ECP 2008 DILI 518002 EUscreen

Exploring Europe's Television Heritage in Changing Contexts

D4.1 - Functional Specifications & Portal Architecture

Deliverable number	*D4.1 - Functional Specifications and Portal Architecture*
Dissemination level	*Public*
Delivery date	*2010*
Status	*Final*
Author(s)	*NTUA & Noterik*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1]
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

1 OJ L 79, 24.3.2005, p. 1.

Document Information

Deliverable number: *D4.1*
Deliverable title: *Functional Specifications and Portal Architecture*
Actual date of deliverable:
Workpackage: 4
Workpackage title: Semantic Access & Integration
Workpackage leader: NTUA
Keywords: Functional specifications

Table of Contents

1 Scope of this document

The aim of this document is to present the functional specifications of the EUScreen system. The functional specifications have been produced based on the 3 regional workshops that have been organized on the framework of WP3 and WP4 and on the collaboration of the 2 technical partners, NTUA and NOTERIK. In a number of face-to-face and virtual meeting between the technical partners and taking into account D5.1 and D5.1.1 especially for the portal design, the functional specifications have been finalized and are presented in this deliverable. The EUScreen system comprises of the back end and front end tools. The back end tool is responsible for ingesting and preparing the metadata to be presented in the EUScreen portal and also make them available to Europeana. The front end tool icorporates the construction of the EUScreen portal functionalities including transcoding and play-out of the digital content. NTUA is responsible for the construction of the back end tool and NOTERIK for the front end tool. Based on the functional requirements, the tools will be proposed tools will be developed. The first prototypes of the tools will be presented in the workshop in Greece on June. The pilot-testing phase will start on July and will end on October 2010.

2 Overall Architecture

Figure 2.1 illustrates the architecture of the proposed system based on the functional specifications for the EUscreen technical infrastructure. The overall process will involve a Content and Metadata Ingestion web service that provides tools to establish and exploit the semantic alignment of metadata schemata that content providers use to annotate their items, to a common machine understandable schema. Preserving a well defined interoperability allows for homogenized presentation of cultural audiovisual objects on the EUscreen portal and creates a communication channel for the mediation of aggregated content to external services such as the Europeana portal. The adopted approach combined with the use of established protocols and standards facilitate the maintenance of the repository's interoperability with evolving data sources, with minimal provider intervention. It reinforces the repository's evolution procedures by providing both versioning and editing capabilities for the digital items.

WP4 has designed an ingestion work flow that will allow for an elaborate, visually guided ingestion of metadata in the repository. The fundamental principles include the disassociation of input metadata from existing metadata standards in order to avoid ambiguity over interpretation and, the ability to create and manage transformations that will apply to the actual metadata records, which subsequently (re-)define the input schema in a semantic, machine understandable way, based on its mapping to the reference model.

EUscreen user requirements and analysis studies, as well as meetings have shown that there is a variety of metadata standards used by the providers and an additional need for metadata manual and automatic enrichment. There is a limited number of key standards often suggested as best practice but there is still a long way to go to achieve interoperability. National standards in some countries are also a factor that needs to be taken into consideration. There is often the case where providers bypass the semantics of a standard through proprietary rules and conventions that are not sufficiently documented or semantically defined, resulting in misinterpretations that can not be straightforwardly detected. Finally, there is a wide, with respect to the level of detail, range in annotation from the use of a small flat-structured set of elements to defining and using complex schemata that support various levels of annotation, item and concept relations and connections with controlled vocabularies and thesauri.

In this context, WP4 has identified the intermediate Semantic Representation for EUscreen that will serve as the point of interoperability between the providers, as well as between the project’s repository and outside sources. These efforts are based on the European Broadcasting Union's EBU Core Metadata Set, an XML Schema proposed to aggregate

information within the EBU community. In the light of recent and future Europeana developments, the reference model also allows for the contribution of metadata along with digital representations/summaries of items. It delivers information in a 'self-contained' way including links to original repository and provides references to controlled vocabularies and authorities. In parallel to the work of WP4 regarding the metadata schema, WP3 will produce the EUScreen semantic elements that best fit in the EUScreen needs. This work will be based also on the Video Active metadata schema. A semantic mapping is required between EUScreen semantic elements and EBUcore. This will enable using EBUcore to store, export and preserve the metadata and using the EUscreen metadata elements in graphical user interfaces of the mapping and annotation tools.

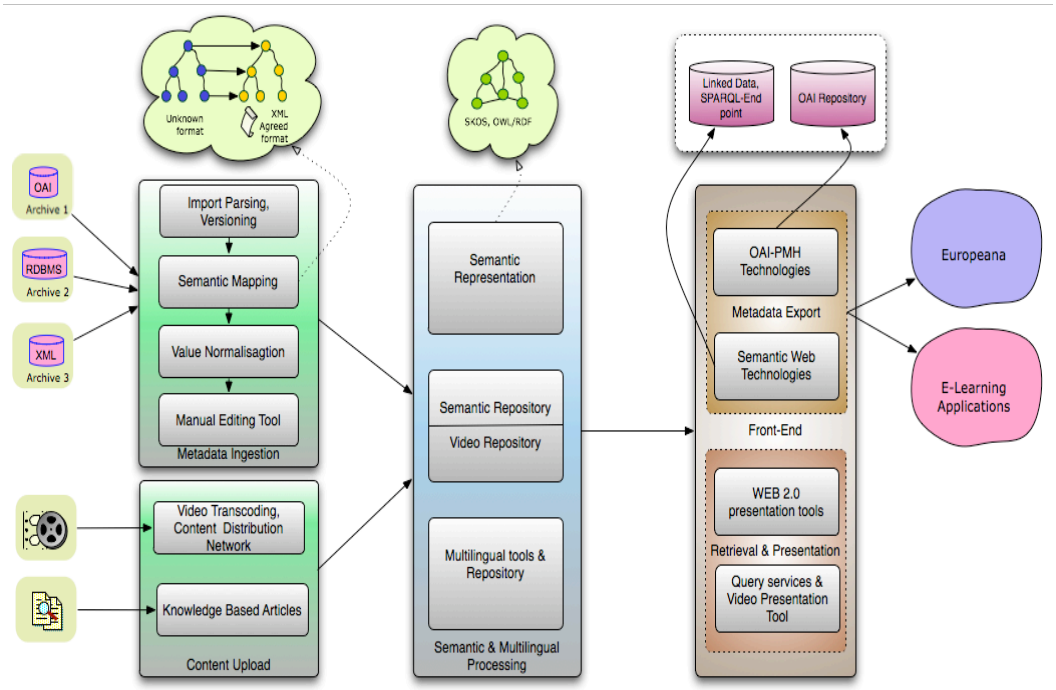

Figure 2.1: EUScreen Architecture

The Metadata Ingestion web service offers a user friendly ingestion environment that allows for the extraction and presentation of all relevant and statistical information concerning input metadata. The Semantic Mapping tool provides all the functionality and documentation required for the providers to define their mappings. Transformations are editable and reusable and can be applied incrementally to user input while providing, throughout all steps previews and visual indications to illustrate and guide user actions. One key issue addressed is the ability to semantically map metadata schemes through conditional mapping of input elements using respective transformation functions (e.g. concatenate) that will allow for the addition and enrichment of semantics even when those are not specifically stated in the input.

The architecture also determines the operational workflow processes to bring the amalgamated content of the partners into Europeana and to create, manage and execute, with the European Digital Library Office, the implementation plan to ensure the content's validity and visibility. The Metadata Export service supports exporting of the aggregated metadata to several established standards for metadata mediation. Primary showcase is the transformation of the aggregated content to the Europeana Semantic Elements schema and the deployment of an OAI-PMH repository to facilitate harvesting by Europeana.

Part of the functional requirements of the front end are presented in D5.1.1. The aim of D5.1.1 is threefold: 1) it aims to map and analyze the user requirements and use cases described in the D5.1 document, and translate these findings into feature list and functional specifications of the EUScreen portal; 2) it presents a set of open issues and questions that need to be solved as soon as possible before moving forward in the design and development process; and 3) it suggests a design process and schedule for the design of the portal.

3 Metadata Ingestion

The Metadata Ingestion Module aggregates, maps and transforms the providers' metadata to EUscreen metadata. Fig 3.1 demonstrates the ingestion workflow. It comprises of a number of steps needed to finalise the ingestion process in the EUscreen system and subsequently to Europeana. It is designed in a way to facilitate and increase the user involvement as well as providing a web-based user friendly environment taking into account the lack of technical knowledge of the people involved.

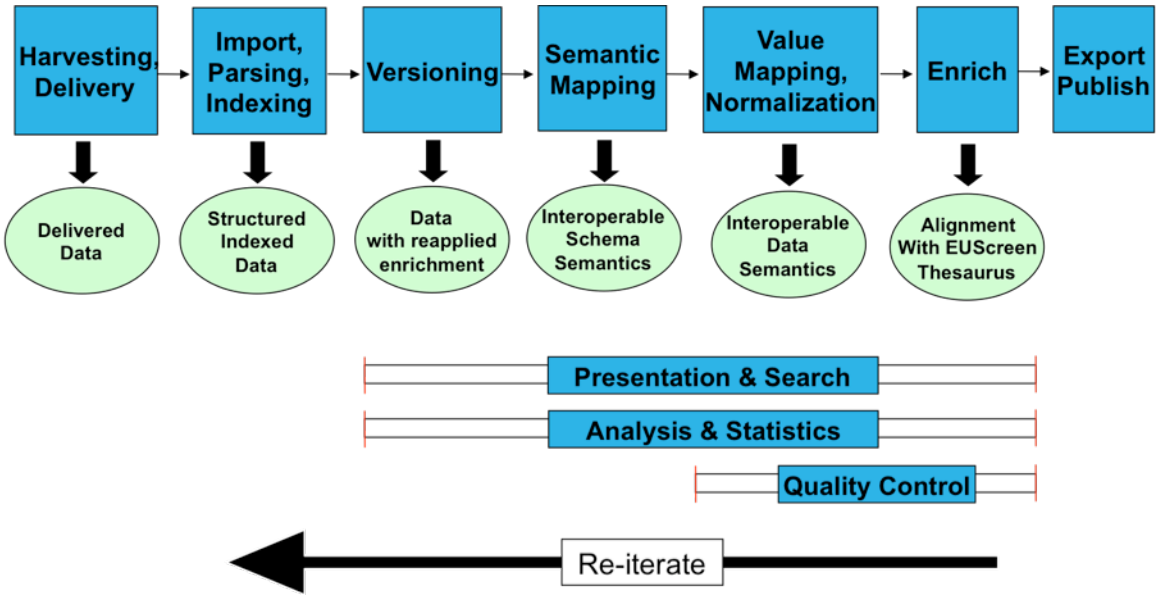

Figure 3.1 Ingestion workflow

The workflow consists of seven phases. Each phase is responsible for specific services needed to ensure the quality of the ingestion process.

3.1 Harvesting & Delivery

It is responsible to set up the environment within which the system will harvest the metadata from the repositories placed at provider's site. It is not using a predefined harvesting format at this phase since there is the need for harvesting, indexing and storage of all the available metadata from the content providers.

The system will be implemented as a web service, where authentication is required to perform a series of tasks that correspond to work flow steps. The service will be an application written in the Java programming language and hosted on a web server by the Tomcat servlet engine. Data is imported into a PostgreSQL database in xml format (as BLOB).

Once uploaded, the xml structure is parsed and represented in a relational database table. As this table can grow quite large it is partitioned into one partition per data upload. All data within one upload is treated as having the same structure, so it is not possible to upload different schemas (or more likely updated schemas) in one upload.

Most of the communication between the application and the database is implemented on the Hibernate framework, a high performance object/relational persistence and query service. This allows for powerful, yet simplified, management of housekeeping objects like Users, Organizations and Data Uploads while also providing additional functionalities such as integration with Lucene for indexing and querying data.

Once data is parsed into the relational table, indexes are built to allow quick access to any part or sub-tree of the xml-tree like data. These are currently constructed as PostgreSQL BTREE indexes; when content full text indexing is implemented, it will be based on Hibernate's search architecture. All further data manipulation such as mapping and transformation, normalization, enrichment, etc. is structured through the addition of extra tables annotating but not altering the original data. This allows easier comparison between uploads and facilitates the versioning strategy.

Content upload

Currently, allowed data formats for uploads are:

- XML in any schema.
- Excel files, based on a template
- zip archives of the above

Uploading of thumbnails or really any binary data will be supported in the future.

The following methods are supported for uploading content:

- HTTP upload; suggested only for relatively small amounts of data (<50MB)
- Upload to a dedicated FTP server
- Remote HTTP or FTP browsing
- OAI-PMH repository harvesting
- SuperUser uploading from local file system (restricted).

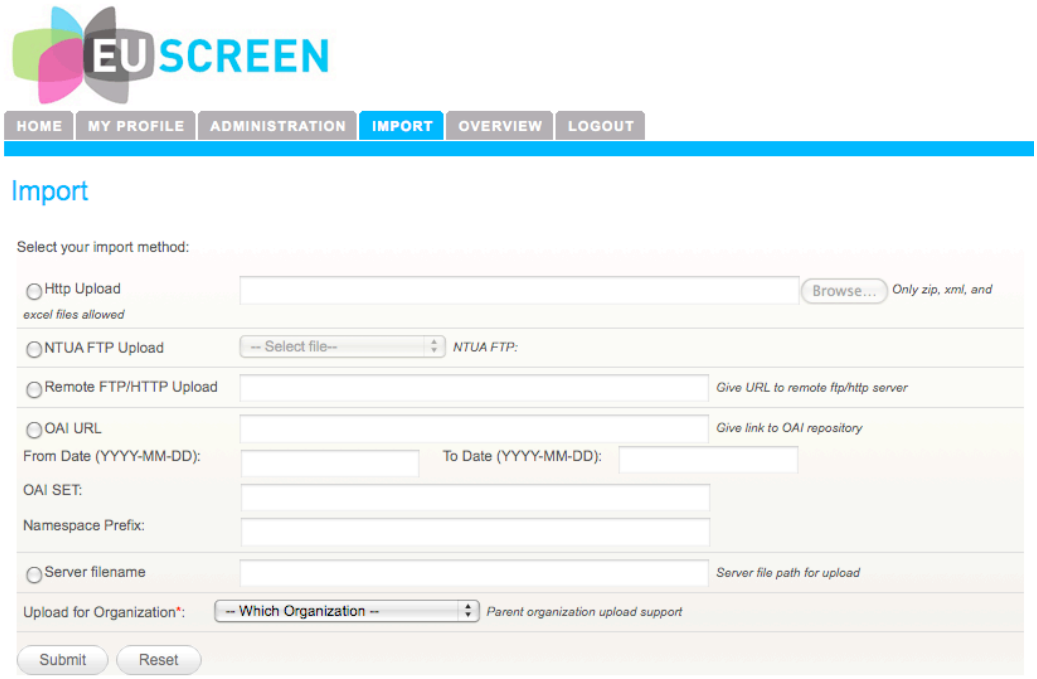

Figure 3.2 Import Methods

3.2 User and organisation management

The users belong exclusively to one organization and can not access data non related to that. They can be assigned with different levels of access, that grant roles ranging from data browsing, over editing and annotating, to being allowed to edit other users' details (administrators). Parent users will be allowed, for organizations that might not have expertise or manpower to use the system and thus, delegate the job to an organization which is then their designated "parent" organization. Parent users extend their rights to child organizations and provide the functionality to build the access hierarchy for any given country/thematic category. The current role set can be easily adjusted to allow more freedom in rights management.

The following rights are currently implemented:

- change/add/delete user
- change/add/delete organization
- edit/upload/delete data
- publish / declare finished datasets
- read-only browsing rights

These rights have been grouped to the following roles:

- Administrators (all user, data and organization rights for the organizations they manage),
- Annotators (data management rights),
- Publishers (publishing data rights),
- Data Viewers (simple viewing rights).
- The system also contains super-users that have full rights to all organizations and their data in the system.

3.3 Import, Parsing and Indexing

It allows for unified access to data. With this tool all Content Provider data is structurally accessible the same way, e.g. xpath-like access for subsequent processing.

3.4 Versioning

It is responsible for the management of uploads and the formation of the provider's input repository with respect to incremental or overlapping uploads. Sophisticated comparison algorithms that detect duplicate or edited items are applied for upload dataset comparison. This functionality allows for future development of services like:

1. Ability to transfer edits & enrichments on old versions of items to the current repository.
2. Merging uploads that require overwriting of existing items that were not edited or enriched.
3. Conflict resolution in cases where multiple edits occur for a specific item.

3.5 Semantic Mapping

It provides the service for assigning semantics to the harvested raw metadata. It is used to manually map providers' fields to a reference schema. This process correspond to the transfer of knowledge about the semantics of any given provider schema, even in the case of information that may not be included in a machine understandable way within a provider data upload, but will be registered by the user through the mapping process.

The mapping transformations are stored and can be reused while the provider is able to manage and edit them at any point. It will support constant value mapping, text element

concatenation (many-to-one mappings) and conditional mappings. Providers that have metadata in supported known formats may be able to omit this step by reusing stored transformations from selected schemas to EBUcore based on existing crosswalks.

Data integration processes comprise of various tasks including data matching, data transformation, and schema/semantic matching. Many solutions have been proposed by the community for each one of those tasks, ranging from applications that rely heavily on the user, to applications that are semi-automatic and in some cases completely automatic depending on the task, thematic category and schema complexity. For the case of schema/semantic matching many techniques and platforms have been developed, enabling the user to complete successfully the task. Notable cases are the schema mapping tool provided by Altova that offers a rich editing environment where the user is able to map any number of arbitrary schemas, but the whole process is totally manual, and the COMA++ platform that offers the user an environment for semi-automatic schema mapping, using state of the art algorithms. Although these approaches attempt to solve the general problem, the case of the EUscreen project has specific characteristics that lead to a more specialized solution to efficiently handle large amounts of diverse data and metadata.

By choosing a semantically rich and well-defined reference schema as the target of the mapping process the user has the opportunity to semantically enrich his data and metadata while quality of the aggregated content is ensured. The mapping process will be manual and the tool offers previewing, assisting and validation capabilities in order to ensure the quality of the result. The design principles of the mapping tool ensure the extensibility of the tool itself on a software level and of the system on a data level. It can be extended to support alternative target schemas that are represented by valid XML documents.

The mapping tool will be designed using a client – server approach. A subset of the functionalities is implemented as server side services, while the user interface is rendered on the client inside a web browser. The communication between the client and the server is achieved using AJAX calls. One of the core design concepts of the mapping tool is that the user should be able to use all the functionality he might need in order to achieve the best possible result with minimal effort. In order to achieve that, the tool must be intuitive and visual aiding and appealing. Another important design concept is that performance must be ensured because the EUscreen back end must be able to perform computationally intense tasks, e.g. metadata transformation and data parsing, without affecting the interaction between the user and the web service. This is achieved in a great degree by separating the interface rendering and the interaction with the user from intense tasks that are executed on the server side.

In order to offer a more user friendly environment to perform the task of schema mapping, the tool can be configured to provide to the user groups of high level elements that constitute separate semantic entities. These top level sets of elements are presented on the right side of the mapping tool. The User Interface is illustrated in Figure 3.3. On the left side, a tree structure is always present showing the target schema. The user is able to interact with this tree, expand or collapse the elements of the tree and retrieve brief statistics for each element and its values.

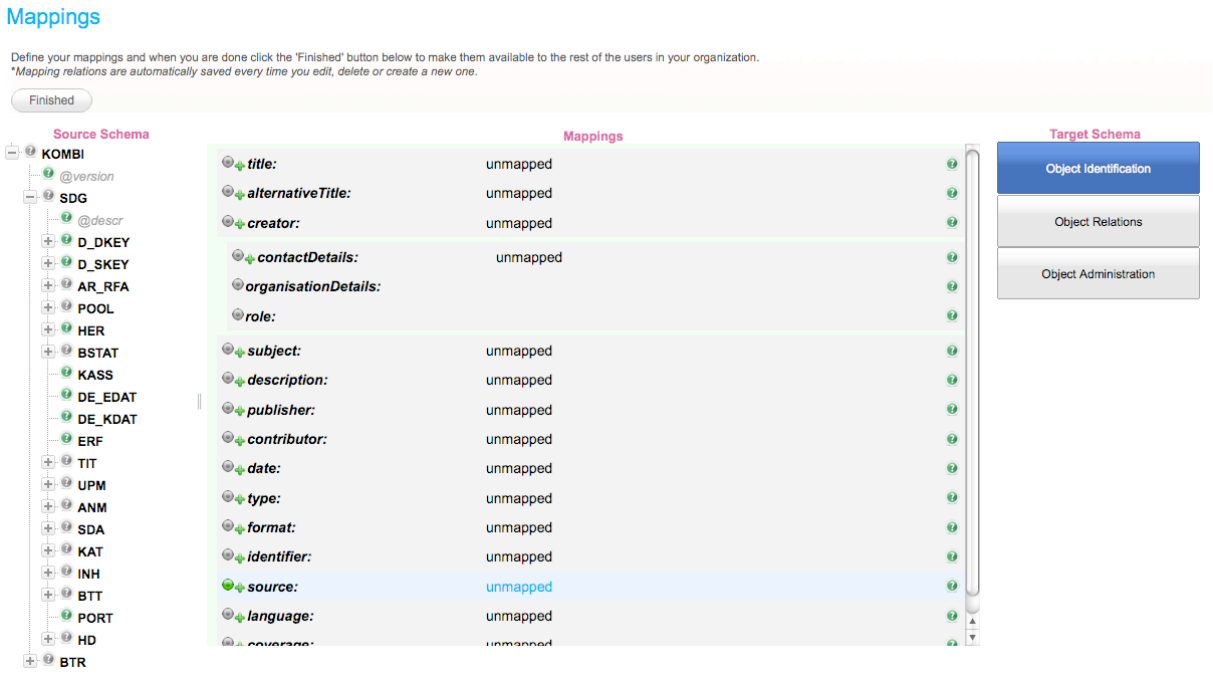

Figure 3.3 Mapping module

3.6 Value Mapping, Normalization

It allows for the alignment of controlled vocabularies between the input and target schema and for the normalization of element values according to required conventions. In particular:

- It enables providers to resolve data issues, e.g. map own terminology list to selected terminology lists.
- It automatically normalizes data e.g. dates, geographical locations, nationality/language.

3.7 Enrichment

It enables the addition of data that is not in the original metadata (e.g. empty fields, fields that take values corresponding to identifiers for controlled vocabularies). This service will also

enable adding values from the EUscreen Thesaurus. The enrichment will happen on item or group level.

3.8 Export-Publish

It will create the OAI-PMH repository with Europeana enabled data. This tool exposes metadata in the ESE schema for the Rhine release and subsequent data submissions to Europeana. To further enhance and test the benefits of data interoperability, WP4 will investigate for the final release of the system, the potential adoption of frameworks that will allow us to expose, reuse and connect to richer structured data. EUscreen partners are actively participating in several working groups within the Europeana family of projects and the project's aim is to establish the ability to adopt emerging guidelines and standards with minimum provider burden. NTUA is following the developments concerning the new Europeana Data Model and the subsequent release of the revised Europeana Semantic Elements schema that will adhere to the former.

3.9 Presentation & Search

It is responsible for previewing items and metadata in the EUscreen test environment. The Content Checker tool developed by Europeana will be also used to assess the validity and quality of the metadata to be provided in Europeana.

3.10 Analysis & Statistics

This service provides detailed analysis and statistics of metadata contributed by a provider (i.e. number of items imported, total values per field etc). The system generates and presents relevant information in any given step of the process, in order to assist users in accessing both data structures and underlying data and, validating data and its transformations. This information is calculated per import and provided based on user's access rights.

Statistics module is responsible to present to the user information from very large datasets in a consistent and visual appealing way. A minimal set of information is transferred to the server when the user invokes the statistics tool; this information consists mainly of a unique ID that identifies the specific import so the invoked module is able to load any data that is needed in order for the statistics to be calculated from the data layer.

When the user invokes the statistics module for a specific import, the first set of statistics is presented. It refers mainly to the Schema and the XML Element information that can be calculated for that import. The screen is separated in two distinct areas. The first one on the right is used to present the various XML schemas and prefixes for each of those that are present in this particular import. The second area of the first page consists of a view separated

with tabs, for each XML Schema presented on the table that was described previously, also one tab is created that holds a table with information regarding the elements found in the specific import and that belong to that Schema. This table has 4 columns where statistics are presented, the names of the columns and their roles are the following:

- **Element.** This is the name of the Element or the attribute of an element found in the import and that belongs to a specific XML Schema.
- **Frequency.** This is the frequency of a specific element compared to the number of distinct items that are present in the import.
- **Unique.** The numbers of distinct/unique values the element or attribute holds.
- **Length.** The average length of the values the element or attribute holds.

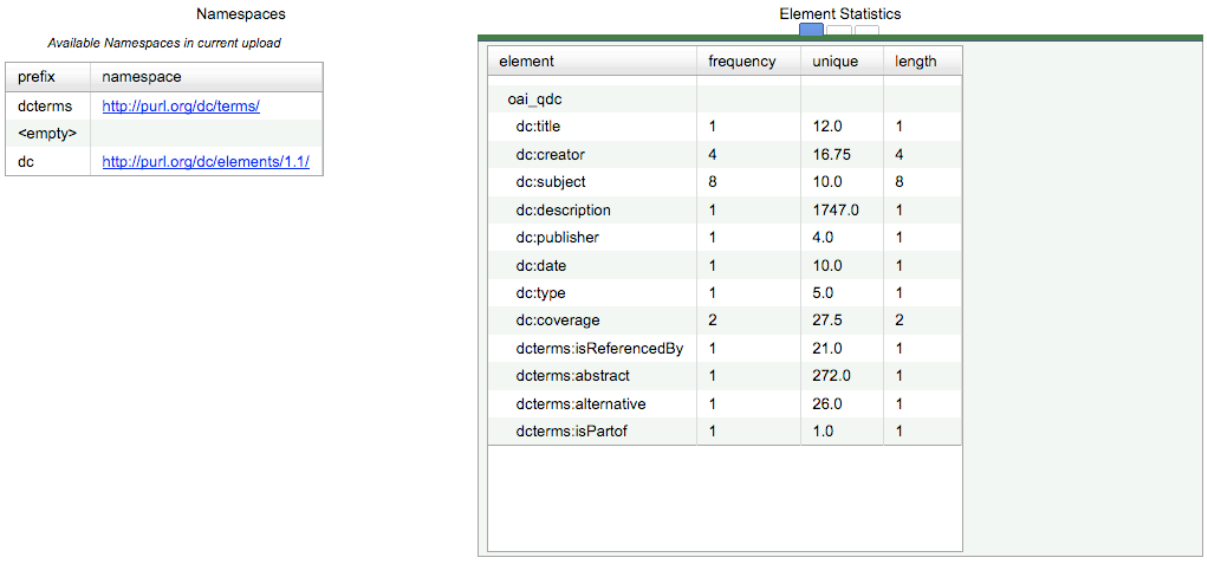

Figure 3.4 Statistics module

3.11 Quality Control

This service will automatically check and report on the Content Providers' data (i.e. missing values, malformed data). Error reports and warnings from the tool can be used to edit semantic mappings, value mappings and/or edit items until the Provider's data successfully passes the Quality control checks.

4 Portal

The EUscreen portal builds on the components used in the Video Active project. The portal contains a variety of services to satisfy the needs from the different user groups. It will provide a multilingual interface, audiovisual asset management options, transcoding functionalities plus miscellaneous search and browsing functionalities.

Part of the functional requirements of the front end are presented in D5.1.1. The aim of D5.1.1 is threefold: 1) it aims to map and analyze the user requirements and use cases described in the D5.1 document, and translate these findings into feature list and functional specifications of the EUScreen portal; 2) it presents a set of open issues and questions that need to be solved as soon as possible before moving forward in the design and development process; and 3) it suggests a design process and schedule for the design of the portal.

4.1 Server-less front-end

Implementation of the front-end services will not be done in the traditional way using server-side programming language like php, java or asp. We will further develop and implement our 'server-less' front-end APIs where a javascript/flash proxy system handles the communication with the back-end services. The result will be a front-end system that can be 'installed' on any plain html web server without any need for server-side technologies. This means it can be hosted and moved to any location or multiple locations. It also means partners can use these APIs to integrate parts of the functionality in their own intranet and internet systems using simple 'embed' ideas. This method is gaining more ground, for example companies like Google who provides these types of APIs for services like Google Maps.

4.2 Resources

The EUscreen portal will contain a rich collection of audiovisual television material and its metadata. The portal will support different types of multimedia resources. We distinguish four different types:

1. video
2. audio
3. images
4. documents

Each of the various resource types will have different requirements. The differences are apparent in the way the resources are treated throughout the system. For example, a document will be viewed differently than videos, images or audio files. Other aspects in which the requirements might differ are upload, view, search and toolset.

Resources of the same type may differ in format and type-specific specifications, such as dimension, colour space, compression, frame rate, bit rate and sample rate. In order to standardize certain portal options, we need some kind of standard format specification and conversion between formats. Additionally, the need for conversion arises from a client's perspective. Certain original resources, such as videos and images, might be really large and will consume a lot of bandwidth. Conversion between formats will be discussed in paragraph 4.4.1.

4.3 Upload

The EUscreen portal will feature two upload options:

1. Web upload
2. Java-based uploading

4.3.1 Web upload

The web upload enables the upload of content by using a web-based tool. The web upload option will be available for each of the four resource types.

4.3.2 Java-based uploading

For each content provider a java-based uploading service can be developed. The advantage of the java based service is the use of a small transcoder at each content provider to transcode the video. Only the transcoded file will actually get uploaded in EUScreen where the raw big-format source material stays at the content provider.

The java-based upload method can be proved the most efficient way to insert video items into the EUscreen system

IT is adviced to reserve a stand-alone computer at each content provider, preferable connected to the internet in a direct way (outside of the network) so Noterik can get remote access to the computer. The content provider can just connect a hard disk with material to the computer and after that Noterik can assist with the rest of the transcoding and uploading process.

The minimum requirements for the content providers' computers are:

- Windows Operating System
- AMD Athlon II X2 or equivalent
- 200GB HD space
- 2GB RAM
- 3Mbit bandwidth

4.4 Transcoding

Transcoding is the process of converting audiovisual material into various formats. Converting audiovisual material is necessary because it will allow us to standardise the formats that will be used in the portal and offered to the end users. The transcoding process will take place when a new asset is uploaded to the EUscreen platform. The process mainly focuses on converting audiovisual material. Additionally, when dealing with videos, screenshots will be extracted from the media files, which will be used in the portal for displaying search results.

4.4.1 Formats

The supported input formats for the various resource types are listed in the following table.

Video:	**Audio:**	**Image:**	**Documents:**
h.264	mp3	jpg	odt
ogg	wav	png	doc
MPEG-1			pdf
MPEG-2			
flv			

The different output formats are listed in the table below.

Video:	**Audio:**	**Image:**

h.264	mp3	jpg
	wav	png

4.5 Search

This section describes the different search options that will be available on the EUscreen portal. The EUscreen portal will provide novel browsing and viewing functionalities, such as faceted search, timeline search and geographical search.

4.5.1 Search filters

The search filters provide the user with faceted browsing capabilities. The user can enter a search string and make a selection of search filters. The results of the query will be specified based on the selected filters.

Filters will be categorized to facilitate the search process. Based on the Video Active project categories as shown below, the EUScreen categories will be defined.

1. type
2. production colour
3. transmission date
4. provider
5. languages
6. genre
7. topic
8. European dimension

4.5.2 Multilingual Search

Multilingual search enables queries and results presentation in multiple languages. The EUscreen platform will make use of the EUScreen Thesaurus to support multilingual search.

A third party software component, Thesaurix[1], will be used to manage the Thesaurus. Each content providers will be responsible for translating all terms.

The work done in the Video Active project, which currently contains around 1500 terms will be used and extended for the additional languages as stated in the DoW.

4.5.3 Timeline search

The EUscreen portal will feature a timeline-based search, which will provide a temporal search experience. All assets that contain temporal information, such as transmission date, will be inserted into a timeline. Users can enter search queries and use filters to limit the amount of results that are shown on the timeline.

4.5.4 Geographical search

The EUscreen portal will contain a geographical search functionality, which visualizes search results using the Google Maps API[2]. Only assets that are enriched with geographical information can make use of this search option. Users can enter search queries and use filters to limit the amount of search results returned.

4.6 Statistics

The EUscreen portal will support several statistics modules, serving different purposes.

1. Google Analytics
2. Custom statistics module

4.6.1 Google analytics

The EUscreen will integrate with the Google Analytics API[3]. It will provide the portal with a standard web analytics solution. It will allow us to analyse the portal traffic, discover how users find your site, where they are from, and how they navigate through it.

The Google Analytics dashboard will be available to certain project administrators.

1 http://www.joanneum.at

2 http://code.google.com/apis/maps/

3 http://www.google.com/analytics/

4.6.2 Custom statistics module

A custom statistics module will be available that will be responsible for storing portal specific information, such as most viewed asset, last viewed asset, most viewed asset (per provider), number of assets (per provider).

4.7 Play-out

The EUscreen portal will provide appropriate views of the four different resource types; videos, audio files, images, textual documents. Videos, images and audio files will be displayed within the portal, and a download link will be provided for the textual documents.

Videos can either be played from the streaming servers provided by Noterik, or from any other external streaming server capable of streaming h.264. However, when streaming from servers outside of the Noterik cluster, certain functionalities and use case scenarios will not be available.

Audio files are regarded as video files without visual content. The same player will be used to play both video and audio.

4.7.1 Display database

A display database/proxy will be implemented–that will allow to obtain the required metadata from external resources and merge them with the resource metadata. The primary interface will be the search interface. Other interfaces will be implemented to merge information created by use-cases or from user-generated content .

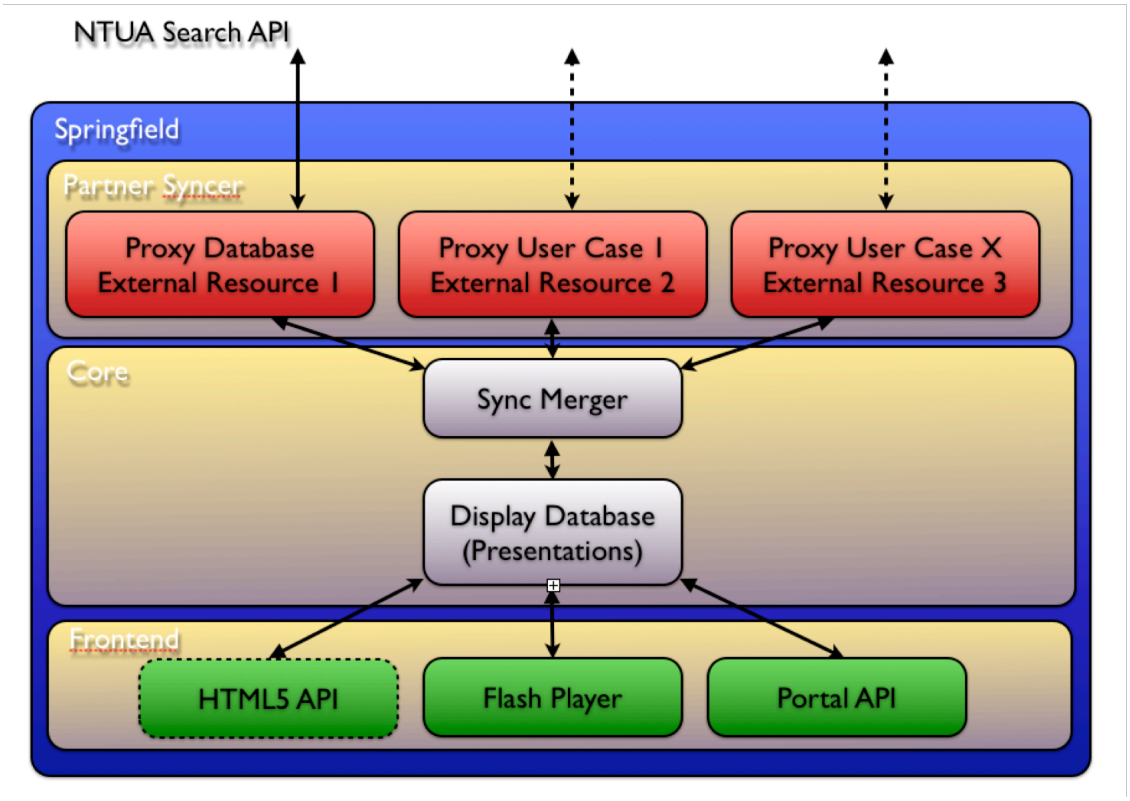

Figure 4.1: Display Database

In Figure 4.1 is illustrated how the display database is integrated into the system.

5 Integration

The EUscreen should be viewed as a collection of web services. Each web service provides different functionalities such as search, browsing, streaming, translation, or supplying metadata.

5.1 Software development method

The software development teams within EUscreen will follow an agile development approach based on the Boehm (1986) spiral life-cycle model, where requirements and solutions evolve through collaboration with interdisciplinary teams like curators and exhibition designers from cultural organisations. Software requirements are planned and implemented in small increments, leading to a higher number of iterations of advancing releases. Within EUscreen we are scheduling major releases for months 14, 24 and 30, but in between there will be several smaller releases (iterations) which will be made available to the stakeholders as soon as they become available.

For the ongoing development a development server (http://euscreen.devel.noterik.com/) will be set up. The development server will be used by all members of the EUscreen consortium to evaluate the work.

Once a functionality is updated, based on the feedback of EUscreen members and approved by the WP leaders, a copy of the software code will be made to the production server.

The development server is strictly for testing purposes, any user input made on this server (registering, uploading, tagging, etc.) will not be transferred to the production server.

5.2 Persistent identifiers

The EUscreen portal will contain many different web services, which all need to be integrated. In order to provide seamless integration of the web services, we need a way to define resources in a global manner. Each web service will have its own unique identification system, which assigns identifiers in its local system. The persistent identifier storage will have the ability to store and map the local identifiers from these systems if necessary.

5.3 User management

Both back end and front end tools are using different user management modules. The OpenID[1] technology will be examined to harmonise the user management modules of the EUscreen system.

[1] http://openid.net/

6 Conclusion

This report documents the technical specifications of the EUscreen system. The EUscreen system comprises of the back end and fronend tools. The back end tool is used to aggregate metadata from a diverse group of cultural heritage content providers, homogenise and align them with an established metadata schema standard that guarantees semantic interoperability. Functionality includes a user and organisation management scheme that supports appropriate user roles and access rights for simple organisations, import of arbitrary metadata schemes used by providers and serialized in xml, a statistical module for input data sources, a visual mapping module that functions as an xslt, transformation of imported data sources, publishing and exposing aggregated metadata in standard metadata schema. The EUscreen portal contains a variety of services to satisfy the needs from the different user groups. It will provide a multilingual interface, audiovisual asset management options, transcoding functionalities plus miscellaneous search and browsing functionalities.