Associazione Investire in Cultura AIC

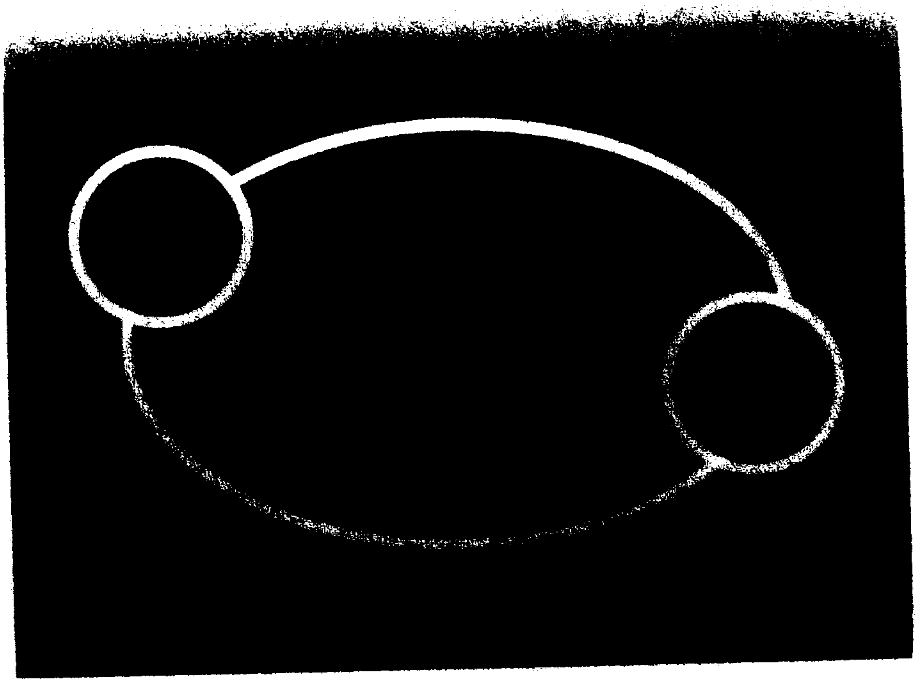
FONDAZIONE ROMA MEDITERRANEO

PROCEEDINGS

4th International Congress on

"Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin"

VOL. I



Cairo, Egypt
6th – 8th December 2009

ISBN 978-88-96680-31-5

9 788896 680315

SEARCHING AND BROWSING FILM ARCHIVES: THE EUROPEAN FILM GATEWAY APPROACH[1]

FRANCA DEBOLE[a], PASQUALE SAVINO[a], GEORG ECKES[b],
[a]*CNR-ISTI, Pisa*
[b]*Deutsches Filminstitut – DIF, Frankfurt*

Keywords: Video Archives, metadata, film, content-based search, semantic content description

1. Introduction

Several countries and film archives are doing significant investments in digitizing existing collections of moving images and cinema-related material (audio documents, photographs, posters, drawings, text documents). At the same time, there is a significant growing interest to have access to film repertoire, composed of movies, documentaries, scientific videos, and all the material related to them, while existing tools and services do not offer a simple and complete access to the videos that are becoming available. This is due, on one side, to the poor description of the videos, and on the other side, to the limited capabilities offered by search and access services. In many cases, videos are described by using very limited information, such as title, author name, etc.. Furthermore, descriptions available in different archives are very different so that it is difficult to have a uniform access to videos coming from many different archives. Finally, the values used to describe videos, e.g. person names, places, film names, etc. are not taken from unique authority files, so that when these data are used to search videos, many relevant items are not retrieved. Another relevant issue is related to the need of efficient tools that simplify the process of query formulation – possibly including multilingual access – support for browsing of relevant object, and effective presentation of retrieved object.

All these requirements conducts to recognize the need for a precise strategy for interoperability between different archives, which supports interoperability at the level of the format used to present object descriptions as well as at the level of its semantic description. This means that we need a metadata schema that is powerful enough to describe the richness of film content, but at the same time is flexible enough to support interoperability of existing archives through a simplified conversion of existing descriptions into the common schema. At the same time, it is necessary to have the possibility to define common authority files for the most relevant metadata elements, and to have the possibility of cleaning the data coming from different archives by using the authority file values.

This problem has been addressed within the EFG (European Film Gateway) Best Practice Network[2] funded by the European Commission under the eContent*plus* programme[3]. The Project started on 1st September 2008 and will have a 3-year duration. It currently assembles 21 partner institutions from 15 European countries and aims at building a single access point to digital collections of film institutions in Europe. As an aggregator project towards the European Digital Library, *EFG* aims at enabling Europe's film archives and cinémathèques to contribute their rich and valuable collections to *Europeana*[4]. It aims to provide direct access to more than 700.000 digital objects including films, photos, posters, drawings, sound material and text documents.

2. Overview of problem solution

The metadata schema should be able to describe at a sufficient level of detail many different types of information, ranging from audio/video objects to non audio/video material such as images, posters, documents, etc. as well as persons, events, access rights. Furthermore, the model must support the description of relationships between different elements. It is also required that importing data from different archives is simple, fast, and without loss of details in the description. The model should allow one to describe the content of each digital object, as well as information about the owner of the object, the property rights for its use, who are the persons or companies involved in its creation, etc.

As already underlined, user satisfaction and the success of any service that will use these data would require a significant number of digital objects available, but also an high data quality, which means high video quality, and uniform metadata values certified by experts. High video quality is guaranteed by the adoption of standardized ways for digitalization, high bit rates, and the use of standard video formats. Uniform metadata values become possible if the system adopts and enforces the use of authority files for the most important metadata values.

A third issue is the need of a specific tools that supports the access to the videos, based on their content. A specific requirement that derives from the integration of different archives in many countries, is that multilingual searches must be supported.

All these issues have been addressed in the EFG Project [2]. Our first activity consisted in the definition of a metadata model which supports all requested needs. The model will be described in some detail in the next section. In order to support interoperability with other archives, i.e. the possibility of importing and exporting metadata and multimedia objects from and to external archives, we developed a specific conversion module from native archive formats into the EFG metadata format and from the EFG metadata format into OAI-PMH [3]. The first component, which has a part that is archives-dependent (i.e. is we need to import data from a new archive, a new component must be implemented), maps each native metadata element into one of the EFG metadata elements. The definition of mapping rules requires the knowledge of the native archive format, apart that of the EFG metadata format. The module that supports the export of metadata represented in OAI-PMH, offers the possibility of transferring the data from the EFG archive into any other archive that accepts data represented in OAI-PMH. In particular, in the Project we will ingest most of our data into Europeana [4], so that the rich video archives can be accessed through the portal of the European Digital Library.

As soon as data are ingested into EFG, they must be "cleaned" and validated before making them available to end users. The cleaning process makes use of a set of thesauri and controlled vocabularies that are used to convert original data values into uniform values. During this phase, if an archive is ingesting a movie with the name of an actor that is not present into the authority file, the system will propose, by using similarity criteria, one or more possible alternative values. The archivist can either accept one of the proposed values, or he can decide to extend the thesauri with the new name. During this phase, the system detects possible connections between the new ingested item and those already existing. All these relationships are verified by the archivist for final approval. Finally, the archivist can create new relationships between the new items and those already present in the archive. All these functionality are offered through a tool for the management of authority files and metadata editing.

After completion, all validated data are ingested into the archive and made available to end users. The archive is based on the D-Net storage service [1]. The system foresees two different user categories: the archivist – who has the possibility to search all objects archived and to edit their descriptions, if necessary – and the generic user – who has the possibility to search, to browse the archive, and to visualize its content (including videos) through a web portal.

Finally, since the interoperability with other systems and platforms is one of the key issues addressed, the system offers the opportunity to export any object or groups of objects by using the OAI-PMH protocol.

3. The EFG metadata schema

The main step in building the strategy described in previous section is given by the definition of a common metadata schema to which the proprietary schema of the EFG content providers, and later on other schema, could be mapped. In order to do this, we took three specific requirements into consideration:

1. The need to meet the specification of the user requirements. These provide an indication of the concepts that are required by the user and the relative importance of those concepts.
2. The need to represent the concepts that are present in the data. By examining the data it is possible to determine the issues that arise when representing concepts relating to film objects.
3. The need for interoperability. It is necessary to be able to map from content provider legacy metadata to current "standard" metadata schemas.

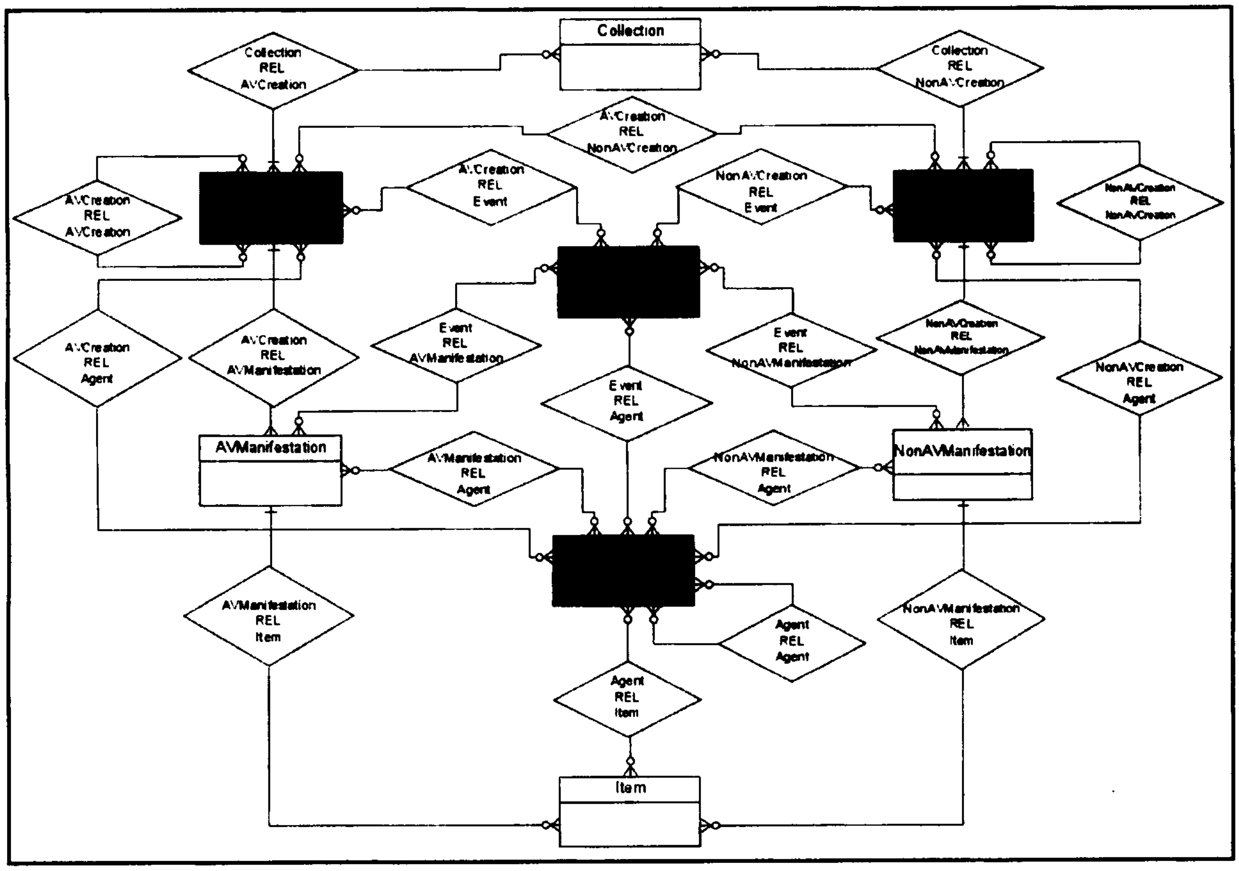

Collection
Collection REL AVCreation
Collection REL NonAVCreation
AVCreation REL NonAVCreation
AVCreation REL AVCreation
AVCreation REL Event
NonAVCreation REL Event
NonAVCreation REL NonAVCreation
AVCreation REL Agent
AVCreation REL AVManifestation
Event REL AVManifestation
Event REL NonAVManifestation
NonAVCreation REL NonAVManifestation
NonAVCreation REL Agent
Event REL Agent
AVManifestation
AVManifestation REL Agent
NonAVManifestation REL Agent
NonAVManifestation
AVManifestation REL Item
Agent REL Agent
NonAVManifestation REL Item
Agent REL Item
Item

Figure 23 - Strucure of the EFG metadata schema

This schema is the result of a study of the metadata schemas and semantic resources widely used within the organizations belonging to this specific domain as well as an analysis of the content and associated metadata to be used in the project, as selected by the Content providers of the EFG consortium. This study took into consideration standards such as FRBR and Dublin Core, the Europeana schema as well as more film-specific standards as the evolving Cinematographic Works Standards EN 15907.

The EFG schema is composed of eight major entities, which are defined as the topmost level of description that can have relationships to other entities. In order to describe objects, the EFG metadata model basically distinguishes between three levels: Creation, Manifestation and Item. "Creation" is the topmost level of description, which can be either an audiovisual work (represented in the "AVCreation" entity) or a non-audiovisual work (represented in the "NonAVCreation" entity). This applies also to the "Manifestation" level which is represented in the entities "AVManifestation" and "NonAVManifestation". The "Item" entity functions as a logical wrapper for the digital file presented in the EFG portal. Moreover, "Agents", "Event" and "Collection" also have entity status.

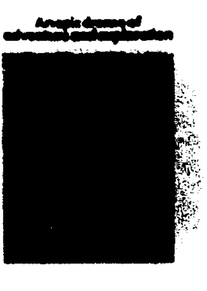
Each major entity of the schema comes with elements, attributes and relationships. Figure 23 illustrates all entities represented in the model and the relationships that may exist among them.

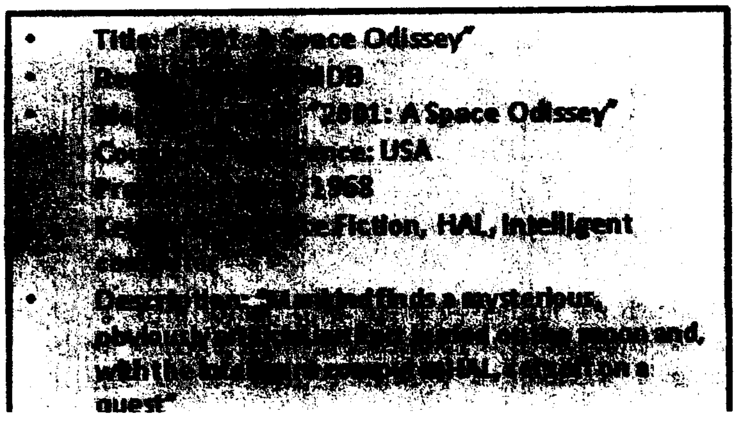
In particular, the eighth main entities of the EFG schema[5] are as follows:

- The AVCreation contains the properties of a cinematographic work. For example, it includes the film title, the record source, 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. Examples of the AVManifestation properties are again Title, Record Source, and Language, as well as 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 Created, Language.
- The NonAVManifestation has the function to keep track of copies made of non-audiovisual objects. It has properties such as Title, Record Source, 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 repository. Its attribute are isShownBy (i.e. the URL references 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, 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 elements: Name (which includes one or more Parts, such as the prefix, the forname, family name, etc., it also include the geographic and temporal scope), Date (which specifies the temporal properties of the person in relation with his activity), Place (where the activity was performed), Sex, Type of Activity. Similar elements are defined for Corporate Body and Group.
- The Event is defined as a primary 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 by a compilation of creations (audiovisual or non-audiovisual).

Relationships are defined between these entities: for example an audio/video creation is linked to persons (actors, movie maker, etc.) and to non audio/video material such as posters, interviewees, newspaper articles. They can also be related to events, such as awards received, public screening.

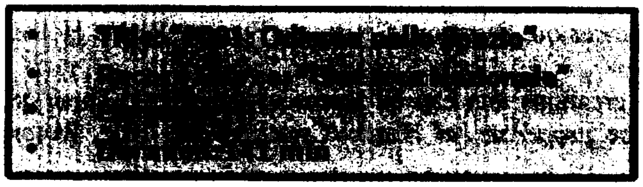
In order to better illustrate the model, we show how it can be used in practice. Let us suppose we want to model the movie "2001: A Space Odyssey" directed by Stanley Kubrik. We may have a record description of the AVCreation which is as follows:



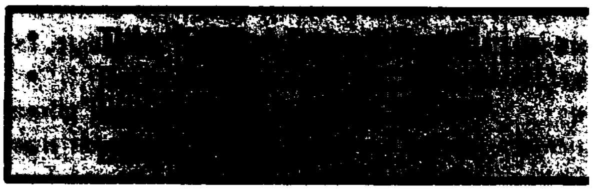


- Title: [illegible] Space Odissey”
- [illegible] DB
- [illegible] “2001: A Space Odissey”
- [illegible]nce: USA
- [illegible] 1968
- [illegible] Fiction, HAL, Intelligent
- [illegible] finds a mysterious, [illegible] and, with [illegible] HAL, [illegible] a quest”

It includes some metadata elements plus a thumbnail describing the AVCreation. We will have severa AVManifestations associated to the AVCreation, such as all national versions of the the movie. As an example we show the Italian and the America version, together with two thumbnails.

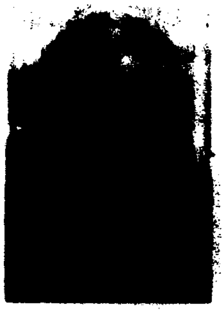


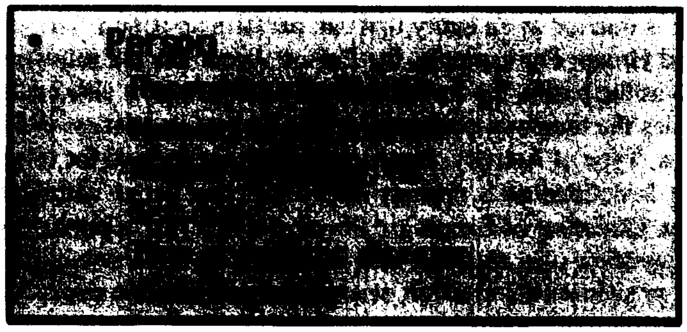
- [illegible]
- [illegible]
- [illegible]
- [illegible]



- [illegible]
- [illegible]
- [illegible]
- [illegible]

We may have several Agents which have relationships to this movie. As an example, we show two persons, Stanley Kubrick – who was the movie director – and Keir Dullea, who played the role of Dr. Dave Bowman.

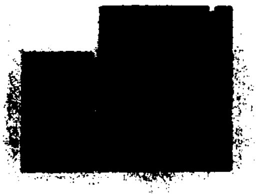




- Person [illegible]

We may have NonAVCreations such as Posters, as shown in the next figure, and film reviews.

- [illegible]
- [illegible]
- [illegible]
- DateCreation [illegible]



All these entities are connected through the relationships that exist among them, as illustrated in Figure 2. For example, we may express that the AVCreation "was advertised" through a certain number of posters, and it "was reviewed" in several articles. The movie "was directed" by Stanley Kubrick and it "was played" by Keir Dullea.

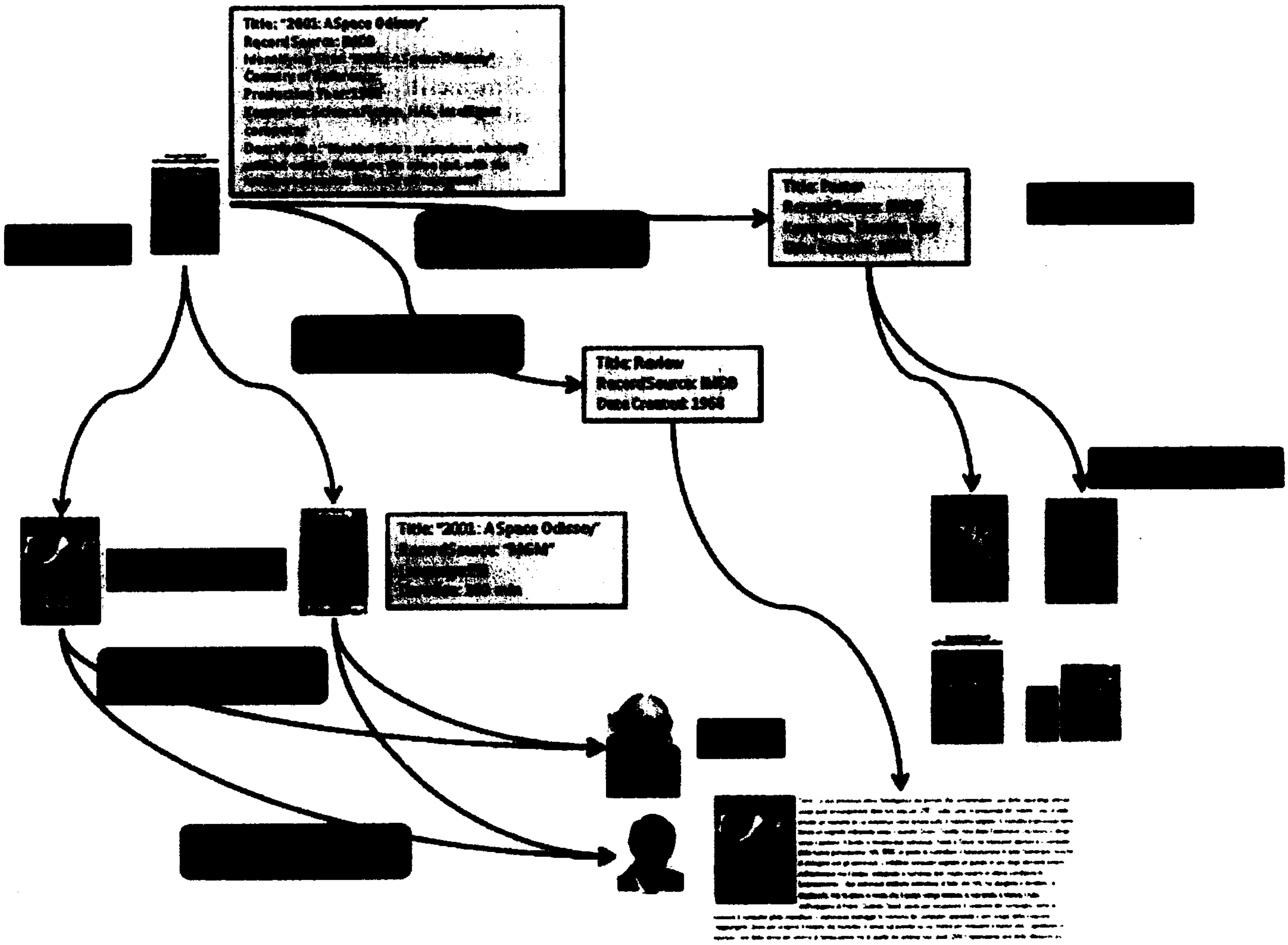


Figure 24 - Example of metadata associated to the movie '2001: A Space Odyssey'

All metadata elements and the relationships can be used to search and browse the archive. For example, it could be possible to search for all movies directed by Kubrick, or for all movies that were reviewed in a given magazine. It could also be possible to browse through the movie directors, movie players. When these movies are displayed, it could be possible to look at all awards received, at biographies of other actors, etc.

4 Conclusion

The *EFG* project has started in September 2008 and will run for three years. Currently the data ingestion has started and a first version of the web portal will be accessible online by end 2010.

References

[1] D-Net V1.0, http://www.driver-repository.eu/DRIVER-News-Events/PR_D-NET_1_0.html

[2] EFG – The European Film Gateway, http://www.europeanfilmgateway.eu

[3] The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] Europeana, http://www.europeana.eu/portal/

[1] This work is partially supported by the European Commission under the eContentplus programme, project EFG, contract 517006. The authors are solely responsible of this paper. It does not represent the opinion of the EC and the EC is not responsible for any use that might be made of data appearing therein.

[2] http://www.europeanfilmgateway.eu

[3] http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm

[4] http://www.europeana.eu

[5] http://www.europeanfilmgateway.eu/downloads/D2 2_Common_Interoperability_Schema_final.pdf