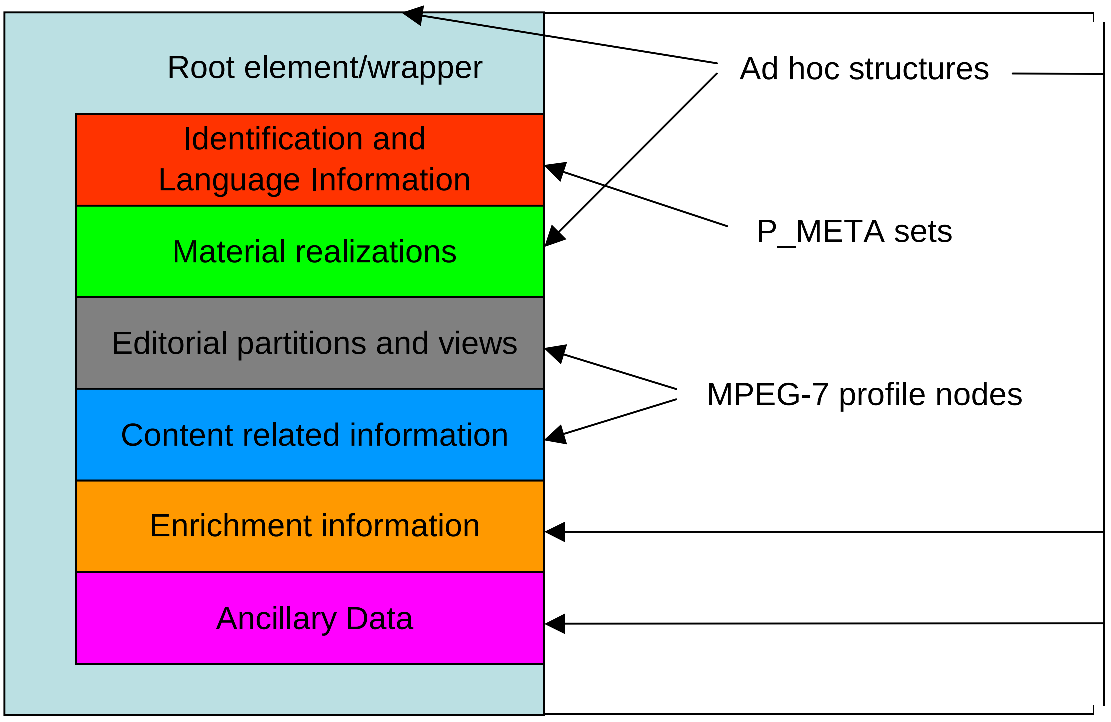


Figure 4: a schema of the MAD document format.

Concerning the PrestoSpace project, given the audiovisual items produced by the preservation and restoration units, we need to develop a software component able to *document* and *deliver* them. This component is called **MAD**, standing for **Metadata Access and Delivery**. The MAD component provides the software modules for documenting and delivering audiovisual information, and it is made up of pluggable GAMPs (Generic Activity MAD Processor) connected to a core Platform for automatic features extraction.

2.1. Architecture of MAD

As mentioned in the Introduction of this document, the MAD 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.

In order to achieve the above goals, the MAD platform adopts a modular, extensible architecture. In detail, it receives in input the digitised media (video and audio files) produced by the Preservation and Restoration units, then it produces several materials,

like key frames, camera motions, and metadata, as output. These materials are then available for information retrieval.
The architecture of the MAD Platform is schematized in Figure 5 below.

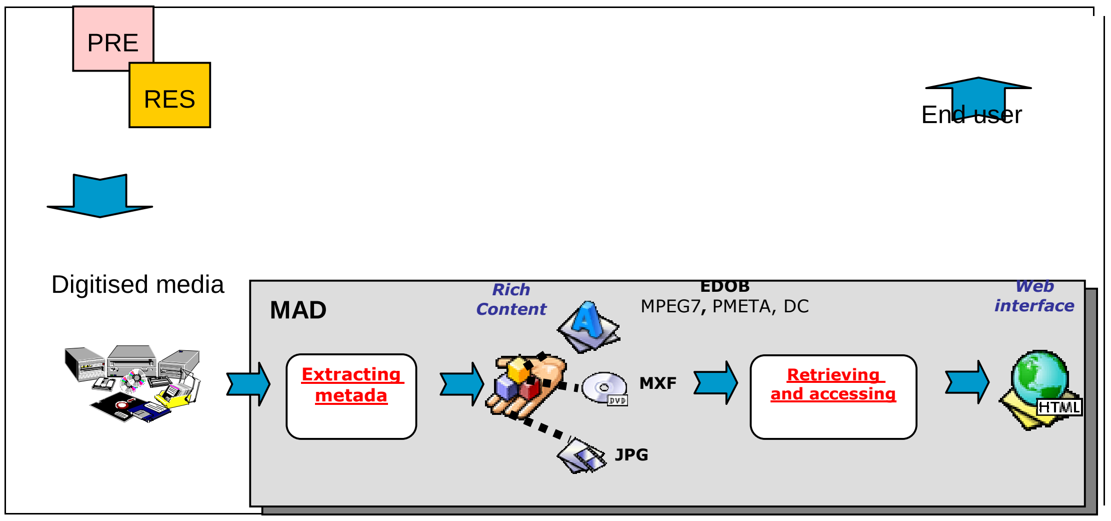


Figure 5: the architecture of MAD.

As cited here above, the MAD Platform adopts a modular and extensible architecture. It consists of two different components:

- The *Documentation Platform*
- The *Publication Platform*

taking care of its two fundamental goals (the red and underlined text in white boxes of Figure 5). The resulting architecture is presented in Figure 6.

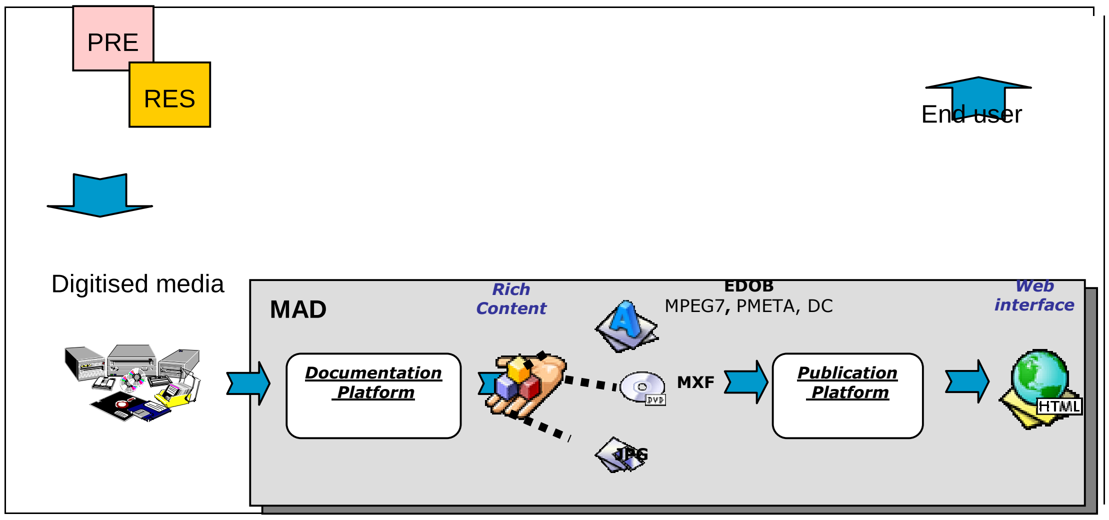


Figure 6: the architecture of MAD (2).

The MAD Platform receives digitised media (audio and video files) as input: these data are processed by the Documentation Platform, which returns different materials as key frames, camera motions and metadata. These materials are then indexed and published on a web server by the Publication Platform. The communication between the Documentation Platform and the Publication Platform is implemented by means of an *Export GAMP*. This export GAMP is responsible of inserting suitable information in the database and of creating an index for the KIM engine. Moreover, it creates the html documents that will be exposed by the Publication Platform. Intuitively, the export GAMP is the component used to implement the communication/interaction between the Documentation and the Publication Platforms.

3. The Turnkey System

In the Introduction we have observed that broadcasters often need to digitised audiovisual big size archives, requiring sophisticated mechanisms of information retrieval. These broadcasters usually have their own Publication Platform, therefore they are interested in enriching the multimedia contents by adding metadata, i.e. they are interested in the Documentation Platform only. The modular structure of the MAD Platform allows to satisfy the requirements of this kind of users, in the sense that one can only make use of the Documentation Platform module, together with its own system for retrieving enriched information.

It is worth noticing that many other users need to digitise very small size archives. Private audiovisual archives and archives of resources of an academic institute typically have a small size, not comparable to the information available to a broadcaster (even local). Users having the goal of managing small size archives essentially do not have their own Publication Platform, i.e. typically they need a *complete* software performing the functionalities of the MAD Platform.

In order to satisfy the requirements of this kind of users, a software component called **Turnkey System** has been developed. The Turnkey System is a lightweight system specifically tailored for small size archives. It is made up of both the Documentation and the Publication Platform with customized features, that is to say it is a fully automatic system for content enrichment and web publishing/searching. Big size archives should use subparts of the Turnkey System because they have their content management systems and web search and publication features.

The Turnkey System can be represented as shown in Figure 7.

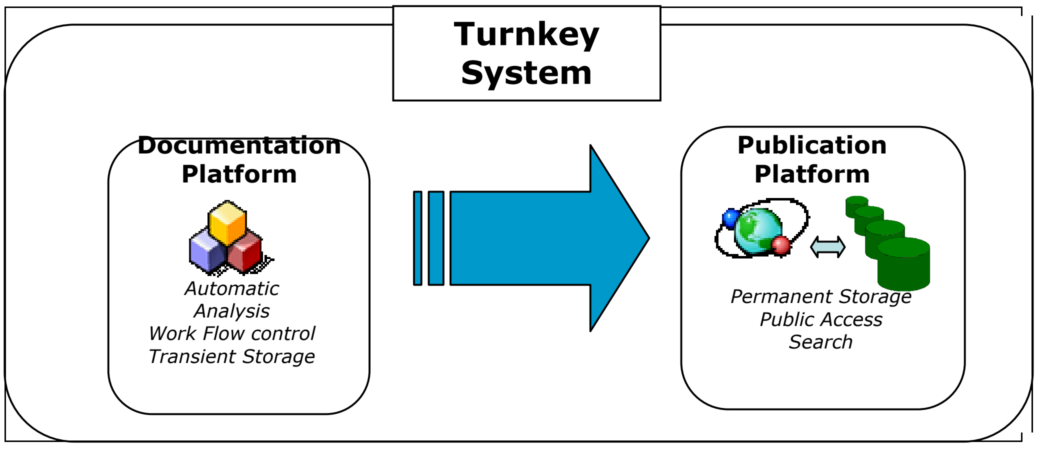


Fig. 7: the Turnkey System.

In the rest of this section we will focus our attention on the two main components of the turnkey system, namely the Documentation Platform and the Publication Platform. These

]

components are also described in specific deliverables **[D18.1]** and **[D18.2]**. Furthermore, in section 3.3 we discuss about the main differences between the turnkey system's architecture and the one of a "standard system", comprising the Documentation Platform and the Publication Platform coordinated by an essential (restricted) implementation of a PSO.

3.1. The Documentation Platform

The documentation Platform is responsible of integrating and bringing together all the components provided within the MAD area, hence it is the core module for building up a MAD System. It provides the essential features for exchanging data and metadata and for running the tools provided by the partners involved in the Area (pull logic).

The Documentation Platform is made up of a core component, called *Core Platform*, and a set of pluggable software processors called *GAMP*s, where GAMP stands for Generic Activity MAD Processor. A GAMP is a software component that extracts the metadata from the digitised material.

As mentioned in the Introduction, the Core Platform offers the following main services:

- It implements a *Workflow management service*, which is responsible for starting processes in the right order and for resolving dependencies between GAMPs;
- It interacts with the component called Essence and Metadata Storage (EMS) system, which stores the audiovisual material sources and the associated metadata;
- It interacts with the component called Concurrent Versioning System, tracking every change to the metadata operated by the GAMPs, built on a standard CVS engine;

As mentioned in the above section dedicated to the MAD architecture, the enriched metadata and related materials created within the Documentation Platform are then delivered by the Publication Platform.
The main features of the Documentation Platform can be represented as shown in Figure 8.

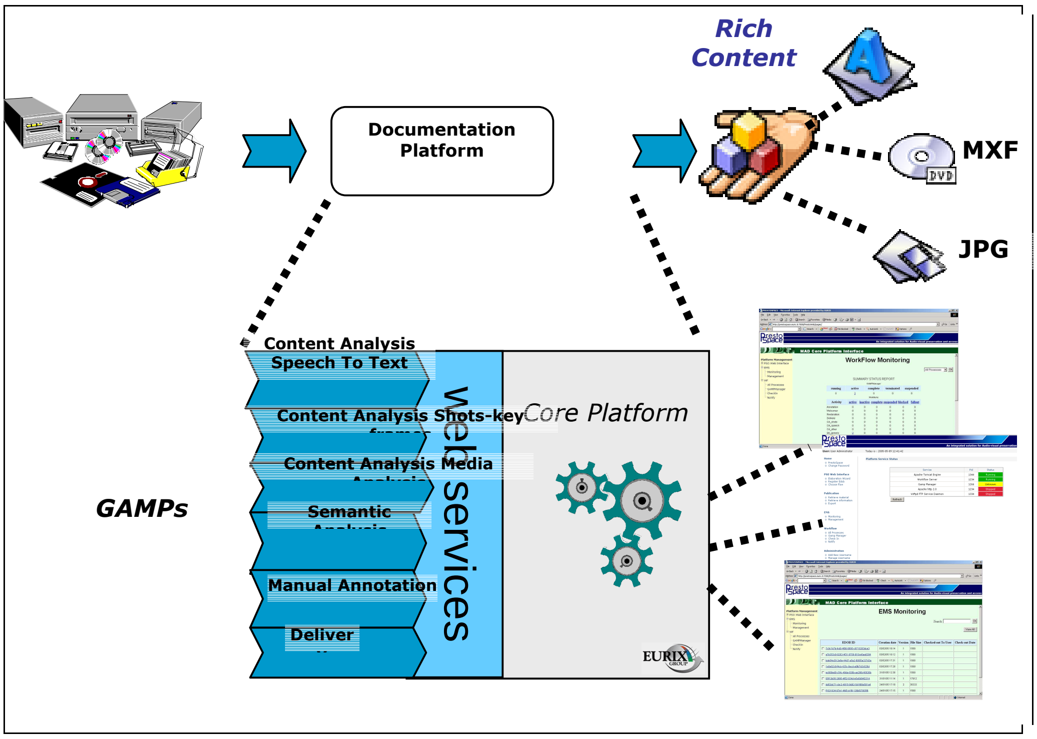


Figure 8: the Documentation Platform

The overall services offered by the Core Platform are available through web services interfaces based on SOAP. Using web services, every GAMP polls the Core Platform asking for a job, and then submits the produced metadata and notifies the completion of its work to the Workflow Manager. By using web services, GAMPs can be implemented by using any programming language supporting SOAP and web services protocols.

The architecture of the Documentation Platform has the following peculiarities:

- it is modular, since GAMPs can interact with the Core Platform even being totally different in implementation details and functionalities;
- it is extensible, in the sense that it is easy and natural to think about the insertion of a new GAMP;
- it is platform independent, since the Core Platform itself is implemented in Java, therefore portable to several architectures and operating systems;
- it is characterized by a multi-tier distribution, in the sense that every GAMP can be installed on a different physical system, provided that a network link to the Core Platform is available.

In the rest of this section we analyze each component of the Documentation Platform in more detail.

3.1.1. The Core Platform

As mentioned above, the Documentation Platform comprises a main component called **Core Platform**. The Core Platform essentially offers a Workflow management service.

The **Workflow management service** is the component used in order to manage the activities of the different GAMPs. In detail, the Workflow management service comprises a *queue for each GAMP* within the platform. Every GAMP polls the Core Platform (i.e. its own queue), asking for a job and related resources: when something is available, then the GAMP starts its process. When it concludes its work, the GAMP notifies the completion to the Core Platform. The Workflow management service is build up using the open source component Open Flow (Zope).

When a GAMP is scheduled by the Core Platform to perform its process, it usually needs to retrieve the EDOBs and/or the files for the elaboration. To this aim, the GAMP contacts the Core Platform, which is responsible to contact the Essence and Metadata Storage component, and retrieves those information.

The **Essence and Metadata Storage (EMS) system** stores the materials on the file system, and tracks their location by means of a relational database. It is possible to maintain several copies of the same material, even located on different machines accessible via suitable protocols (HTTP, FTP, SMB, file, and so on).

It could be the case that a GAMP operates in a wrong way, thus wrongly updating files and/or metadata stored in the EMS. In order to avoid this situation, therefore ensuring that a consistent, sound version of the set of information is always available to the factory, the Core Platform also interacts with the component called Concurrent Versioning System.
The **Concurrent Versioning System (CVS)** tracks every change to the metadata that takes place during the execution of the GAMPs. It is build on a standard CVS engine. In this way, if an unacceptable update has been performed by a GAMP, then the Core Platform asks the CVS to perform a rollback to a consistent version of the system.

It is worth noticing that the EMS and the CVS are components directly managed by a specific, limited implementation of a PrestoSpace Orchestrator (PSO, see **[D19.0.2]** for details), which can be intuitively seen as a coordinator of the activities of the Turnkey system. The Core Platform of the Documentation Platform only interacts (via web services) with EMS and CVS as mentioned above.

Let us conclude this section with a brief remark. It is worth noticing that the interaction between GAMPs and the Core Platform is based on web services. This implies that GAMPs can be developed by using different programming languages supporting SOAP and deployed on totally different platforms and operating systems.

3.1.2. Generic Activity MAD Processors (GAMPs)

The Documentation Platform is able to connect the components provided by the partners involved in the MAD area, namely to so called GAMPs (Generic Activity MAD Processor). The GAMPs are software units that extract metadata from the digitised materials. The Core Platform maintains a queue in the workflow for every GAMP, which will poll it in order to become aware of any activity to be done. In order to achieve their goals, the GAMPs ask the Core Platform for the materials and the related (associated) metadata produced up to the request time.

The Documentation Platform makes use of three different kinds of GAMPs, namely:

- Content Analysis tools
- Semantic Analysis tools
- Manual Annotation tools

The basic idea of the turnkey system is that it can be executed even on a small machine; therefore, the turnkey system only makes use of some GAMPs. Obviously, if the turnkey system is installed on a small machine, only the GAMPs performing a “lightweight” metadata extraction are used.
Here below is the list of actually implemented GAMPs[1]:

- Welcomer: demux MXF (RAI)
- Semantic Analysis (University of Sheffield, University of TorVergata)
- Annotation GAMP (JRS)
- Shot boundary detection tools (RAI, content analysis)
- Key frame detection and extraction tools (JRS, content analysis)
- Stripe Images extraction tools (JRS, content analysis)
- Camera motion detection tools
- Visual activity extraction tools
- Speech to text transcription tools (RAI, content analysis)
- Audio structuring and segmentation tools (RAI)
- Multimedia structure detection tools
- Editorial parts segmentation tools (University of Sheffield)
- Reference video clips detection tools
- Low-level visual features extraction tools

[1] It is worth noticing that any kind of new GAMP can be easily added in the future. In order to insert a new GAMP it is sufficient to add a new process queue to the Workflow engine of the Core Platform, as discussed in section 3.3.

Among the others, the Annotation GAMP allows the user to enrich an audiovisual item by means of a manual addiction of metadata.
For further details on all the GAMPs developed within the PrestoSpace project, we remind the reader to deliverables **[D15.4]**, **[D15.5]**, and **[D15.6]**.

3.1.3. Generic GAMP

In order to add a new GAMP to the Documentation Platform architecture, e.g. performing the extraction of further metadata, two different alternatives are available:

1. the GAMP can be implemented following the guidelines about its functioning. Obviously, the new GAMP must fill all these specifics, namely it has to publish all needed web services and it has to implement the operations required by any GAMP (for a detailed description, see section 4.2 of this Deliverable);
2. the GAMP can be implemented by means of a **Generic GAMP**.

The Documentation Platform provides a Generic GAMP, a Java component that can be used in order to build a GAMP component in an easy way. In order to add a new GAMP to the Documentation Platform, the following steps have to be performed:

- the new metadata extracted by the GAMP must be MPEG7 compliant;
- in order to deliver the new kind of information, the Publication Platform has to be modified, taking this new metadata into account;
- a new queue related to the new GAMP must be added to the Documentation Platform.

Obviously, one can think of replacing an existing GAMP with another one, producing the same kind of matadata. This could be the case in which a more efficient implementation of the GAMP is provided. In this case, one can replace the existing GAMP with the new generic GAMP, making use of the same queue and producing a specific metadata, rather than assignining a new queue to the GAMP within the Platform.

A broader discussion on how to use the generic GAMP is presented in section 6.2 of this deliverable.

3.2. The Publication Platform

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

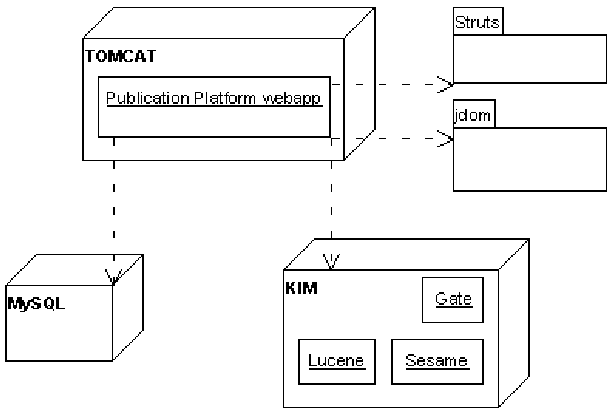


Figure 9 : The Publication Platform - architecture

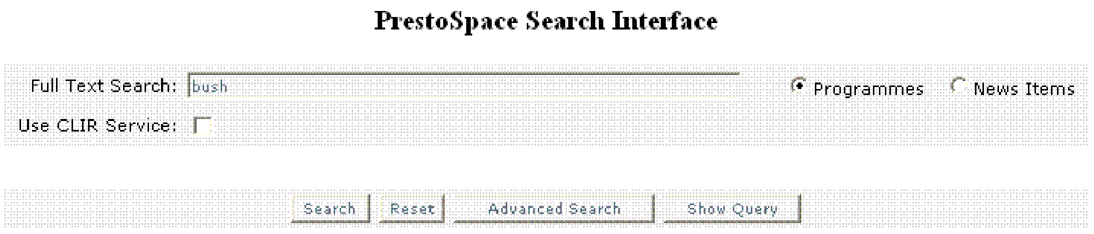
The platform architecture is based on three main components: a web application to allow user interaction; a database (MySQL) to store data about available programmes and so to make easy searching and selections; the KIM platform (provided by Ontotext) to perform semantic functionality through semantic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

3.2.1. The 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 10:


PrestoSpace Search Interface

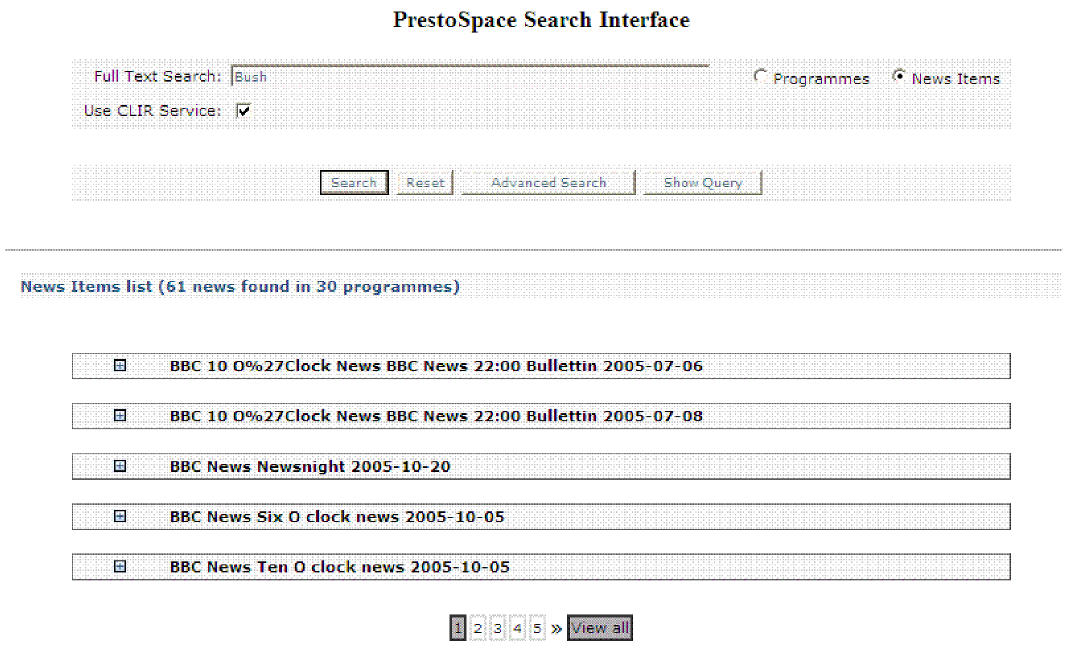
Full Text Search: bush [✓ Programmes] News Items

Use CLIR Service: ☐

Figure 10: 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 11) 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

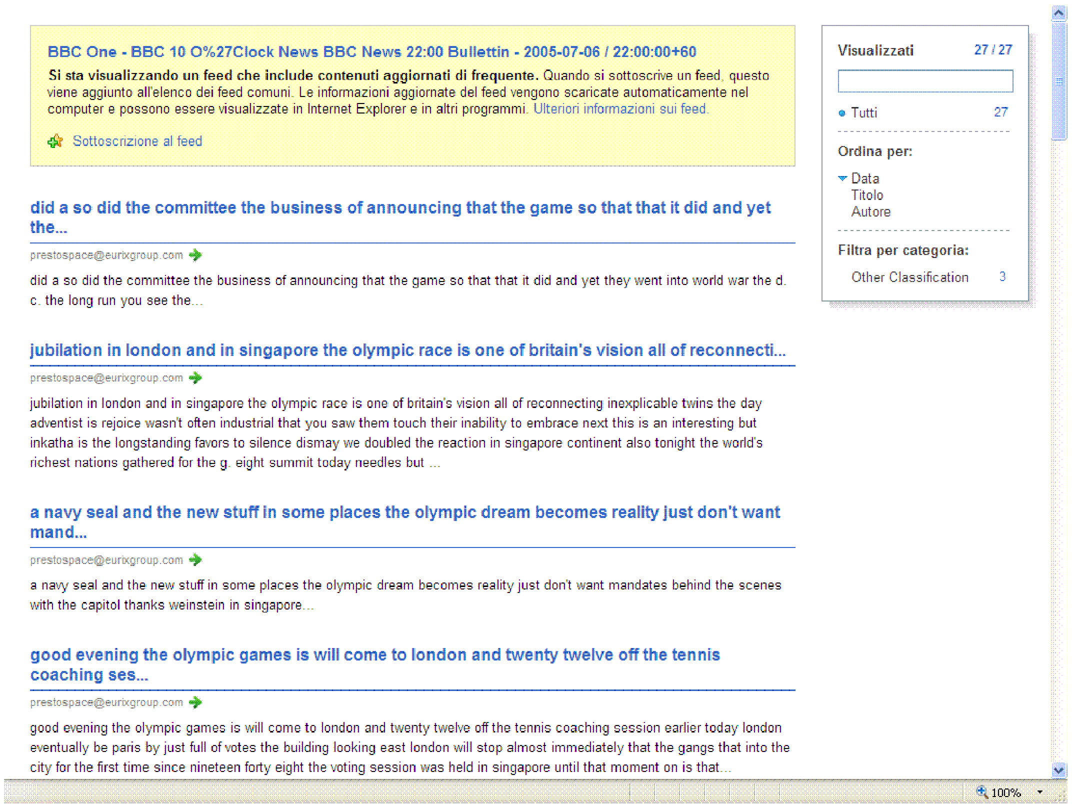
⊞ BBC News Six O clock news 2005-10-05

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

Figure 11 : the list containing the results of a query

3.2.2. The 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 12).



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 12: RSS export feature

3.3. Communication between the Documentation Platform and the Publication Platform: the Export GAMP

The Documentation Platform also contains an Export GAMP, which is responsible of the communication between the Documentation and the Publication Platform. In detail, the Documentation Platform makes use of the Export GAMP in order to forward to the Publication Platform the results produced by the Documentation Platform.

The Export GAMP essentially performs the following activities:

- it populates the database used by the Publication Platform
- it generates suitable directories containing javascript files used by the Publication Platform
- it creates suitable indexes for retrieving information in the Publication Platform.

All the above activities are performed in order to transfer information (i.e. the enriched metadata) produced by the Documentation Platform to the Publication Platform.

3.4. Inputs, Processing, Outputs

In this section a schematization of inputs, processings and outputs of the Turnkey System is presented.

3.4.1. Inputs

Input data:
1. Editorial Object Identification
2. Preservation and Legacy metadata (data provided during the Preservation phase)
3. Digitised Material. Essence submission from preservation factory (PRE). The essences are expected to be published on file, samba, ftp, http, https (so far implemented or others) protocols by the preservation system. They can be in some of the planned formats (MXF as default).

Input standard

4. Metadata model (defined in WP15): as default standard in the MAD platform we can assume that EDOB is the reference one (EDitorial Object) and Dublin Core or PMeta can be used as well.
5. Delivery format definition:
   We can assume that the exchange format used for delivering data will be the same used internally. Actually WP15-16 will define it. As a starting reference we can use the EDOB schema.

Physical connections constraints

6. The Core Platform will publish its services on SOAP. More precisely it will publish them as web services on a wsdl interface. For using it, a system will need an http connection and API for xml/soap message marshalling/unmarshalling. For submitting/uploading essences, a system needs file/samba/ftp/http/s server for publishing every document it is planning to send to the platform.

3.4.2. Processing

Documentation platforms will process the EDOB schema internally and will deal with some well defined standards as MXF, MPEG7, PMETA, DC.
These are the default document formats the platform is expected to manage.
Furthermore the platform will handle the following protocols: file, samba, ftp, http, soap.

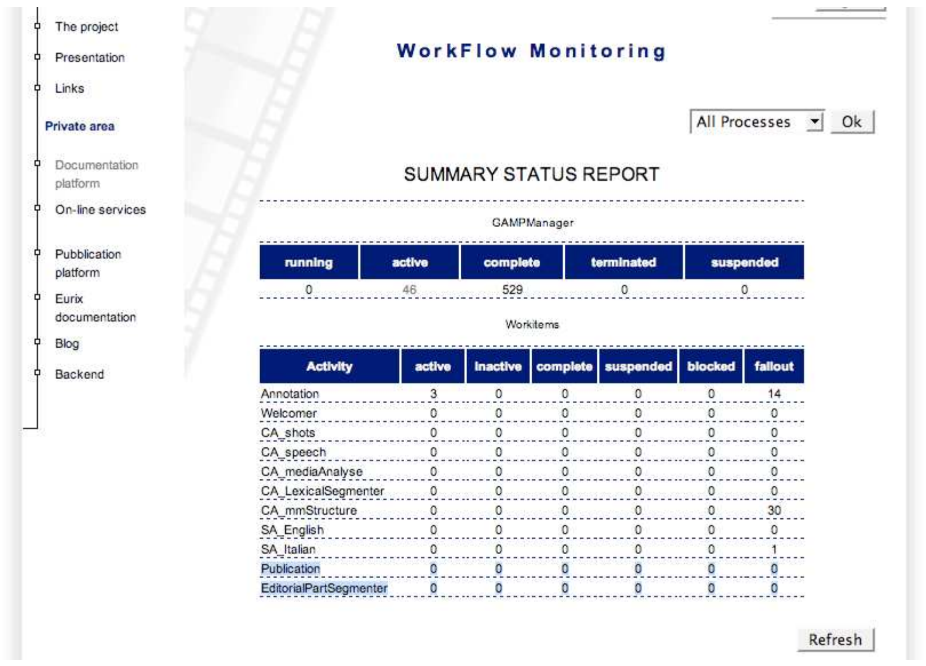
3.4.3. Outputs

The documentation platform will provide the enriched metadata and digitized material. As final outputs we have a complete export in some format WP16 has to identify. MXF and further attachments (as MPEG7 and other Content Analysis formats) are the input for the Publication Platform. The Publication Platform provides a web interface for searching and retrieving information produced by the Documentation Platform. This web interface is the output of the Turnkey system.

3.5. The ADMIN Component

The Turnkey System also comprises an additional component, called **ADMIN**, that can be used in order to manage the operations of the Paltform.
The ADMIN component consists in a web application, which allows the administrator to manage and control the work flow activities.

As an example, in Figure 13 the web page summarizing the status of the work flow system is presented:



SUMMARY STATUS REPORT

GAMPManager

running	active	complete	terminated	suspended
0	46	529	0	0

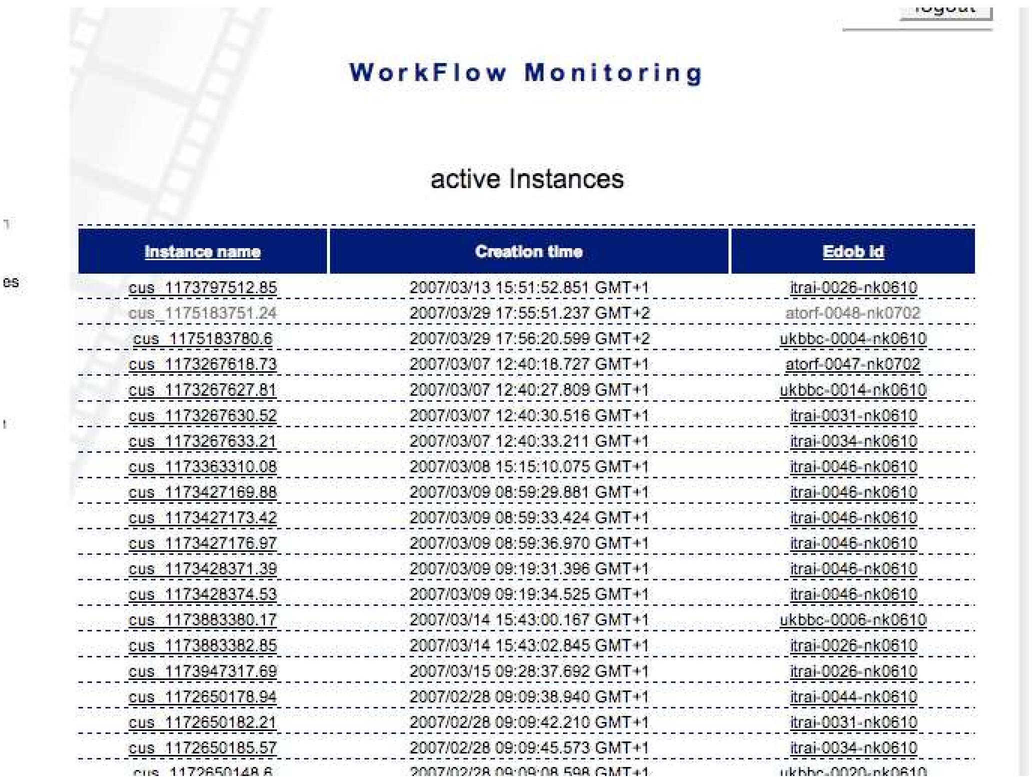
Workitems

Activity	active	inactive	complete	suspended	blocked	fallout
Annotation	3	0	0	0	0	14
Welcomer	0	0	0	0	0	0
CA_shots	0	0	0	0	0	0
CA_speech	0	0	0	0	0	0
CA_mediaAnalyse	0	0	0	0	0	0
CA_LexicalSegmenter	0	0	0	0	0	0
CA_mmStructure	0	0	0	0	0	30
SA_English	0	0	0	0	0	0
SA_Italian	0	0	0	0	0	1
Publication	0	0	0	0	0	0
EditorialPartSegmenter	0	0	0	0	0	0

Figure 13: the web page of the Work Flow Monitoring System.

The administrator can then obtain further information on a specific work item. For instance, in order to take a look at the active items, the administrator can click on the number of active items in the table called GAMPManager. The ADMIN interface will show the list of active instances being processed by the work flow manager; for each instance, three information are given:

- the identifier of the active instance;
- the date of creation
- the associated EDOB

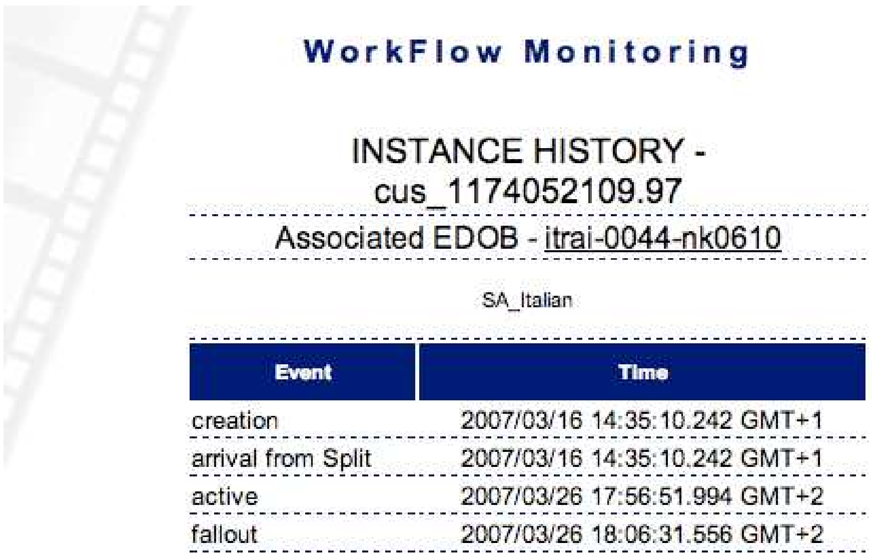

WorkFlow Monitoring

active Instances

Instance name	Creation time	Edob Id
cus_1173797512.85	2007/03/13 15:51:52.851 GMT+1	itrai-0026-nk0610
cus_1175183751.24	2007/03/29 17:55:51.237 GMT+2	atorf-0048-nk0702
cus_1175183780.6	2007/03/29 17:56:20.599 GMT+2	ukbbc-0004-nk0610
cus_1173267618.73	2007/03/07 12:40:18.727 GMT+1	atorf-0047-nk0702
cus_1173267627.81	2007/03/07 12:40:27.809 GMT+1	ukbbc-0014-nk0610
cus_1173267630.52	2007/03/07 12:40:30.516 GMT+1	itrai-0031-nk0610
cus_1173267633.21	2007/03/07 12:40:33.211 GMT+1	itrai-0034-nk0610
cus_1173363310.08	2007/03/08 15:15:10.075 GMT+1	itrai-0046-nk0610
cus_1173427169.88	2007/03/09 08:59:29.881 GMT+1	itrai-0046-nk0610
cus_1173427173.42	2007/03/09 08:59:33.424 GMT+1	itrai-0046-nk0610
cus_1173427176.97	2007/03/09 08:59:36.970 GMT+1	itrai-0046-nk0610
cus_1173428371.39	2007/03/09 09:19:31.396 GMT+1	itrai-0046-nk0610
cus_1173428374.53	2007/03/09 09:19:34.525 GMT+1	itrai-0046-nk0610
cus_1173883380.17	2007/03/14 15:43:00.167 GMT+1	ukbbc-0006-nk0610
cus_1173883382.85	2007/03/14 15:43:02.845 GMT+1	itrai-0026-nk0610
cus_1173947317.69	2007/03/15 09:28:37.692 GMT+1	itrai-0026-nk0610
cus_1172650178.94	2007/02/28 09:09:38.940 GMT+1	itrai-0044-nk0610
cus_1172650182.21	2007/02/28 09:09:42.210 GMT+1	itrai-0031-nk0610
cus_1172650185.57	2007/02/28 09:09:45.573 GMT+1	itrai-0034-nk0610
cus_1172650148.6	2007/02/28 09:09:08.598 GMT+1	ukbbc-0020-nk0610

Figure 14: a list of active instances.

The administrator can then require more information about the selected instance (Figure 15) or the associated EDOB (Figure 16).


WorkFlow Monitoring

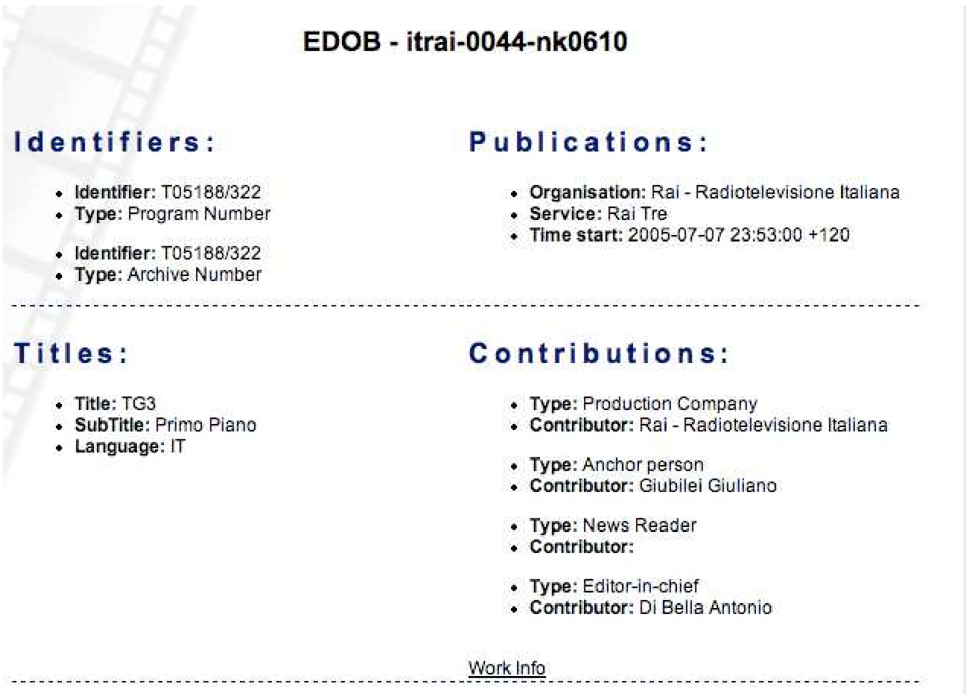
INSTANCE HISTORY - cus_1174052109.97

Associated EDOB - itrai-0044-nk0610

SA_Italian

Event	Time
creation	2007/03/16 14:35:10.242 GMT+1
arrival from Split	2007/03/16 14:35:10.242 GMT+1
active	2007/03/26 17:56:51.994 GMT+2
fallout	2007/03/26 18:06:31.556 GMT+2

Figure 15: information about an active instance



EDOB - itrai-0044-nk0610

Identifiers:

- **Identifier:** T05188/322
- **Type:** Program Number

- **Identifier:** T05188/322
- **Type:** Archive Number

Publications:

- **Organisation:** Rai - Radiotelevisione Italiana
- **Service:** Rai Tre
- **Time start:** 2005-07-07 23:53:00 +120

Titles:

- **Title:** TG3
- **SubTitle:** Primo Piano
- **Language:** IT

Contributions:

- **Type:** Production Company
- **Contributor:** Rai - Radiotelevisione Italiana

- **Type:** Anchor person
- **Contributor:** Giubilei Giuliano

- **Type:** News Reader
- **Contributor:**

- **Type:** Editor-in-chief
- **Contributor:** Di Bella Antonio

Work Info

Figure 16: information about an EDOB

In this section we have only presented an introduction to the ADMIN component, with the aim of giving an overview of its functionalities. We do not present details on the implementation of this component, since it is a standard web application. However, in section 6.1 we will give a detailed discussion on how to use the ADMIN component by means of the web application.

4. Implementation of the Turnkey System

In this section we give further technical details about the implementation of the Turnkey system. We present detailed information about the implementation of its components, namely the Documentation Platform and the Publication Platform.

4.1. Implementation of the Documentation Platform

4.1.1. Physical Environment

In order to manage the activities of different GAMPs, the Documentation Platform makes use of a **Work Flow system**.
The Work Flow system can be either a commercial or an open source system. Our current implementation adopts the OpenFlow engine, running on the Zope platform. However, in order to allow possible future changes to the adopted work flow system, the Java components interacting with the work flow management system has been developed in an abstract way. More precisely, suitable interfaces are provided. Moreover, classes implementing those interfaces and referring to OpenFlow are also provided.
In order to adopt another work flow system (thus replacing OpenFlow), one does not need to change the overall architecture: it is only needed to implement all the work flow interfaces, providing suitable classes. These classes will replace the ones provided for the OpenFlow engine.

Let us conclude this section with a brief remark on the physical environment of GAMPs. The GAMPs (the clients) involved in the process can be thought either as running on the same machine or as performing on different ones. It is worth noticing that the whole MAD System could be made up of a Rack system where every single machine will handle some specific task. Furthermore, as discussed in section 3.7 above, one can think of having a limited number of GAMPs running on a “lightweight” Documentation Platform, requiring results from other GAMPs running on other machines. A GRID-like architecture can be provided to ensure the connections among these components.

4.1.2. Interaction between GAMPs and Core Platform

In section 3.2 we have recalled the main features of GAMPs. GAMP stands for Generic Activity Metadata Processor. A GAMP implements a specific process of metadata extraction within the Documentation Platform.

In this section, we describe how a single GAMP interacts with the Core Platform (of the Documentation Platform), in order to offer its functionalities.

As a first step (1), a GAMP polls its own queue in the Core Platform. If it founds some work to do, i.e. an active job belongs to the GAMP's queue, then the GAMP gets its job and starts its process. As a second step (2), the GAMP asks the Core Platform to checkout the EDOBs related to the job in analysis. Before starting its own metadata extraction, the GAMP also needs to retrieve all the files linked to the EDOBs; to this aim, another invocation to the Core Platform is performed (3). In order to physically retrieve the requested file, the Core Platform asks the EMS.

At this time of the process, the GAMP has all the information needed to perform its own metadata extraction (4). This process could require to store additional files and/or metadata in the factory; in this case, the GAMP asks to insert new material (5) by means of a request to the Core Platform. The Core Platform forwards the GAMP's request to the EMS component. The metadata (EDOB) built by the GAMP are then registered on the Core Platform (6). Finally, the GAMP notifies the Core Platform that the elaboration of the current job is over (7).

The communication between a GAMP and the Core Platform is based on web services, and it is performed by exchanging XML documents. As an example, suppose that a GAMP needs a specific file in order to perform its process: in phase 3, it asks the Core Platform to retrieve this file (by forwarding this request to the EMS component). The result of this request consists of an XML document containing all the information necessary to recover the file, that is to say the protocol to use, the port to adopt, and so on.
As another example, consider phase 5, and suppose that a GAMP needs to store some files in the factory. In this case, the GAMP returns an XML document to the Core Platform: this document contains all the parameters needed to access the produced data. The Core Platform will then send this XML document to the "restricted" PSO managing the activities of the Tunrkey System, which will download these new materials according to the directives of the EMS.

The interaction between a GAMP and the Core Platform can be summarized as follows:

1. The GAMP asks a **getJob(queueName):XML** to the Core Platform; queueName is the name of the queue associated with that GAMP. This service returns an XML document with all the information about the job to be processed by the GAMP;
2. The GAMP asks a **checkoutEDOB(idEDOB):XML** to the Core Platform; by means of this operation, the GAMP asks the Core Platform to return the EDOB (an XML document), whose identifier (idEDOB) is passed as an argument;

3. The GAMP invokes a **getMaterial(UMIDs):XML**, aiming at recovering some files needed to its metadata extraction. As usual, the Core Platform answers with an XML document. The identifiers of all needed files are stored in the list UMIDs, which is passed as an argument;
4. elaboration: the GAMP performs its own metadata extraction
5. if the GAMP, during the elaboration of phase 4, has generated some new files, then it needs to store them in the factory. In this case, the GAMP asks an **insertMaterial(XMLDocument)** to the Core Platform. XMLDocument contains all information on the new files. The Core Platform forwards this file to the EMS in order to make it available to the factory;
6. the GAMP asks the Core Platform the **checkinJob(XMLDocument)** in order to register the EDOB (the argument XMLDocument) produced by the GAMP;
7. The GAMP notifies the Core Platform that it has successfully concluded its work. This is made by an invocation of **notifyJob(XMLDocument)**.

4.1.3. Man-Machine, Control, Software, Files and database interface

4.1.3.1. Man-Machine interfaces

The interaction between human beings and the Documentation Platform is performed by means of the ADMIN component.
In section 3.6 we have mentioned that the Documentation Platform comprises an ADMIN interface, which is a standard web application. The ADMIN component offers a GUI for adding annotation and representative information to the Metadata. The web interface offers these main functionalities:

1. submitting essences and providing the metadata.
2. managing the work flow and for controlling the entire work cycle.
3. metadata browsing: a web interface for browsing the essences and the metadata is provided. It will be useful for getting a quick view of the work done.

4.1.3.2. Control Interfaces

The Core Platform provides a control interface based on SOAP protocol. We are currently analyzing SNMP protocols for controlling the machines involved.

4.1.3.3. Software interfaces

The Core Platform provides Web Services interfaces in a WSDL format.

The main interfaces are:

1. PSO
2. EMS
3. Work Flow
4. Admin
5. Browse

Some features could be provided as Java Web Start applications, published on http (jnlp mime type).

4.1.3.4. Files and Databases interfaces

Files can be accessed by the following protocols provided by the Core Platform (by means of an interaction with the SO component called EMS):

1. file://
2. smb:// (samba)
3. ftp://
4. http-s://
5. soap://

4.2. Implementation of the Publication Platform

In previous sections we have introduced the Publication Platform, which is a web application allowing users to retrieve and use audiovisual information within the Factory. In this section we describe in detail the software components used by the Publication Platform in order to perform its functionalities.

4.2.1. 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.

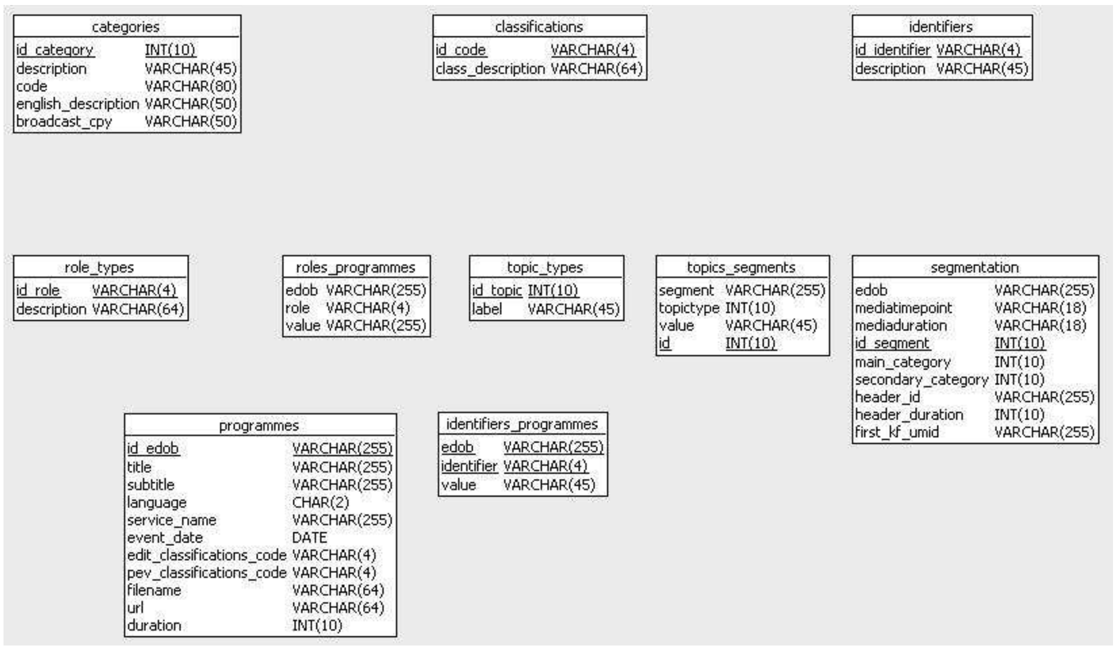
4.2.2. The Kim Platform

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.

4.2.3. 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 17: 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.

PART B: Utilization of the Turnkey System

In this part of the Deliverable some details on how to use the Turnkey System and its components are provided.

5. How to use the Turnkey System

In this section we give details on how to use the components of the Turnkey System. This part can be seen as a sort of user manual for this component.

5.1. How to use the ADMIN component

In this section, we describe how to use the web application called ADMIN, which is used by the administrator in order to monitor the work flow management system.

After being logged in the factory, the above operations can be performed by the main web page of the Documentation Platform section:

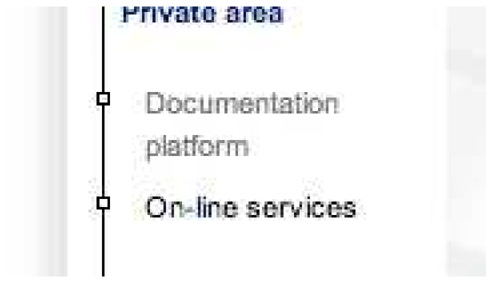
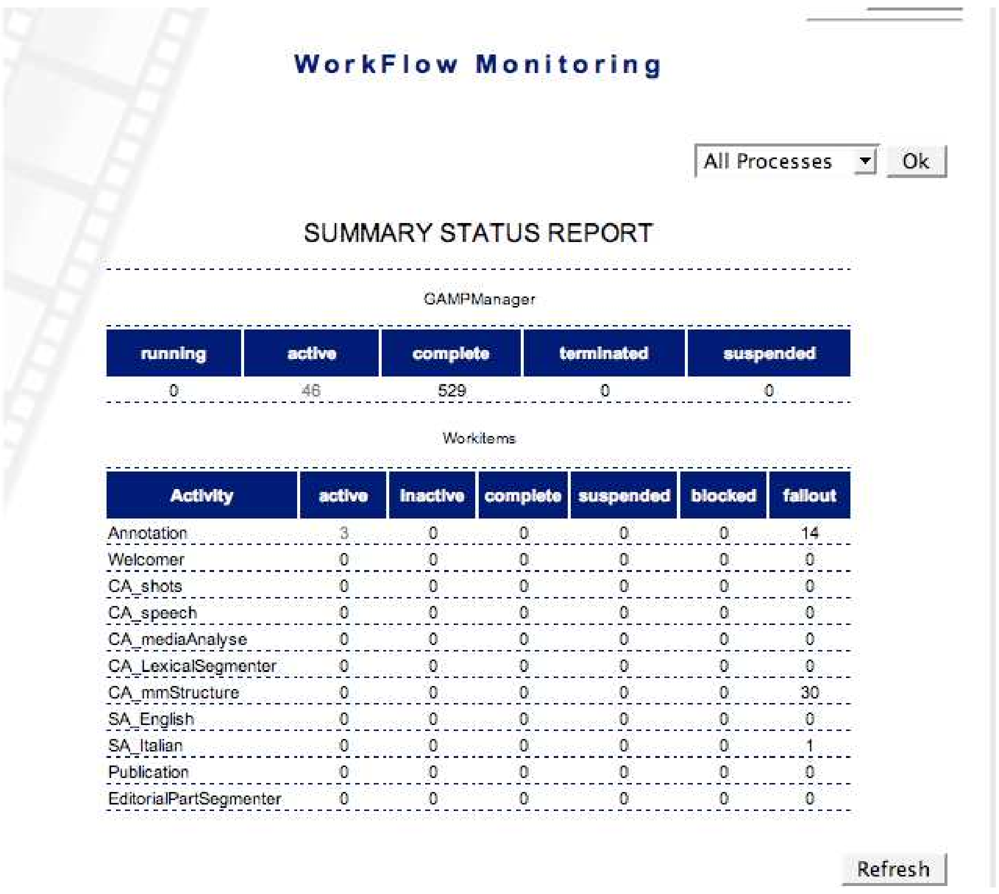


Figure 18: the link to the ADMIN component in the Documentation Platform

The main features of the ADMIN component can be selected by the third link of the main page, i.e. the link called “Work Flow Monitoring”. The page below corresponds to this section:



SUMMARY STATUS REPORT

GAMPManager

running	active	complete	terminated	suspended
0	46	529	0	0

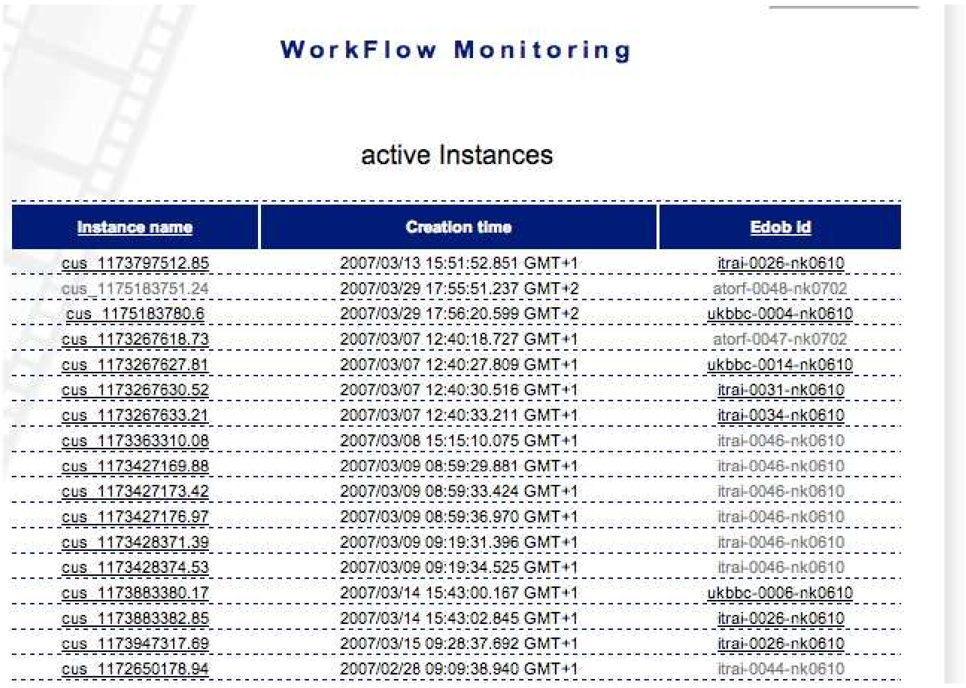
Workitems

Activity	active	inactive	complete	suspended	blocked	fallout
Annotation	3	0	0	0	0	14
Welcomer	0	0	0	0	0	0
CA_shots	0	0	0	0	0	0
CA_speech	0	0	0	0	0	0
CA_mediaAnalyse	0	0	0	0	0	0
CA_LexicalSegmenter	0	0	0	0	0	0
CA_mmStructure	0	0	0	0	0	30
SA_English	0	0	0	0	0	0
SA_Italian	0	0	0	0	0	1
Publication	0	0	0	0	0	0
EditorialPartSegmenter	0	0	0	0	0	0

Figure 19: the Work Flow Monitoring page

Two tables are presented:

- the GAMP manager table, summarizing the status of all the items involved in the system, namely the items running, active, completed, terminated and suspended. By clicking on each number, the list of items of the selected category is presented. As an example, Figure 20 shows the list of active items.
- the Workitems table, whose rows contain the different activities that can be performed on a work item (Annotation, Welcomer, CA_shots, and so on), whereas the columns contain the status of the items, namely active, inactive, completed, suspended, blocked, fallout. By selecting each item of the table, the ADMIN component also offers the opportunity to check all the information about the specific work item, monitoring its evolution within the work flow process. An example is presented in Figure 21.

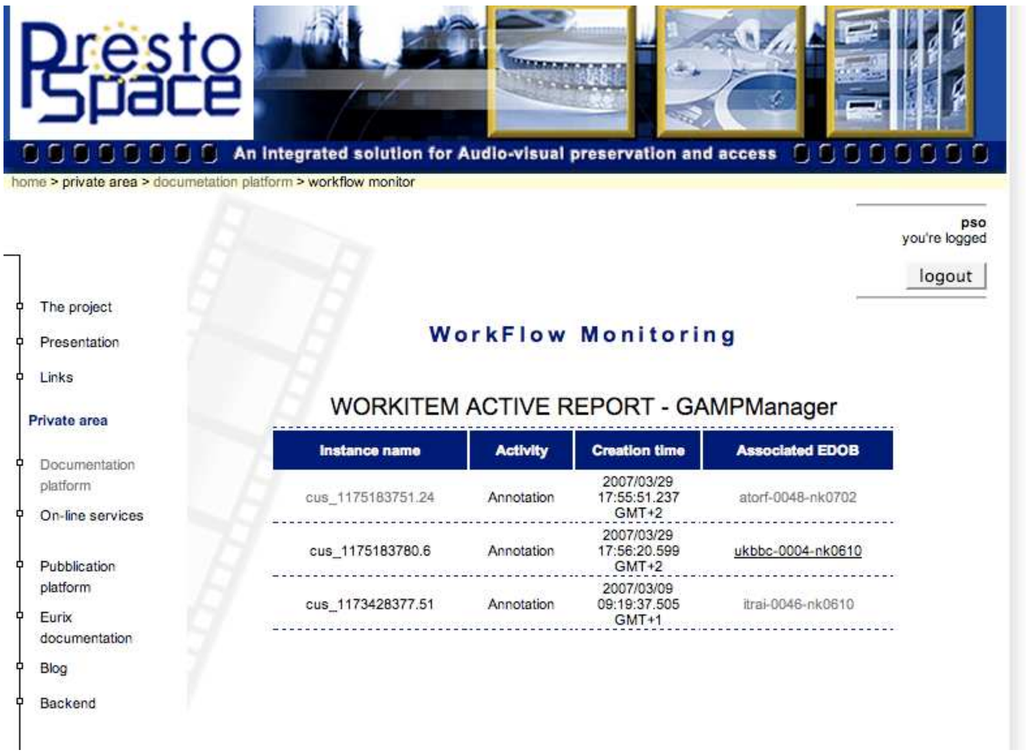


WorkFlow Monitoring

active Instances

Instance name	Creation time	Edob id
cus_1173797512.85	2007/03/13 15:51:52.851 GMT+1	itrai-0026-nk0610
cus_1175183751.24	2007/03/29 17:55:51.237 GMT+2	atorf-0048-nk0702
cus_1175183780.6	2007/03/29 17:56:20.599 GMT+2	ukbbc-0004-nk0610
cus_1173267618.73	2007/03/07 12:40:18.727 GMT+1	atorf-0047-nk0702
cus_1173267627.81	2007/03/07 12:40:27.809 GMT+1	ukbbc-0014-nk0610
cus_1173267630.52	2007/03/07 12:40:30.516 GMT+1	itrai-0031-nk0610
cus_1173267633.21	2007/03/07 12:40:33.211 GMT+1	itrai-0034-nk0610
cus_1173363310.08	2007/03/08 15:15:10.075 GMT+1	itrai-0046-nk0610
cus_1173427169.88	2007/03/09 08:59:29.881 GMT+1	itrai-0046-nk0610
cus_1173427173.42	2007/03/09 08:59:33.424 GMT+1	itrai-0046-nk0610
cus_1173427176.97	2007/03/09 08:59:36.970 GMT+1	itrai-0046-nk0610
cus_1173428371.39	2007/03/09 09:19:31.396 GMT+1	itrai-0046-nk0610
cus_1173428374.53	2007/03/09 09:19:34.525 GMT+1	itrai-0046-nk0610
cus_1173883380.17	2007/03/14 15:43:00.167 GMT+1	ukbbc-0006-nk0610
cus_1173883382.85	2007/03/14 15:43:02.845 GMT+1	itrai-0026-nk0610
cus_1173947317.69	2007/03/15 09:28:37.692 GMT+1	itrai-0026-nk0610
cus_1172650178.94	2007/02/28 09:09:38.940 GMT+1	itrai-0044-nk0610

Figure 20: the list of active instances



WorkFlow Monitoring

WORKITEM ACTIVE REPORT - GAMPManager

Instance name	Activity	Creation time	Associated EDOB
cus_1175183751.24	Annotation	2007/03/29 17:55:51.237 GMT+2	atorf-0048-nk0702
cus_1175183780.6	Annotation	2007/03/29 17:56:20.599 GMT+2	ukbbc-0004-nk0610
cus_1173428377.51	Annotation	2007/03/09 09:19:37.505 GMT+1	itrai-0046-nk0610

Figure 21: information about a work item

5.2. How to use the generic GAMP

In this section we describe in detail how to make use of the Generic GAMP in order to add a new GAMP component to the Documentation Platform.

Here below are the instructions about the usage of the Generic GAMP via command line:

```
JAVA_HOME:/usr/lib/java
JAVA:/usr/lib/java/bin/java
PRESTOSPACE_HOME:/shared/prestospace/GenericGAMP/linux
LIB_DIR:/shared/prestospace/GenericGAMP/linux/lib
```

USAGE:

ClGamp.sh {XMLConfigFile | -h | -h CHECKOUT_EDOB | -h CHECKIN_EDOB | -h NOTIFY | -h GET_JOB | -h GET_MATERIAL | -h INSERT_MATERIAL}

where:

XMLConfigFile is an XML file defining a GAMP Operation.

To see all xml templates run

ClGamp.sh -h

-h option to read all xml templates
-h CHECKOUT_EDOB option to read xml template for CHECKOUT_EDOB
-h CHECKIN_EDOB option to read xml template for CHECKIN_EDOB
-h NOTIFY option to read xml template for NOTIFY
-h GET_JOB option to read xml template for GET_JOB
-h GET_MATERIAL option to read xml template for GET_MATERIAL
-h INSERT_MATERIAL option to read xml template for INSERT_MATERIAL

QUEUES for GET_JOB command:

Welcomer	(Demux and so on)
CA_shots	(ContentAnalysis jobap)
CA_speech	(ContentAnalysis jobap)
CA_mediaAnalyse	(ContentAnalysis jobap)
CA_other	(ContentAnalysis jobap)
Restoration	
SA_generic	(SemanticAnalysis jobap)
SA_other	(SemanticAnalysis jobap)
Annotation	
Delivery	

5.3. How to use the Publication Platform’s Interface

In this section we describe how to use the interface of the Turnkey System, namely the interface provided by the Publication Platform.

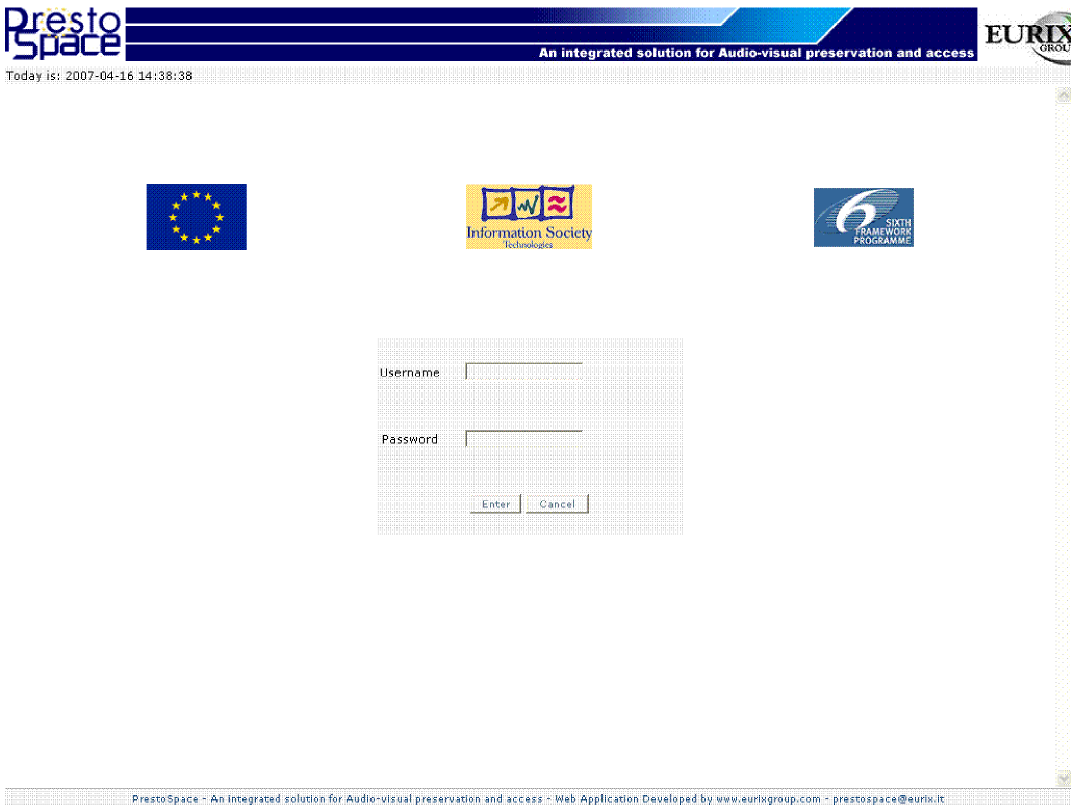
This part can be seen as a sort of user manual of the Turnkey System. It is quite obvious that this interface is a restricted version of the one of the Publication Platform for the MAD factory described in **[D18.2]**; indeed, the interface allows to access only the specific information offered by the GAMPs supported by the Turnkey System.

5.3.1. 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:



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

Username

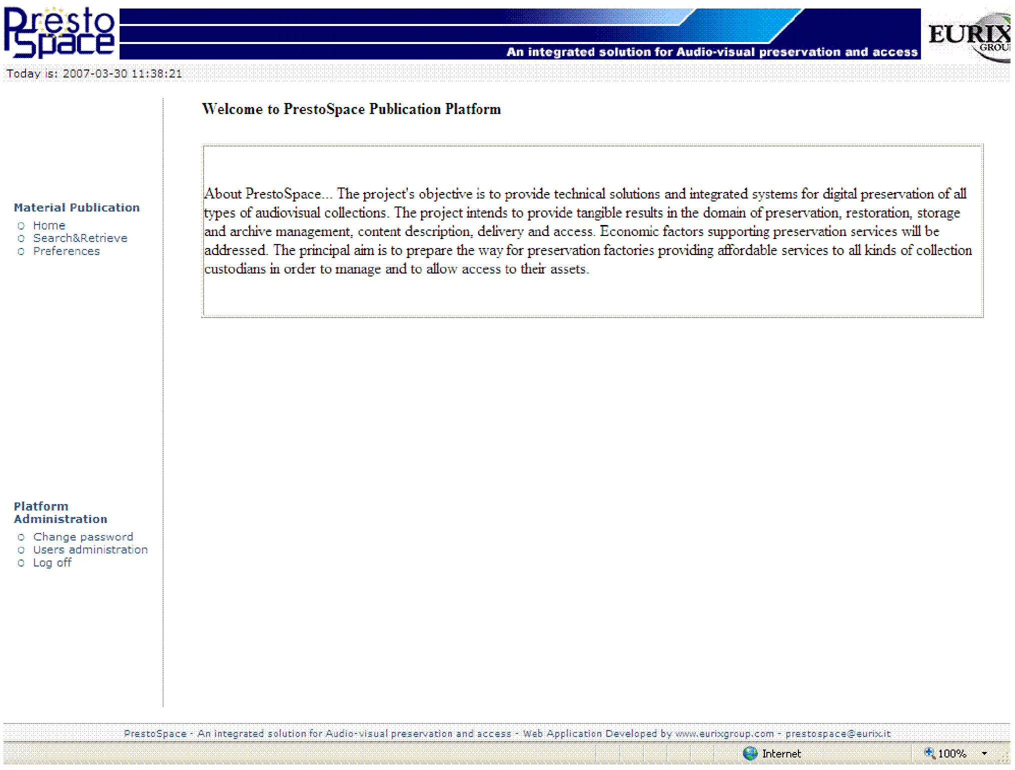
Password

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

An integrated solution for Audio-visual preservation and access
Enter
Cancel

Figure 22: The Welcome Page of the Publication Platform

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



An integrated solution for Audio-visual preservation and access

Today is: 2007-03-30 11:38:21

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.

Material Publication
Home
Search&Retrieve
Preferences
Platform Administration
Change password
Users administration
Log off
PrestoSpace - An integrated solution for Audio-visual preservation and access - Web Application Developed by www.eurixgroup.com - prestospace@eurix.it
Internet
100%

Figure 23: The Publication Platform Web Interface

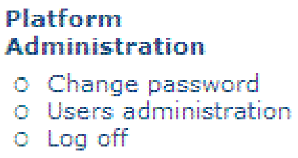
The user can access the documented programme/news or manage the user accounts.

5.3.2. Platform Administration

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

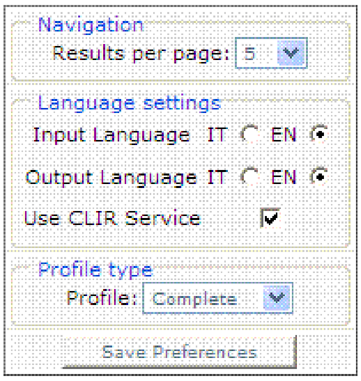
- change its own password
- access to the administration of users
- log off the system.

The web page is shown here below.



5.3.3. Material Publication

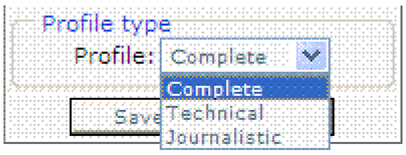
Preferences



Navigation
Results per page: 5
Language settings
Input Language IT [✓ EN]
Output Language IT [✓ EN]
Use CLIR Service [✓]
Profile type
Profile: Complete

Figure 24: Preferences

The user can set the number of results displayed in a page (default: 5) 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.

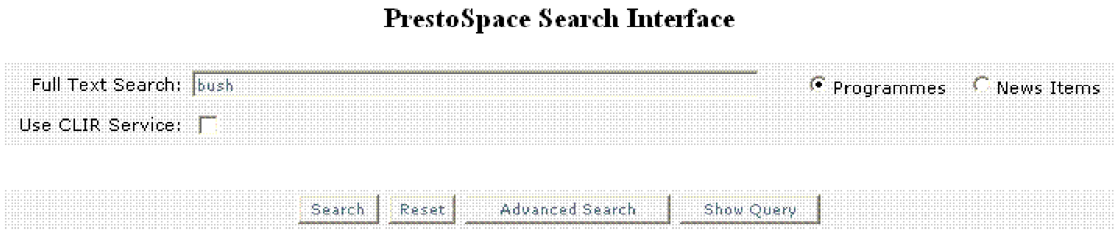


Profile type
Profile: Complete

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)**.**

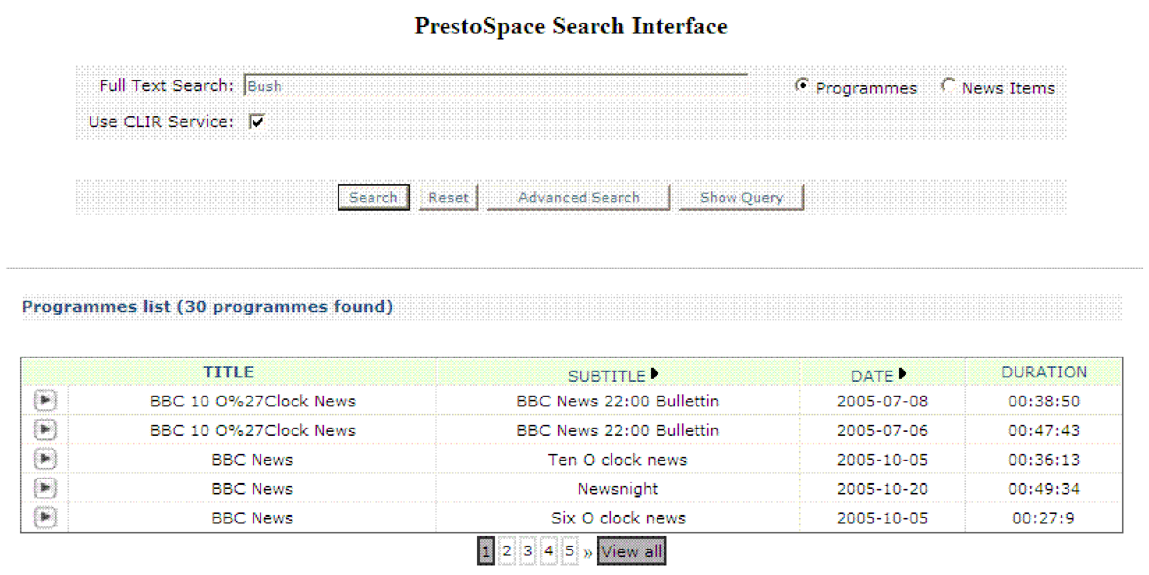


PrestoSpace Search Interface

Full Text Search: bush [✓ Programmes] News Items
Use CLIR Service:

Figure 25: 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 [✓ Programmes] News Items
Use CLIR Service: ✓

Search | Reset | Advanced Search | Show Query

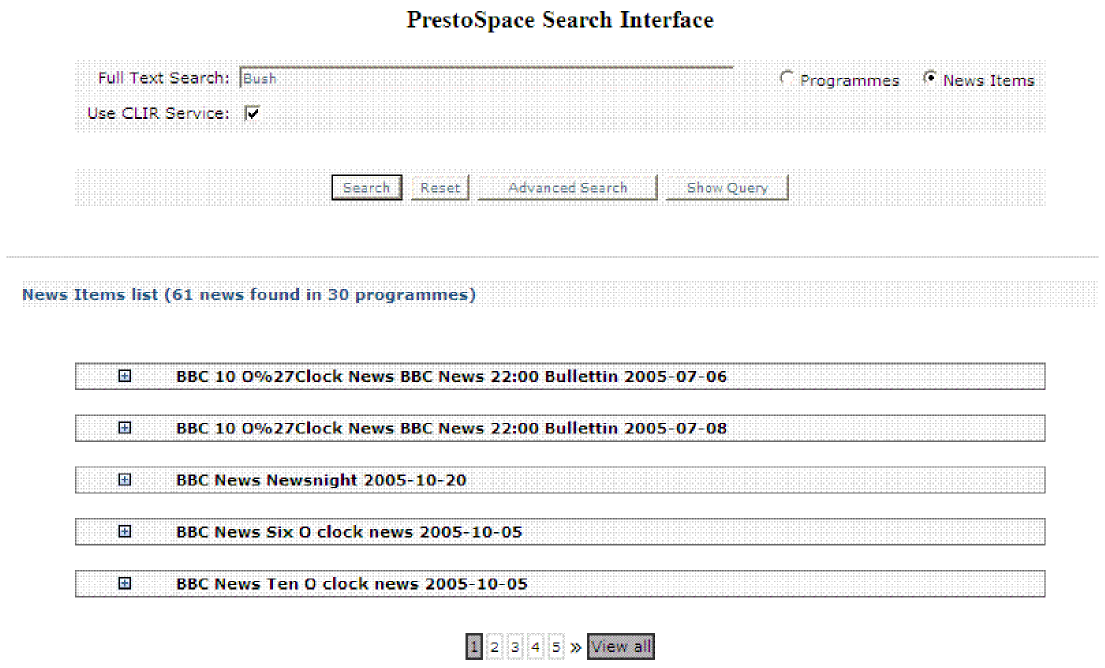
Programmes list (30 programmes found)

	TITLE	SUBTITLE	DATE	DURATION
▶	BBC 10 O%27Clock News	BBC News 22:00 Bulletin	2005-07-08	00:38:50
▶	BBC 10 O%27Clock News	BBC News 22:00 Bulletin	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

1 2 3 4 5 » View all

Figure 26: 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 Programmes [✓ News Items]
Use CLIR Service: ✓

Search | Reset | Advanced Search | Show Query

News Items list (61 news found in 30 programmes)

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

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

⊞ BBC News Newsnight 2005-10-20

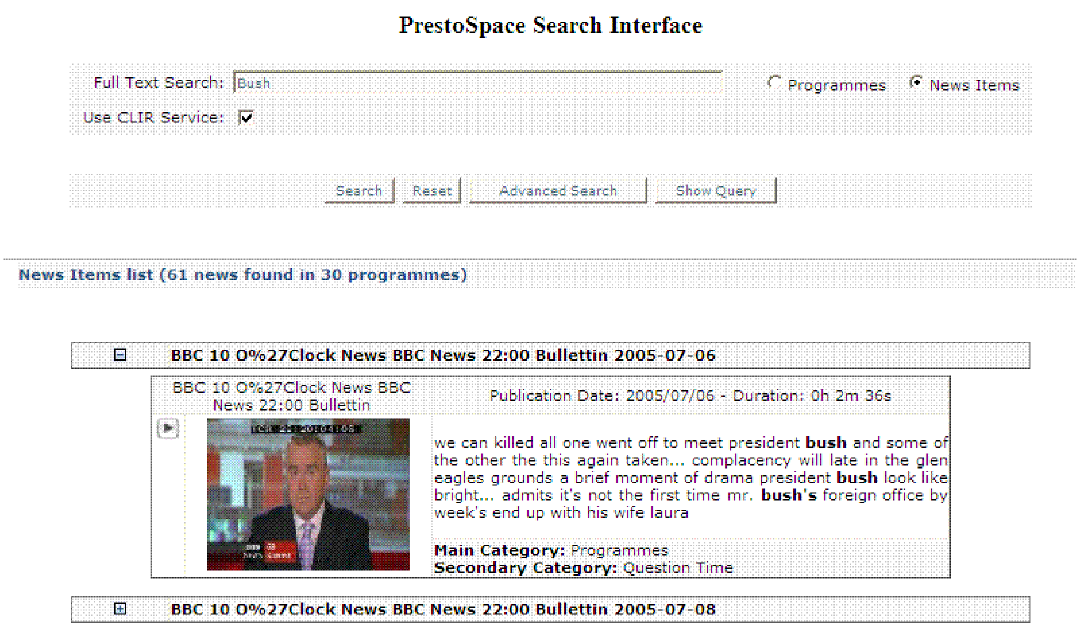
⊞ BBC News Six O clock news 2005-10-05

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

1 2 3 4 5 » View all

Figure 27: 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 Bulletin 2005-07-06

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

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

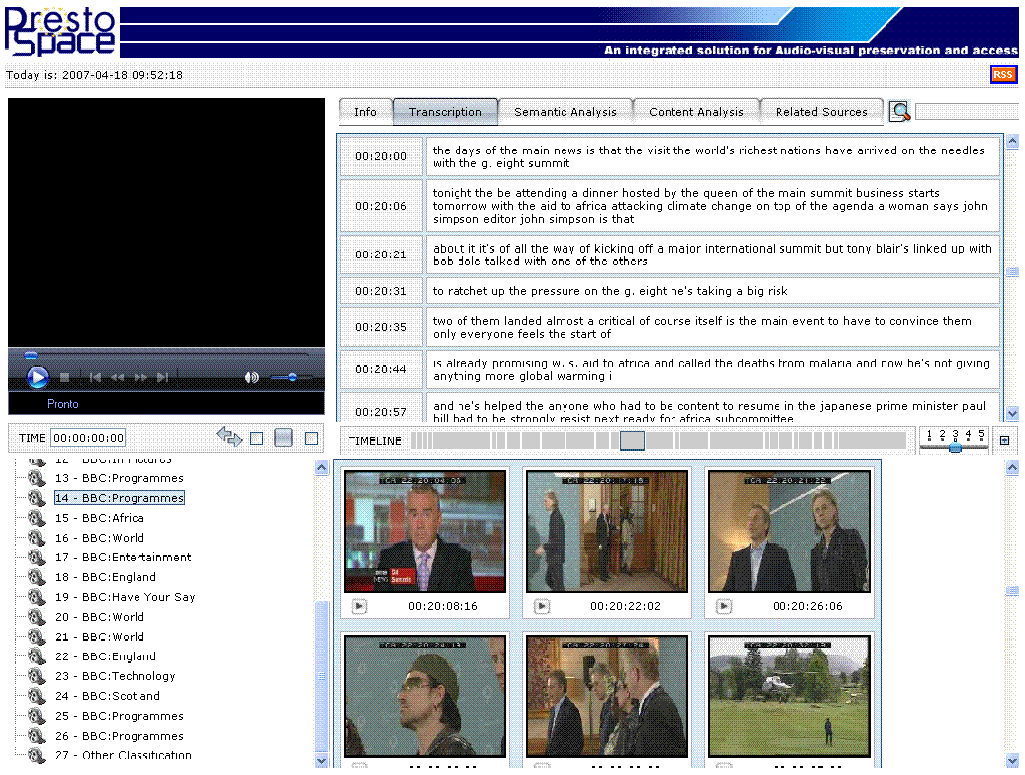
we can killed all one went off to meet president **bush** and some of the other the this again taken... complacency will late in the glen eagles grounds a brief moment of drama president **bush** look like bright... admits it's not the first time mr. **bush's** foreign office by week's end up with his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

Main Category: Programmes
Secondary Category: Question Time

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

Figure 28: 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 (Fig. 29):



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

TIMELINE

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 29: 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 (**Erreur ! Source du renvoi introuvable.**) 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 4.2.2–)
- 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).

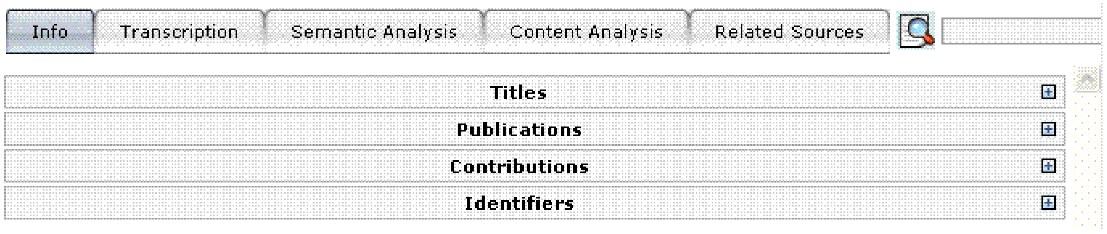
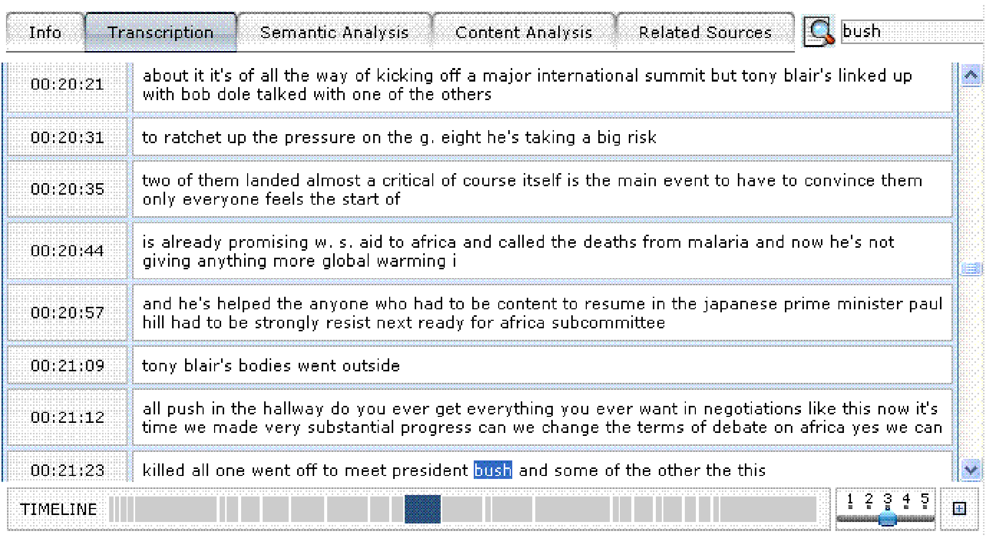


Figure 30: legacy data – The Info tab

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 31: 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

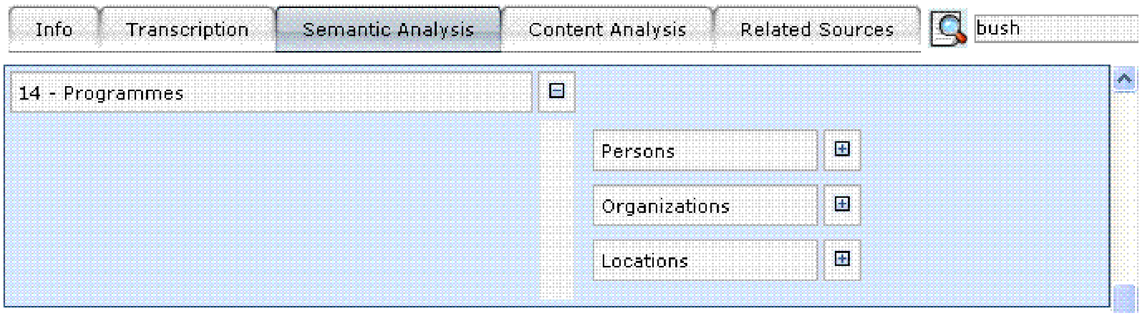
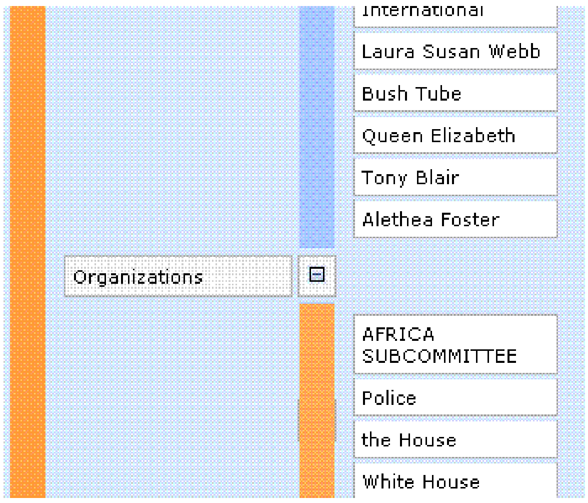


Figure 32: Semantic Analysis Tab

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 33), 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 34).

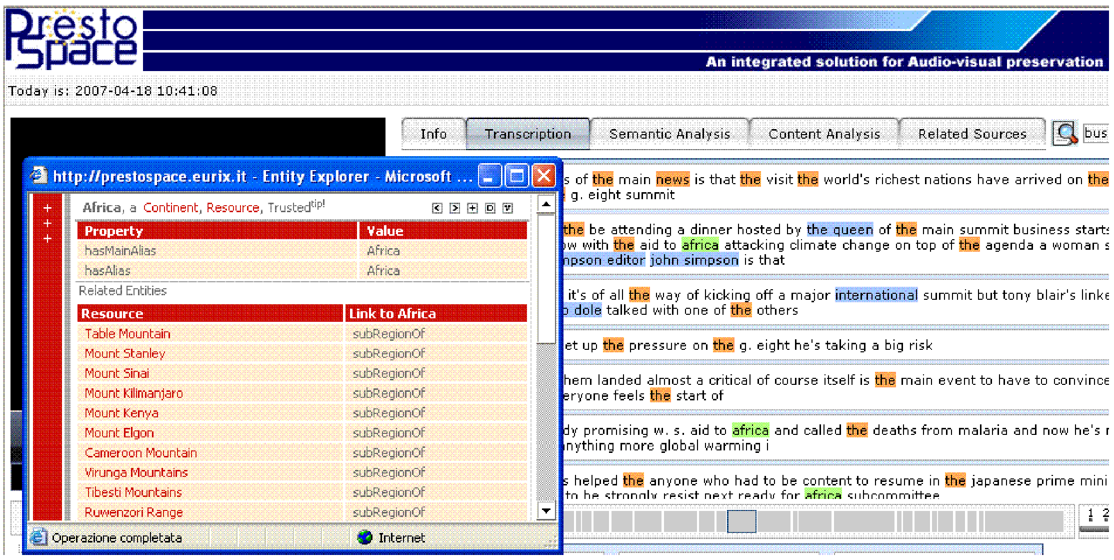


International
Laura Susan Webb
Bush Tube
Queen Elizabeth
Tony Blair
Alethea Foster

Organizations

AFRICA SUBCOMMITTEE
Police
the House
White House

Figure 33: named entities



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

http://prestospace.eurix.it - Entity Explorer - Microsoft ...

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

Operazione completata Internet

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
w 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
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
nything more global warming i

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

Figure 34: The Entity Explorer

Content Analysis Tab

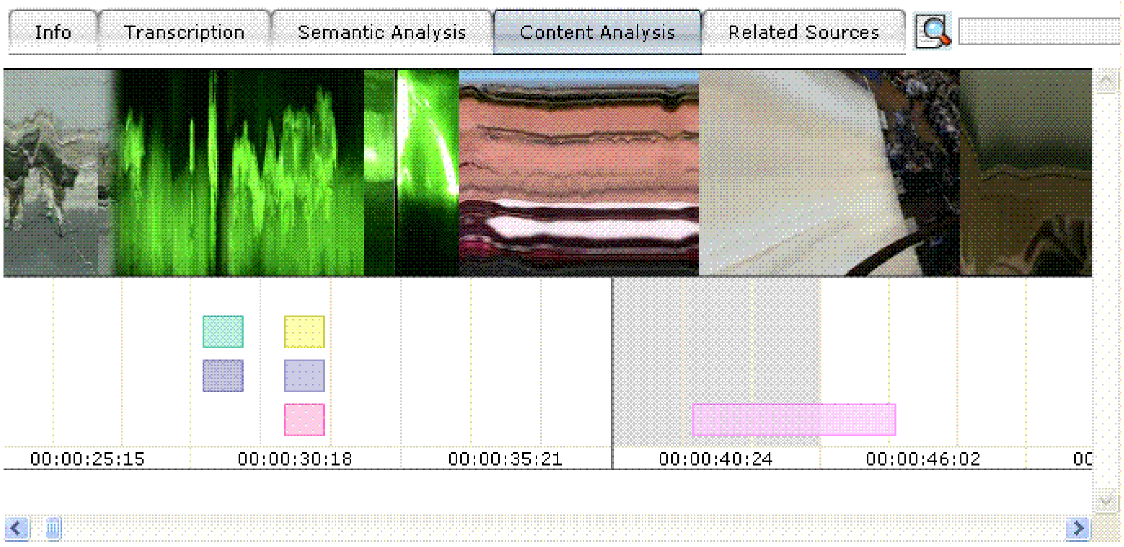
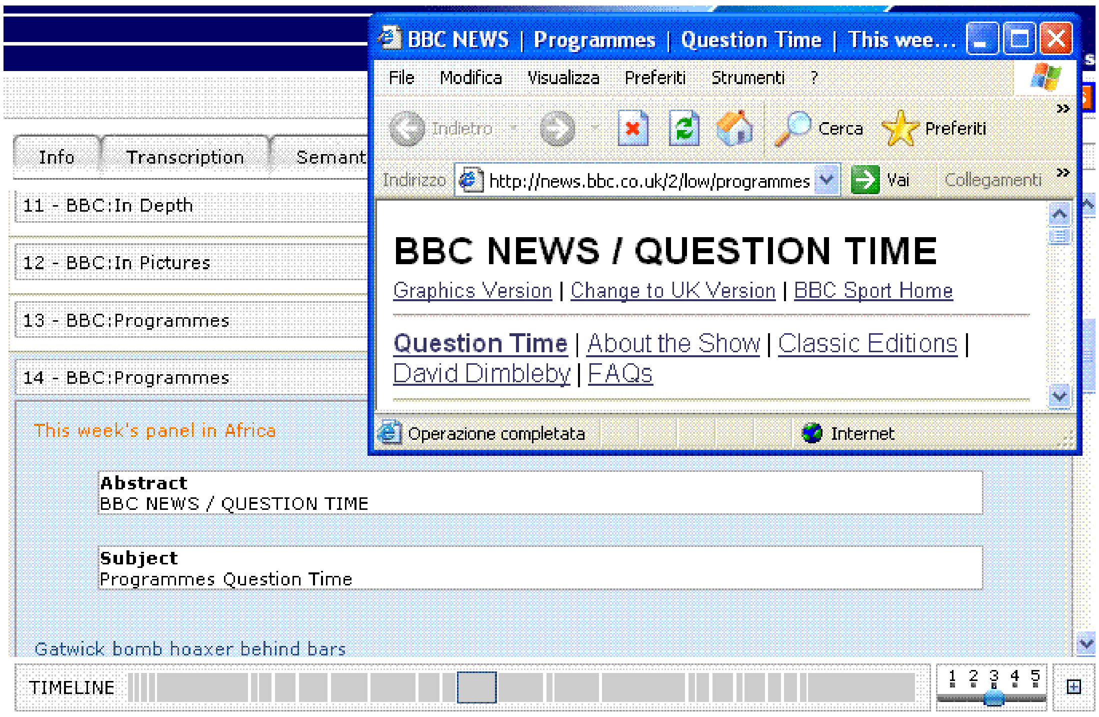


Figure 35: Content Analysis tab

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


BBC NEWS | Programmes | Question Time | This wee...
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti
Indirizzo http://news.bbc.co.uk/2/low/programmes Vai Collegamenti
Info Transcription Semant
TIMELINE
1 2 3 4 5
Operazione completata
Internet

11 - BBC:In Depth
12 - BBC:In Pictures
13 - BBC:Programmes
14 - BBC:Programmes

This week's panel in Africa

Abstract
BBC NEWS / QUESTION TIME

Subject
Programmes Question Time

Gatwick bomb hoaxer behind bars

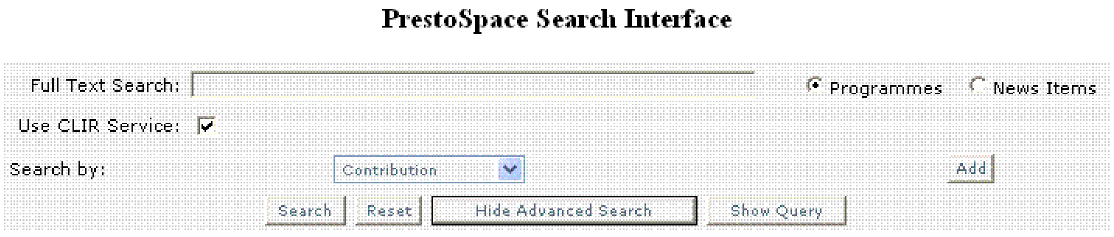
BBC NEWS / QUESTION TIME
Graphics Version | Change to UK Version | BBC Sport Home
Question Time | About the Show | Classic Editions | David Dimbleby | FAQs

Figure 36: 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

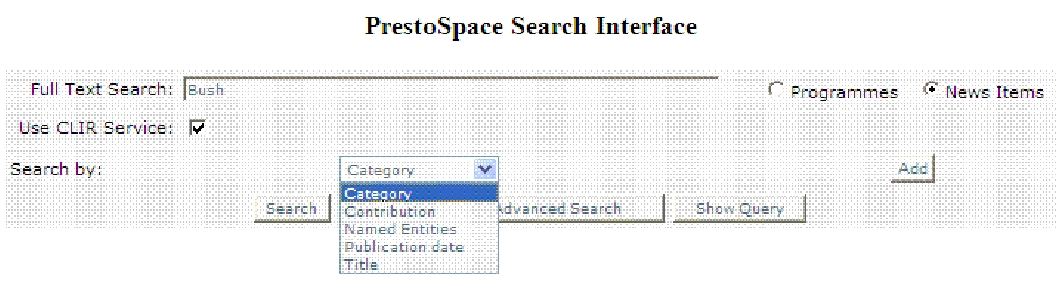
Clicking on the Advanced Search button the user can submit queries more complicated than a simple full text searching task.



Full Text Search: [✓ Programmes] News Items
Use CLIR Service: ✓
Search by: Contribution

Figure 37: Advanced Search

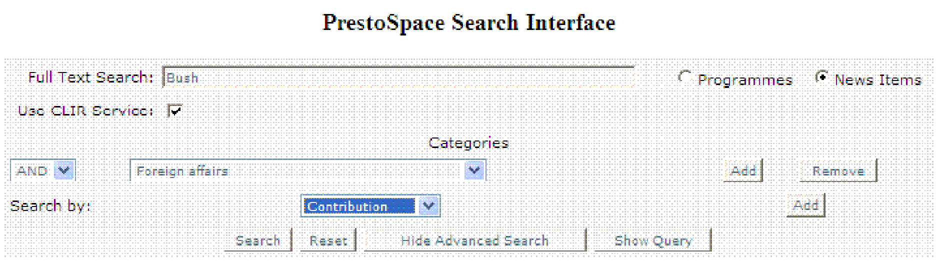
The user can use filters searching by Category (only for News items), Contributions, Named entities, Publication Date, Publication Service (only for Programmes) and title.


PrestoSpace Search Interface

Full Text Search: Bush
Programmes [✓ News Items]
Use CLIR Service: ✓
Search by: Category

Figure 38: Filters of the queries in the Advanced Search

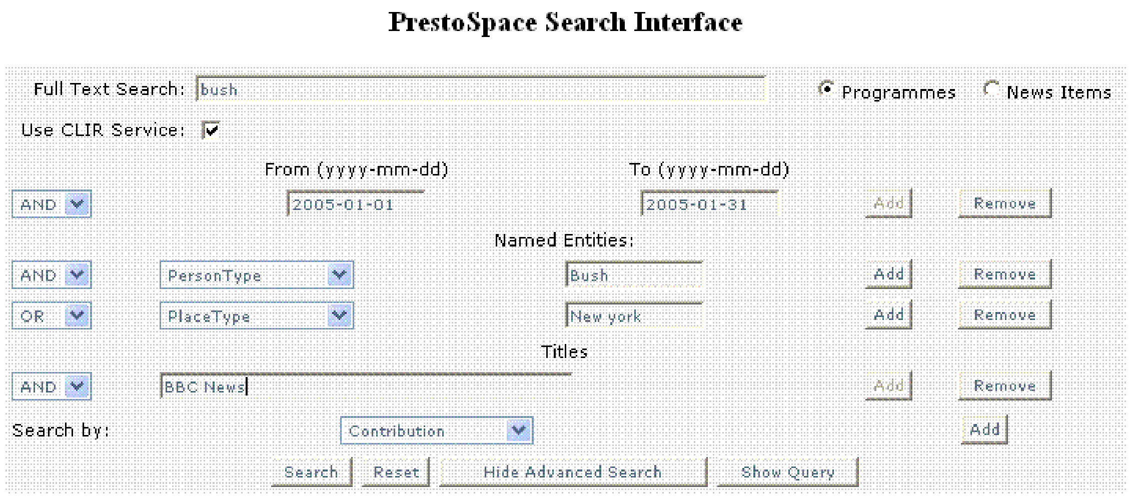
Clicking on the 'Add' button the filter is added to the query. It is also possible to make logical operation on the filters of the same type (AND/OR):


PrestoSpace Search Interface

Full Text Search: Bush
Programmes [✓ News Items]
Use CLIR Service: ✓
Categories: AND Foreign affairs
Search by: Contribution

Figure 39: Filter of the query

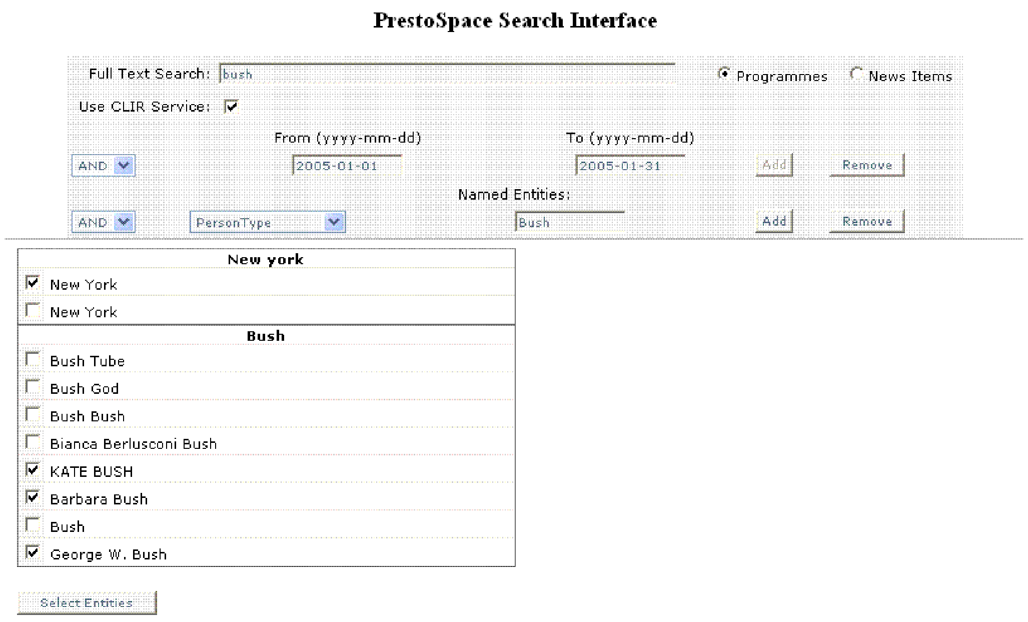
In the following Figure it is selected to search any BBC News that contains the word 'bush', published in the January of 2005, containing the Person 'Bush' OR the Place New York:


PrestoSpace Search Interface

Full Text Search: bush
[✓ Programmes] News Items
Use CLIR Service: ✓
AND From (yyyy-mm-dd) 2005-01-01 To (yyyy-mm-dd) 2005-01-31
Named Entities:
AND PersonType Bush
OR PlaceType New york
Titles
AND BBC News
Search by: Contribution

Figure 40: 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: ✓

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

Named Entities: AND PersonType 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 41: Named Entities of the query

PART C: Conclusions

6. Licensing

The software product described in this deliverable is a prototype in an advenced state of implementation. In order to engineer this software, some further implementation steps have to be performed. As an example, an improvement of the feedback machinery is needed.

7. Bibliography

[D15.4]	Content Analysis Tools.
[D15.5]	Cross-Linguistic IE Tools Analysis.
[D15.6]	Semantic Interpretation Tools.
[D16.2]	Conceptual Search.
[D16.4]	Delivery Models.
[D18.1]	The Documentation Platform for the MAD Factory.
[D18.2]	Publication Platform for the Results of Digitization and Documentation
[D19.0.1]	External and Internal Models and Protocols for the PrestoSpace Factory.
[D19.0.2]	The PrestoSpace Orchestrator (PSO)

8. Glossary

Term	Description
ADMIN	The component of the `Documentation Platform` which allows the user to manage and monitor the activities of the `Documentation Paltform.` It represents the interface between human users and the Documentation Platform.
Core Platform	The component of the `Documentation Platform` offering a workflow management service and interacting with `PSO`'s components `EMS` and `CVS`. This software component represents the middleware which is publishing `web services` interfaces. It has a built in work flow engine for managing all the activities done within the MAD platform (content analysis, semantic analysis, annotation, delivery, etc...).
CVS	Concurrent Versioning System, the system which is responsible for tracking every change to the metadata that takes place during the execution of a GAMP. It is a component of the `PSO.`
Enhanced Metadata	Medatata and structuring information that are generated within the `MAD factory`, in the view of improving the accessibility to digitised contents.
GAMP	This label stands for Generic Activity MAD Processor, which represents the Generic Client communicating to the `MAD Core Platform`, using SOAP (WebServices) protocol.
Generic GAMP	Software component which allows to simplify the creation of a new GAMP.
MAD Factory	Facilities where massive documentation, metadata enhancement, and preparation of publication for audiovisual contents are performed.
Mass Storage	Storage solution in which all the assets (programs, recordings etc.) are kept on common media (disks, tapes etc.) and access is managed through a file management system.
Publication Platform	The component of the MAD Factory which is responsible of delivering enriched audiovisual contents.
Preservation Factory	Facilities where massive A-to-D migration of audiovisual contents is performed.
PSO	This is the PrestoSpace Orchestrator, which is the `administrator` of the PrestoSpace factory, coordinating all its components `PRE, RES,` and `MAD`.
Queue Filler	Software component used to test the `Documentation Platform`. It gives the opportunity of inserting jobs in the GAMP's queues.
Turnkey system	Complete name: `Turnkey system for delivering to small archives`. A small scale stand alone production

	quality system suitable for small archives and already configured for the publication of the preserved material. Given the intrinsic modularity of the MAD Platform, the functionalities deployed in a `Turnkey System` installation can be modulated according to the user needs. The `Turnkey System` will be derived from the test bed that the project is setting up to run all the experiments needed to define the specifications of the final platform.
EMS	Essence and Metadata Storage System, the system which is responsible for storing the essence within the `PSO`.
Documentation Platform	The component of the MAD Factory which is responsible of extracting metadata from audiovisual content by means of different GAMPs.