Project Acronym: EFG1914
Grant Agreement number: 297266
Project Title: EFG1914

D6.1 EFG Metadata Schema Extension: Documentation

Revision	1.0
Date of submission	30 November 2012
Author(s)	Paolo Manghi, CNR-ISTI With input from Jan Kenter, Deutsches Filminstitut - DIF
Dissemination Level	Public

Project co-funded by the European Commission within the ICT Policy Support Programme

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

INDEX

1. Introduction

EFG1914 is a film digitisation project with three main aims:

- To digitise 654 hours of film and ca. 5.600 film-related documents on the theme of the First World War
- To give access to the material through the European Film Gateway and Europeana
- To build a virtual exhibition using selected objects digitised in EFG1914

Online access to the web files will be granted via the partner's own websites and via the European Film Gateway. Backbone of the EFG portal is a hierarchical database that can store extensive filmographic and object-related information. This database was built from scratch in the EFG project, which ran from September 2008 – August 2011. As it allows the storage of comprehensive metadata sets for various types of AV and Non-AV material as well as for persons, companies and events already now, it is fit to represent the data delivered for EFG1914 quite easily. Hence, after an evaluation of the expected content and needs for EFG1914, it was decided to carry out only minor changes regarding the data storage, affecting value lists in use rather than the creation of new entities. Instead of extending the schema further, CNR-ISTI, technical partner in the EFG and EFG1914 project, worked re-designed the Metadata Editor tool that was originally developed in the EFG project.

Chapter 2 of this deliverable summarizes the relevant aspects of the EFG XML schema defined in the course of the EFG project. Specifically, it describes the different interrelated types of entities that are part of the metadata schema. Subsequently, Chapter 3 introduces the metadata ingestion architecture in the different phases of ingestion, cleaning, editing and publishing, highlighting which parts have been refined or extended in EFG1914. The re-design of the Metadata editor tool will also be described in more detail in this chapter. The enhancement of an existing metadata editor of EFG considers the special needs of EFG1914 and the wishes of partner institutions.

Finally, Section 4 will then define the metadata schema modification policies to be undertaken in EFG1914, while Section 5 will be an ongoing section where such changes will be recorded in the months to come.Due to the development of a virtual exhibition (WP 7) the re-usage requests of metadata may change in EFG1914. Section 5 will cover the new requirements that are currently in development.

2. EFG Common Metadata Model and XML Schema

The EFG Common Metadata Model was designed after the analysis of the metadata models and schemas adopted within various organizations operating in the audio/video domain, starting from the data providers of the EFG consortium. This study took into consideration standards such as FRBR and Dublin Core, as well as more film-specific standards such as the Cinematographic Works Standards EN 15907. As a result, the EFG Common Metadata Model includes eight interrelated entities:

The *AVCreation* contains the properties of a cinematographic work: the film title, the record source (archive), the country of reference, the publication year, etc.

The *AVManifestation* contains the information about the physical embodiment of an audiovisual creation. Examples are archival copies (analogue or digital) and database files.

Properties of an AVManifestion include language, dimension, duration, coverage, format, rights holder, and provenance.

The *NonAVCreation* describes all non-audiovisual creations that can be represented in EFG. These are pictures, photos, correspondence, books or periodicals. The properties of NonAVCreations are: title, record source, keywords, description, date of creation and language.

The *NonAVManifestation* entity keeps track of copies of non-audiovisual objects. It has properties such as type (e.g. text, image, sound), specific type (e.g. photograph, poster, letter), language, dates (i.e. a date or period associated with the issue of the manifestation), digital format (including its status, size, resolution), physical format, geographic scope, rights holder.

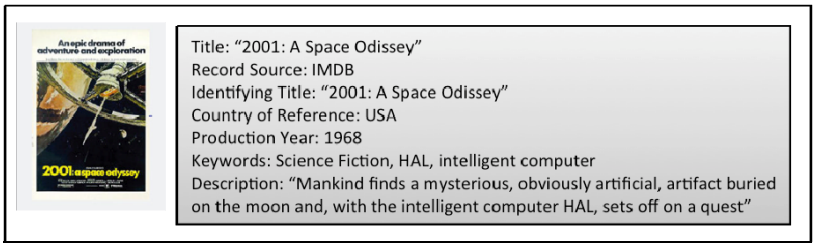
The *Item* entity points to the digital file held in the source archive. Its attributes are isShownBy (i.e. the URL reference to the digital object on the content provider's web site), isShownAt (i.e. the URL reference of the object in its information context), digital format, provider and country.

The *Agent* is defined as an entity that can perform an action. The model includes three agent types: Person, Corporate Body and Group. For example, the Person Agent has the following properties: name (composed of prefix, forename and family name), type of activity, date (which specifies the temporal properties of the person in relation with his activity), place (where the activity was performed), sex. Similar properties are defined for Corporate Body and Group.

The *Event* is an entity that can occur within the lifecycle of an audiovisual or non- audiovisual creation. Examples of Events are Physical Event (e.g. a public screening or a broadcast), Decision Event (e.g. when a manifestation of a creation was evaluated by a censorship body), IPR registration, Award (i.e. the award obtained by an audiovisual creation or an agent), Production event (e.g. dates and places where castings took place, dates and locations of shooting).

The *Collection* is defined as a compilation of creations (audiovisual or non- audiovisual).

In order to better illustrate the model and the relationships it defines among the above entities, we show a real-case example about the film "2001: A Space Odyssey" directed by Stanley Kubrik. We may have a record description of the AVCreation as follows:



The record description includes some metadata elements plus a thumbnail describing the AVCreation. We will have several AVManifestations associated to the AVCreation, such as all national versions of the movie, for example the Italian and the American versions. At the same time we may have several Agents related to this movie. As an example, we show a record description for the movie director, Stanley Kubrick:

Furthermore we may have NonAVCreations such as posters and film reviews. All these entities are connected through relationships (see Figure 1). The metadata record associated to each entity will be used to retrieve the archived object, while the relationships will be used to support browsing. As an example, it is possible to search for all movies directed by Stanley Kubrick in the '50s and browse all received awards, biographies of actors, etc.

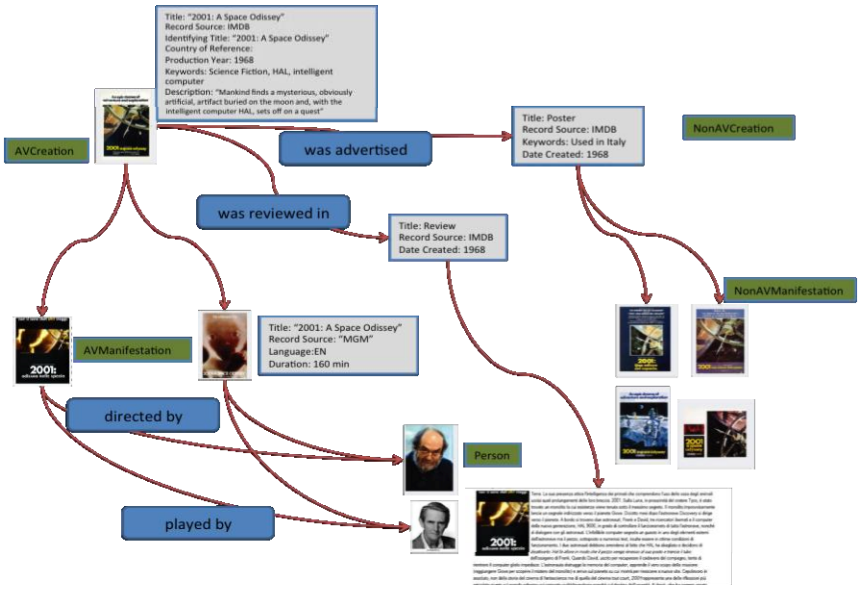


Figure 1 - Example of metadata associated for the film "2001: A Space Odyssey"

The EFG Common Metadata XML Schema implements the common model as described above. It defines XML element types and attributes for all the eight entities and their relevant properties. The common schema is conceived as the type union of eight XML schemas (one for each entity) in such a way that one EFG XML record represents one entity of the model together with its relationships (references) to other entities. Furthermore, the schema defines the so-called "controlled elements", which are the XML elements whose values must comply with a given vocabulary of terms.

2.1 EFG Schema and Europeana Data Model

Today, in order to adhere to the requirements of the EFG project, the infrastructure is configured to transform EFG metadata records into ESE metadata records and to export these to Europeana. ESE records are harvestable by Europeana via OAI-PMH as OAI sets where records are grouped by country. ESE has been replaced in 2012 with the new Europeana Data Model and EFG1914 must export records describing its digital objects that comply with this new model. On this respect, since ESE is a subset of EDM that includes all mandatory metadata fields (e.g., edm:type, edm:rights, edm:isShownBy, edm:isShownAt), the current infrastructure already guarantees an immediate compatibility towards Europeana. In fact, the current mapping from EFG records to ESE records will function as a starting point

to deliver basic EDM metadata records and meantime reach a richer and complete mappings. Indeed, thanks to the richness of the EFG data model, the final mapping will deliver EDM records complete of various optional and contextual entity fields. Examples are edm:ProvidedCHO, edm:WebResource and ore:Aggregation, or edm:Agent, edm:Event, edm:Place. The documentation available at http://www.pro.europeana.eu/web/guest/edm-documentation will be taken into account while defining semantic and structural correspondences among EFG and EDM entities. Until the final mapping to EDM will be completed, the EFG infrastructure will deliver EFG records as ESE records to Europeana applying the simple default ESE-to-EDM mapping described at http://www.pro.europeana.eu/web/guest/technical-requirements.

3. EFG1914 Data Infrastructure

The EFG1914 data infrastructure continues and extends the existing EFG data infrastructure to include new content providers (repositories or archives) and refine the possibilities of data curators willing to edit and publish content from such providers. More specifically, the infrastructure collects metadata records conforming to different schemas from different repositories, transforms them into records of the EFG common schema, and publishes them through the same index via APIs or the EFG portal.

As shown in Figure 2, for each repository joining the infrastructure the infrastructure implements separate workflows for *ingesting* the records (i.e., the action of collecting the metadata records from the repository), *transforming* the records (i.e., the action of transforming the records into the common EFG metadata schema structure) and *cleansing* such records (i.e., the action of aligning the semantics of the transformed records with the one of EFG schema semantics). Once the records are cleansed, data curators may decide to *edit* such records in order to change or fix or enrich their content before being published. As we shall see in the following, the EFG1914 infrastructure has changed the record editing policies, switching from an edit-and-freeze approach to a dynamic update approach. The metadata editor (MET) of EFG does not fit in every detail to the needs of EFG1914 for several reasons. One argument to establish a new version of the MET was to change the edit-and-freeze approach to a more flexible method for EFG1914 needs. For EFG1914 it is absolutely necessary to provide possibilities to enrich metadata records, which was not possible in the old MET in an easy way. Further details are given in section 3.3. of this document.

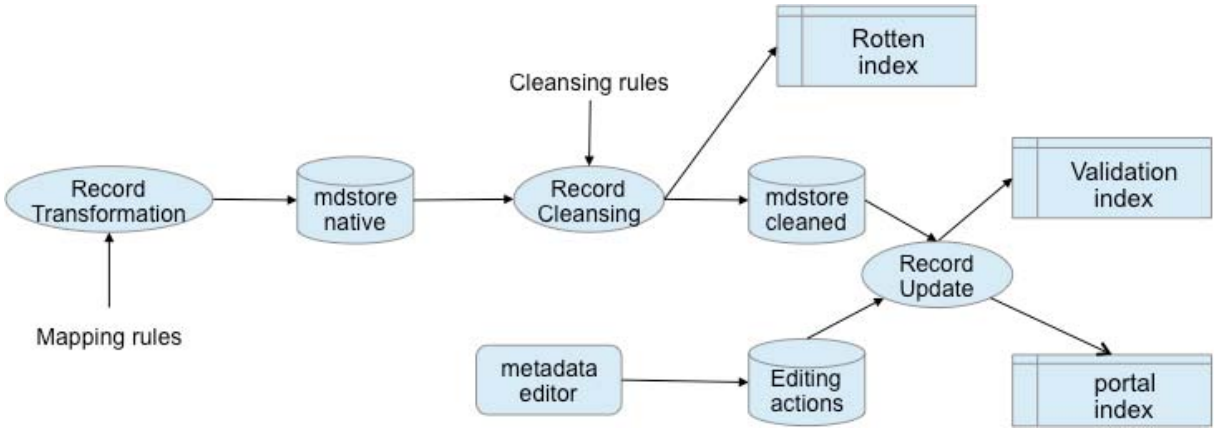


Figure 2 – EFG1914 Workflows: ingestion, transformation, cleansing, editing.

3.1 Record Ingestion and Transformation

Original (XML) records are exported by content providers according to various methodologies, including automated collection via standard access protocols (OAI-PMH) or upload of XML files into a common document repository (BSCW). Records from the same repository conform to the same XML schema, while different repositories generally feature different schemas. As a consequence, the infrastructure must consider custom mapping rules (i.e. XML scripts) to transform XML records from a repository/archive into EFG records conforming to the common EFG XML schema.

Record transformation is performed by data curators, who are system administrators in charge of (*i*) constructing and refining along time the XML scripts according to the inputs provided by content providers, and (*ii*) applying such scripts to supervise the effect of the transformation. Mappings from local XML schemas onto the EFG schema are to be defined by content providers; such information is delivered to data curators using special correspondence tables, associating input XML elements to EFG XML elements.

Data curators execute both ingestion and transformation of records for a given archive manually. The result of the process is stored into a so-called *Native MDStore* (MD stays for MetaData). Any change to the local XML schema as well as any modification to the mapping requires a relative refinement of the XML scripts and a consequent re-execution of the ingestion/transformation workflow.

Note: currently the ingestion and transformation workflow cannot be fully automated (given XML scripts) due to frequent errors in the data (e.g., records may change in structure, XML structure may be corrupted) and to the mechanisms used to export the records, which are generally unstable (e.g., unreachable servers, partly OAI-PMH compatibile). We envisage some experimentations of automated ingestion/transformation with those archives featuring a consistent and regular setting of record exports.

3.2 Record Cleansing

The next action after record transformation is that of record cleansing. Records in *Native MDStores* are processed in order for the values therein to be converted onto the formats and vocabularies provided by the EFG schema. The action does not affect the structure and focuses on value normalization, in order to enable meaningful searches and browsing over a uniform information space. Besides normalization, the action identifies back-links missing between records and adds them to the records, so that two-ways browsing can be implemented. For example, AVcreation records can be associated to the relative Person records and vice-versa. The cleansing process redirects “rotten” records, i.e., those records whose values are missing or cannot be converted according to the EFG schema, into a special index, where content providers can check them out and apply the required fixes. Records that are successfully cleaned are stored into a so-called *Cleaned MDStore*, where they can be edited and validated.

3.3 Record Editing

In EFG1914, content providers are provided with a Metadata Editor tool (MET), through which they can edit the cleansed records before they are made public. In particular, given one record of interest, they can define a set of XML element updates or XML element addition actions to be applied to the last harvested copy of the record. Such changes are

applied before indexing to the last ingested version of the edited record and their effect is limited to the XML elements touched by the editing.

An important use case of the MET in the context of EFG1914 is the ability to enrich AV and nonAV metadata in the aggregated information space with Library of Congress Subject Headings (LCSH). Since most archives do not use LCSH headings in their local databases, the relative metadata exports do not include this information. The new MET will allow to enter LCSH headings to metadata without loosing them again after an metadata update from the local archive. Other examples for the usage of the editing possibilities are changing URLs to thumbnails or digital objects or missing keywords.

Note: the dynamic update approach described above is less invasive and blocking than the one adopted by the previous Metadata Editor Tool in EFG. Such tool was designed to support content providers with interfaces to edit the elements of a given record A taken from (i.e. looked up) the EFG information space. Editing A would generate a new edited record A'. The harvesting and indexing workflow, taking records from archives to the EFG information space, would always intercept versions of A and replace them with A' in the EFG information space. As a consequence, if the archive hosting A would modify its elements, not necessarily the elements once edited to generate A', such changes would not propagate to the EFG information space after recollecting content from the archive. In EFG1914 such an approach was considered too rigid and the tool was re-designed to re-apply edits dynamically and element-wise to the incoming records.

3.4 Record Validation

Once the records have been cleansed (and possibly edited), they are submitted for validation to the relative content providers. Content providers are provided with an index through which they can verify the structure and content of the records before committing the current ingestion/transformation/cleansing workflow and make them available to the public. The possibility for content providers to commit the records to be indexed has been realized in the context of EFG1914 as it was missing in EFG.

4. EFG Schema: Modification Policies

Although no changes to the EFG schema have been identified so far, we envisage these will be required as new content providers may join the infrastructure and new end-user requirements will surface. By observing the workflows described above it is clear how changes to the schema may have a strong (and expensive) impact on the overall ingestion/transformation/cleansing workflows. In fact, mappings and cleansing rules may be invalidated for all repositories registered to the infrastructure and new interactions between content providers and data curators should take place to adjust them. In addition, the Metadata Editor would require adaptation to such changes. Such a process has a high human cost and simply cannot be fired whenever a schema change is identified and required. It is therefore good practice to concentrate schema updates in a well-anticipated period of time, which in EFG1914 has been established every March, starting from the 1st of the month. Before such a date all changes to the EFG schema collected during the previous year should be collapsed into a new schema version, which will be valid for the year to come. In particular, the Consortium opted to limit schema changes to those strictly required by EFG1914 end-users and by Europeana end-users. More specifically:

1. EFG1914 end-users: among such changes, the Consortium decided to accept those "compatible" with the portal, i.e. that do not require re-coding of the EFG portal; such

a constraint has been temporarily imposed due to the costs implied by changes to the portal code, which are currently not sustainable by the project.

2. Europeana end-users: these include changes required to deliver richer information to Europeana; such information may not therefore be visible through the EFG1914 portal, but make more appealing and therefore increase the visibility of EFG1914 objects through the Europeana information space.

5. EFG Schema: Modifications

The richness of the EFG data model made it possible to define direct mappings from records exported by new EFG1914 content providers into the EFG schema. After our initial analysis it seems that only minor changes will be required and mostly in the vocabularies rather than in the XML schema structure. In this section we shall describe and collect such changes as they are surfacing in the course of the project, in order to upgrade the infrastructure accordingly every year. So far, the following descision regarding the EFG1914 Virtual Exhibition was taken.

Virtual Exhibition

The *virtual exhibition tool* will offer end-users the tools to create and visually access a set of *virtual exhibitions* (VE). These are intended as “learning objects”, which combine new text (written by the creator) with objects from within the EFG1914 information space to form a narration or a story. The aim of such “stories” is to deliver an entertain-and-teach logic to end-users. As such VEs will be new EFG1914 objects to be preserved in a dedicated repository and to be searchable through the EFG1914 information space. After discussions with ATHENA and DIF it was decided to treat VE elements as non-AV textual objects in the EFG information space.

- The EFG schema should enrich the description of item types for non-AV objects with an entry “Virtual Exhibition”; optionally, the item schema could be extended to capture the narrative essence (e.g. references to other objects) of VEs.

It is not necessary to change the schema for VE items while they could be stored as nonAV entities in the EFG information space. VE items become searchable through the EFG portal webpage. To display the VE objects on the EFG portal just minor changes are needed.

Vocabularies

EFG1914 will bring in new terms in the vocabularies as well as new vocabularies. CNR-ISTI will implement user friendly interfaces for vocabulary management. The time of delviery of such tools is to be decided, considering the availability of CNR-ISTI resources (the tool was not included as a task in DoW).