

ECP-2007-DILI-517006

EFG – The European Film Gateway Final report on type and quantity of archival resources tagged

Document number	<i>D3.2</i>
Dissemination level	<i>Public</i>
	<i>Revised public version: 02.11.2011, internal confidential version: 02.10.2011</i>
Delivery date	
Status	<i>Final</i>
Author(s)	<i>Francesca Schulze (Deutsches Filminstitut), Pernille Schütz (Det Danske Filminstitut), Uffe Smed (Det Danske Filminstitut)</i>



eContentplus

This project is funded under the *eContentplus* programme, a multi-annual Community programme to make digital content in Europe better accessible, usable and exploitable.

Table of Content

Institutional abbreviations and acronyms	3
1 Executive summary	4
2 Introduction	6
3 Cataloguing within EFG Data Ingestion and Editing Workflow	7
4 Data Cleaning and Enrichment Activities in the Source Databases	12
4.1 <i>Evaluation Questionnaire</i>	12
4.2 <i>Achievements: Local Cataloguing Work</i>	13
5 EFG's Vocabulary and Matching Work	15
5.1 <i>EFG Controlled Vocabularies: Scope</i>	16
5.2 <i>Value Lists</i>	17
5.4 <i>Semantic Relationship Vocabulary</i>	18
5.3 <i>Vocabulary Matching Experience</i>	19
5.5 <i>Achievements and Lessons Learned</i>	23
6 Cataloguing on EFG level by using the Metadata Editor	25
6.1 <i>Achievements: Cataloguing with the Metadata Editor</i>	25
6.2 <i>A Cataloguing Example</i>	27
7 EFG's Approach to Authority File Building	29
7.1 <i>Challenges</i>	29
7.2 <i>Doublet Cleaning with the Authority File Manager</i>	30
7.3 <i>Achievements and Lessons Learned</i>	34
8 EFG Data Quality Workshop	36
9 Evaluation of EFG WP 3 Work by the Partners	38
9.1 <i>Questionnaire for Partner Archives</i>	38
9.2 <i>Results Question 1: Impact on Local Cataloguing</i>	38
9.3 <i>Results Question 2: Partners' Feedback for WP3-leader team</i>	39
9.4 <i>Results Question 3: Long-Term Impact on Local Cataloguing</i>	42
9.5 <i>Results Question 4: Personal Feedback from Partners</i>	44
10 Conclusions – Lessons Learned	44
11 References (WP 3 Guidelines and Reports)	47
Annex I: An EFG Partner's Cataloguing plan	49
Annex II: EFG WP 3 Evaluation Questionnaire	53
Annex III: Evaluation Data Cleaning and Enrichment in Source Databases	58

Institutional abbreviations and acronyms

Short	Name
CCB	Cineteca del Comune di Bologna
CEN/TC 372	European Committee of Standardization, Technical Committee 372
CF	La Cinémathèque française
EN 15907	Metadata Standards for Cinematographic Works: Film identification - Enhancing interoperability of metadata - Element sets and structures
ISTI-CNR	Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Consiglio Nazionale delle Ricerche
CP	Cinemateca Portuguesa – Museu do Cinema, Lisbon
DFI	Det Danske Filminstitut, Copenhagen
DIF	Deutsches Filminstitut – DIF, Frankfurt
EYE	EYE Film Institute Netherlands, Amsterdam (formerly Nederlands Filmmuseum)
FAA	Filmarchiv Austria, Vienna
FIAF	Fédération Internationale des Archives du Film
FN	Filmoteka Narodowa
FUH	FernUniversität in Hagen
LUCE	Cinecittá Luce (formerly Istituto Luce)
KAVA	Kansallinen audiovisuaalinen arkisto
LCA	Lietuvos Centrinis Valstybės Archyvas -Lithuanian Central State Archive, Vilnius
LKB	Lichtspiel - Kinemathek Bern
MNFA	Magyar Nemzeti Filmarchívum – Hungarian Film Archive, Budapest
NFA	Národní filmový archiv, Prague
NNB	Nasjonalbiblioteket, Oslo
WP3-leader team	Work Package 3 leader team (consisting of DFI as WP leader and DIF as project co-ordinator)
TTE	Tainiothiki tis Ellados – Greek Film Archive, Athens

1 Executive summary

Work Package 3 had a valuable impact on the EFG network and beyond. This WP was very successful with respect to the challenges and obstacles it had to deal with. The situation in the beginning of the project was that film archives did not use common standards for cataloguing rules, vocabularies and metadata schemas. WP 3's achievements in these fields are summed up under the following bullet points:

➤ **EFG Vocabularies**

EFG vocabularies were established as keyword lists with the main purpose to allow for a uniform display of cataloguing terms in the EFG Portal as well as in Europeana. They had to be fit to express the film archive data consistently and to contextualise them in the common EFG database. The WP3 members compiled two kinds of vocabularies: a) value lists for attributes and b) value lists for semantic relationships. All in all 45 of these controlled vocabularies were established and translated into 13 languages. The vocabularies raised a lot of interest in the film archive community. Outside the EFG consortium, the FIAF Cataloguing Commission expressed a strong interest not only in the vocabularies but also especially in the translations of them. The FIAF Cataloguing Commission then re-used EFG's multilingual filmographic vocabularies for the translation of the FIAF Glossary of Filmographic Terms. The EFG vocabularies were published on the project website so that they can freely be used by other interested parties and will also be made available on the wiki filmstandards.org which is a common platform for the film archive community to discuss metadata issues.

Matching their local cataloguing terms to the English vocabularies was successful and led to a harmonised display of catalogue information in the EFG Portal. While the multilingual display could not be achieved, the matching work carried out in EFG guarantees a harmonised display of cataloguing terms in English. A remarkable result from the archives' matching work e.g. is that some of the partners introduced controlled vocabularies for database fields, which were formerly managed as free-text.

➤ **Impact of EFG on local cataloguing practices**

The questionnaire sent out by DFI on the experiences of the EFG archives with the WP3 work revealed that about 80% of the EFG archives considered the work as very relevant to their local cataloguing practices. They clearly stated that the joint approach of harmonising data in EFG had a considerable impact on the quality on their local databases. The archives considered the established general guidelines on how to clean and enrich data for EFG needs "EFG Data

enrichment and cleaning guidelines” as especially useful. The guidelines are available publicly as part of the “EFG Data Provider Handbook” in the “Outcomes” section of the EFG project website.

In order to enrich their digital collections the archives added more EFG-relevant data to existing records or catalogued newly digitized archival resources. Since collections often have only been catalogued partially in the past a full revision and further indexing of the digital objects according to EFG needs was carried out for in total around 750.000 records. Furthermore, the archives enriched authority records with EFG-relevant data or added new Person, Film Work or Corporation records to their filmographic databases. The archives also cleaned their data for EFG purposes. Most of the cleaning activities were dedicated to the cleaning of digital object records that included checking and correction of spelling mistakes. Authority data cleaning concerned the identification and merging of actual doublets in the local databases for the most part. The archives enriched and cleaned a total of 750.000 records and established around 270.000 relevant relationships from digital objects to person, film work or corporation entities.

The outcomes and benefits of WP3 work for the archives were also detectable on the level of work routine. Conventional routines were challenged, metadata got into the focus at the institutions and the collaboration and networking across the borders of the respective countries.

➤ **Bring forward standardisation**

Work carried out in WP3 was very collaborative not only between the EFG archives, but also with respect to the networking aspects of WP3 across other initiatives and the film archive domain in general. Holding an open workshop on data quality and semantic interoperability issues in Frankfurt as well as the close linkage to the CEN/TC 372 standardisation work group were part of the manifold networking activities of the WP. Through this EFG actively participated in the process of making cataloguing practices within film institutions across Europe more homogenous and sustainable.

➤ **Ground work for a common European filmography**

EFG managed to successfully lay the groundwork for a joint European filmography by bringing together and harmonising filmographic data from 16 archives. Even though EFG was not able to establish reliable authority files in the EFG database with 140.000 film work, 251.000 person and 32.000 corporation records the EFG database contains rather comprehensive filmographic information already. Partners used the EFG Authority File Manager tool to clean their person and film work data locally, which in return also improved the quality of the EFG database.

2 Introduction

This deliverable is a follow up on deliverable 3.1 “Report on type and quantity of archival resources tagged” finalized in September 2010. D3.1 is a very comprehensive report, which describes WP 3’s groundwork during the first two project years very detailed. This deliverable sums up the overall achievements and lessons learned of this WP and focuses on its activities in the last year.

Chapter 3 gives an overview of the metadata ingestion workflow in EFG with special regard to the cataloguing activities carried out within WP 3. Cataloguing in the context of EFG means all data cleaning and enrichment activities that were performed at the local level (inside the partner archives’ cataloguing and content management systems) and the EFG level (in the EFG Information Space of aggregated data). Cataloguing was partially done manually by film archive staff and was partially supported by automatic procedures either on the local or on the EFG level. Results from the partners’ local cataloguing work in the project term are summed up in chapter 4 while the forthcoming chapters concentrate on the last year’s activities to improve the contributed data in the common EFG database by means of vocabularies and matchings (chapter 5), metadata editing (chapter 6) and authority file building (chapter 7). Each of the last three chapters reports on the achievements, lessons learned and respective work’s impacts for the EFG network and beyond.

Chapter 8 sums up the results from the open workshop "Data Quality and Semantic Interoperability Issues in European Film Archives" carried out by DFI and DIF in Frankfurt. This workshop served as an important platform to bring forward the discussion on standardised cataloguing work and vocabularies in the film archive domain.

The last two chapters contain conclusions regarding the WP 3 work from the point of view of the cataloguing teams at the partner institutions (9) and the overall conclusions and lessons learned summarized by the WP3-leader team (10), consisting of DFI as WP leader and DIF as project co-ordinator.

3 Cataloguing within EFG Data Ingestion and Editing Workflow

Figure 1 displays the complete EFG data ingestion and editing workflow with special regard to the cataloguing activities carried out in WP 3. It moreover emphasises WP 3’s decisions for practical solutions applied to enrich and clean the film archive data. The different tasks were carried out either by WP 3 or 2 (“Technical Interoperability and Access”). The overall process was co-ordinated by DIF. The following pages introduce each step briefly while the tasks co-ordinated by the WP3-leader team (framed in red in the figure below) are described in more detail.

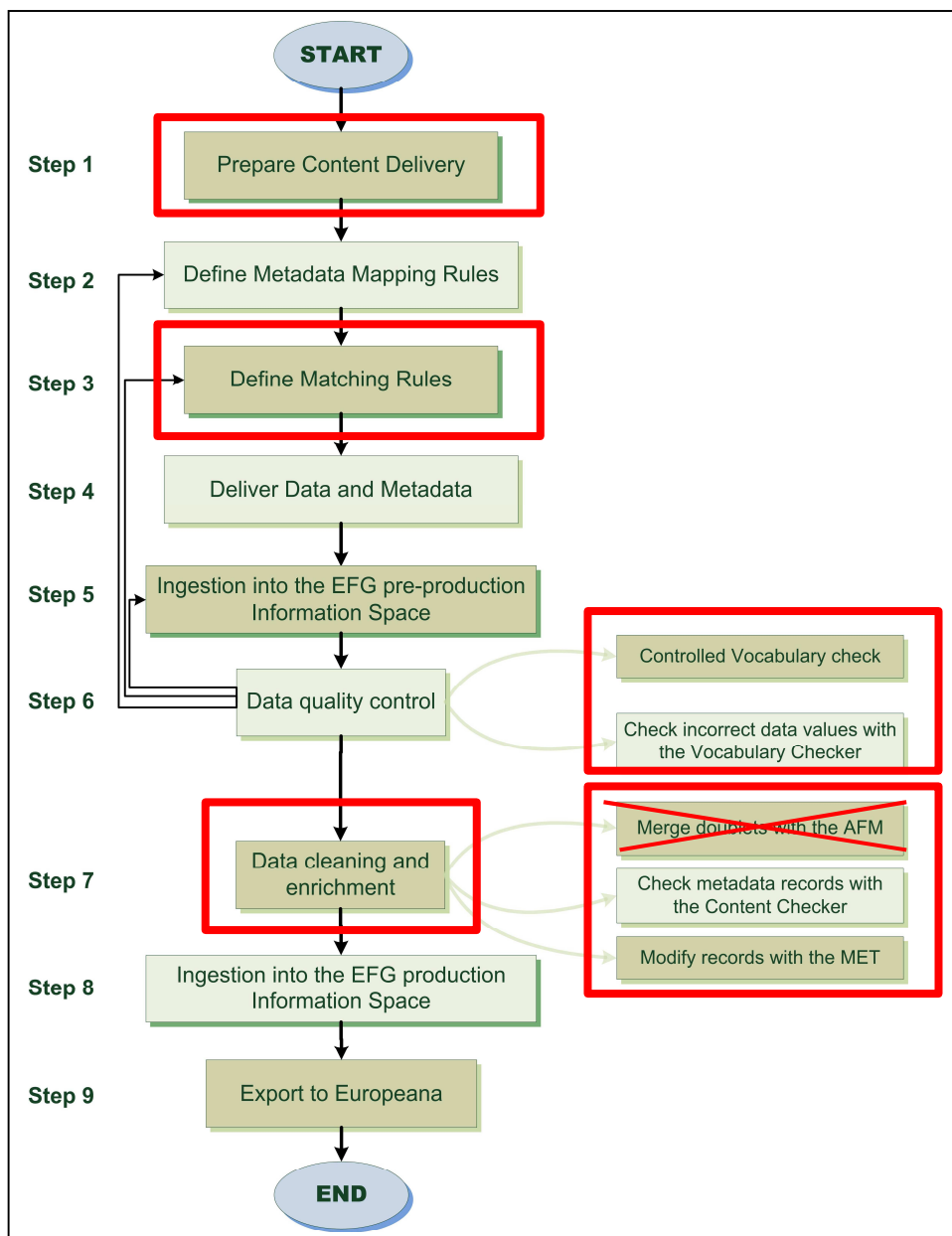


Figure 1: EFG Cataloguing Activities within Data Ingestion Workflow

Step 1: Prepare Content Delivery

Before the partner archives delivered their data contributions to EFG, the WP3-leader team guided them to enrich and clean their metadata records for EFG needs in their local database(s). Cataloguing on the local level was necessary predominantly because many of the catalogues used were originally not intended to supply data to other portals than the local internal catalogue or the individual website, such as filmportal.de or the Danish Filmography. Furthermore, a survey among the partner archives carried out by the WP3-leader team in April 2009 revealed the consequences of the fact that they do not apply common cataloguing rules. The results of this survey, which focused especially on how the archives index names and titles, were reported in Milestone 3.3 “Best Practices for Filmographic Editing and Authority File Administration” [EFGM332009]. The presentations confirmed what was known before but revealed to what extent the following observations are true:

- Archives are inspired by FIAF Cataloguing Rules and Anglo-American Cataloguing Rules but have developed local adjustments.
- Archives have different ways to distinguish between two or several persons with the same name.
- Different technical solutions dictate differing approaches to cataloguing names and titles.
- In general there are no consistent cataloguing rules applied across the film archive domain
- In general there are no consistent vocabularies applied across the film archive domain

As a result of the survey and discussions on the “WP 3 Workshop on Cataloguing Rules and Vocabularies” held in Copenhagen in May 2009, the WP 3 members decided not to apply common cataloguing rules to enrich and harmonize their data for EFG purposes. The amount of work that would have been necessary to standardize institutional cataloguing practices and processes was considered as disproportionate in relation to EFG’s overall cataloguing aims.

For this reason, the WP3-leader team established general guidelines on how the partner archives should clean and enrich data for the specific needs of the EFG Portal. These “EFG Data Enrichment and Cleaning Guidelines” name priorities for the archives' cataloguing of digital resources and filmographic information (meaning: film-relevant information about person, film works, corporate entities) for EFG. Most of the partners considered the EFG cataloging priorities when they enriched and cleaned their data contributions. However, it was not always possible for them to follow all recommendations since, like mentioned above, the EFG guidelines were not totally in line with the local cataloguing priorities. The “EFG Data Enrichment and Cleaning Guidelines” were integrated into the “EFG Data Provider Handbook” which is publicly available in the “Guidelines & Standards” section of the project website [REF EFGDPH2011, see chapter 4: “Preparing data for EFG”]. It is worth mentioning that they contain relevant information for other film institutions outside the EFG consortium as well since they recommend the use of previously

established cataloguing standards whenever possible. It is optional for film institutions joining EFG after the end of the project to catalogue their data contributions according to these guidelines.

As a next step the WP3-leader team established individual work plans which listed the recommended EFG cataloguing activities for each partner in the third year of the project. For a cataloguing plan example please refer to *Annex I: An EFG Partner's Cataloguing plan*. Based on these plans, the WP3-leader team agreed individually with each partner on the priorities for the outstanding cataloguing activities until the end of the project. Through this procedure the archives managed to accomplish all their EFG-related cataloguing work in time until August 2011.

Step 2: Defining Metadata Mapping Rules

In this step metadata-mapping rules had to be defined and the local database fields had to be mapped to the elements of the common EFG Metadata Schema [REFEFG2009]. For this purpose WP 2 developed a mapping form, which all 16 archival partners completed. Deutsches Filminstitut – DIF together with each archive providing data, established the mapping rules¹. Because the film archives cannot use standardized formats to export their metadata for EFG, mapping rules needed to be established for each different kind of data export, which amounts to a total of 64 mappings in August 2011.

Step 3: Defining Matching Rules

Step three concerns vocabulary matching activities and matching rules. For the purpose of aggregating data in EFG and, subsequently Europeana, multilingual entries in several database fields that had been used and filled with data in an unsystematic manner at the local level, needed to be streamlined, or “cleaned” so that they could fit into the EFG vocabularies used for the matching of filmographic terms. Therefore, matching tables were established in which the local values were assigned to the terms of the EFG vocabularies which were defined for certain elements and relationship types of the EFG Metadata Schema [REFEFGS2009]. Under guidance of the WP3-leader team each partner archive established one matching table for its data. For more information about this EFG data cleaning activity please refer to chapter 5 *EFG's Vocabulary and Matching Work*.

¹ For more information on the mapping process in EFG, please refer to Deliverable 2.4 „Report on inclusion of archives' repositories“

Step 4: Deliver Data and Metadata

As a next step, the providers delivered their local XML-exports to the EFG pre-production Information Space. The pre-production Information Space is a part of the technical EFG infrastructure that contains the actual EFG database. Six archives developed an OAI-PMH interface on top of their databases, through which data for EFG was "harvested" on a regular basis. Other archives submitted their data contributions via XML exports. This delivery process was managed between by Deutsches Filminstitut – DIF.

Step 5: Ingestion into the Pre-production Information Space

In step five, WP 2 leader ISTI-CNR ingested the XML-exports into the EFG Pre-production Information Space. This meant that the data from the local database exports was converted to the EFG metadata schema according to mapping rules defined in step 3. Once the ingestion had been completed, ISTI-CNR reported that the data was available in the EFG Pre-production Information Space. The new data, which was not available to end users through the EFG Portal at that stage, could be viewed through a tool specifically developed for EFG: the EFG Content Checker Tool. With the help of this tool, DIF and the partner archive checked if the data was properly ingested into the EFG pre-production Information Space. If a revision of the mapping rules was necessary the procedure was repeated starting from step 2.

Step 6: Data Quality Control

After the XML-exports were ingested into EFG the source values were automatically converted to the EFG vocabulary terms according to the matching tables the partners delivered in step 3. During the data quality control phase all incorrect metadata records, where a transformation from the local value to the EFG vocabulary term was not performed, were identified. This happened whenever a value was not included or not included correctly in the matching table. In order for the partner archives to be able to verify for which local values the vocabulary matchings were not established correctly, ISTI-CNR designed a special web tool: The Vocabulary Checker. This tool displayed all metadata records containing incorrect metadata values to the archival users. If incorrect data values were discovered, the matching tables were revised accordingly. DFI and DIF supported the archival partners in verifying and correcting their matching tables according to the results of the Vocabulary Checker tool. For the revision of the Matching Rules the procedure is repeated starting from step 3.

Step 7: Data Cleaning and Enrichment

Step seven, Data Cleaning and Enrichment, starts after the ingestion and correction of matching tables are finalized. The WP3-leader team informed the partner archives when further cataloguing was needed according to the individualized cataloguing plans. The archives used the EFG Metadata Editor tool to correct the metadata values and to enrich their metadata manually. The additional cataloguing activities with the Metadata Editor was carried out on the EFG level. The Authority File Manager Tool was used to identify possible duplicates of persons and film works within an archives' data contribution. Originally, it was planned that the archives can merge all duplicates directly within the EFG Information Space of aggregated metadata. This means that doublets occurring between different institutions should have been merged by the cataloguers on the EFG level. However, this task turned out to be too time consuming for the present project, so that it still remains to be accomplished in the future. In order for the archival cataloguers to clean their authority data locally, they exported lists of possible duplicates within their data - after detection with the aid of the tool - into an excel file, and then deleted the actual duplicates in their local database(s). In this case, a new ingestion of the locally cleaned data was necessary and the procedure was repeated starting from step 4. Please refer to chapter 7 for more information about EFG's approach to authority file building and chapter 6 for the cataloguing work with the Metadata Editor tool.

Step 8: Ingestion into EFG Production Information Space

At this step the data was ingested into the EFG Production Information Space which meant that it was rendered visible in the EFG Portal. This work was performed by ISTI-CNR after the provider had given permission for the publication of the data.

Step 9: Export to Europeana

The final step was the data export to Europeana in the ESE v3.4 format. This step was managed between DIF and the Europeana Ingestion team. The Europeana OAI-PMH delivery facilities were maintained by ISTI-CNR. The Europeana export filter runs stable so that little intervention by ISTI-staff is considered necessary for further possible Europeana exports in future.

4 Data Cleaning and Enrichment Activities in the Source Databases

In the third project year the WP3-leader team continued to monitor which kind of data cleaning and enrichment work the EFG archives carried out in their local cataloguing systems. For more information on how the WP3-leader team guided the archives through the local cataloguing process please refer to the first step “Prepare Content Delivery” of the EFG ingestion workflow (see also figure 1). In five project periods the archives were asked to report on their cataloguing activities by means of an evaluation questionnaire which the WP3-leader team established for this purpose. *Annex III* provides an overview of all data cleaning and enrichment activities performed by the partners from 1-SEP-2008 until 31-AUG-2011. Since the local activities are already described in detail in deliverable 3.1 [EFGD312010], this chapter provides only a summary of the overall achievements.

4.1 Evaluation Questionnaire

The WP3-leader team adapted minor changes in the questionnaire according to feedback from EFG partner archives and Europeana staff, based on the results reported in D3.1 [EFGD312010]. The changes in relation to the initial questionnaire were:

1) The initial question 1 “Local cataloguing work on digital collections” was split into two questions in order to get more precise indications on the amount of enriched digital objects (Question 1 “Data Enrichment Digital Objects”) and cleaned digital objects (Question 2 “Data Cleaning Digital Objects”).

2) Furthermore, it was considered relevant to get more information about the methods applied by the archives to clean or enrich their data. In order to evaluate whether the respective number refers to a cataloguer’s activity or to an automatic procedure the archives were asked to choose one of the following options for questions 1 – 4:

- *Hand-edited*: This was indicated when an archival cataloguer carried out the respective enrichment or cleaning activity manually by editing each record separately. For example, the Lietuvos Centrinis Valstybės Archyvas (LCVA) further enriched 130 film object records for EFG needs.
- *Automatically*: Partners chose this option when the respective cleaning or enrichment activity was performed completely automatically in the local cataloguing or content management system(s). For example, gender of film-related persons was identified and indexed automatically at DFI.
- *Both*: This option applies when the respective manual cataloguing activity was supported by an automatic procedure. For instance, partners used the EFG Authority File Manager Tool

to identify duplicates within their own authority data contributions (automatically) and cleaned them accordingly in their local cataloguing systems (hand-edited). Partners which used the Authority File Manager for the local doublet cleaning are Deutsches Filminstitut, National Norwegian Library and Det Danske Filminstitut. Other partners introduced semi-automatic procedures locally to clean or enrich their data. For instance, DFI defined a semi-automatic process to harmonize heterogeneous entries in the field for Person's type of activity, and to build up a controlled vocabulary.

4.2 Achievements: Local Cataloguing Work

Table 1 provides a summary of the numbers provided in *Annex III* regarding the archives' local data enrichment and cleaning activities. 14 out of 16 partner archives replied to the evaluation questionnaire. Two institutions did not reply to the questionnaire because these institutions did not receive funding for cataloguing in EFG and therefore for them local cataloguing was not mandatory.

In the course of the project all 14 archives enriched and cleaned in total around 750.000 records – either manually, semi-automatically or fully automatically. More than half of the activities were enrichments (58,9 %). In order to enrich their digital collections (24,5 %) the archives added more EFG-relevant data to existing records, or catalogued newly digitized archival resources. Since collections often have only been catalogued partially in the past, a full revision and further indexing of the digital objects according to EFG needs was carried out. Furthermore, the archives enriched authority records with EFG-relevant data or added new person, film work or corporate entity records to their filmographic databases (34,4 %).

The archives also cleaned their data for EFG purposes (41,1 %). Most of the cleaning activities were dedicated to the cleaning of digital object records (34,3 %) which included checking and correction of spelling mistakes. Authority data cleaning (6,8 %) mainly concerned the identification and merging of actual doublets in the local databases. Three archives used the EFG Authority File Manager Tool (further described in chapter 7) to support this.

Note on the numbers reported in this chapter: It is possible that there there are some overlaps between the indications provided by the archives regarding enriched and cleaned data. The WP3 leader team undertook best efforts to detect these overlaps by asking the archives to specify their indications. However, it cannot be ruled out that there are still some remaining overlaps, meaning some enriched records could have been listed under cleaned records and vice versa. But this is insignificant in relation to the overall activity.

Activity	Number	Percent
1. Total Digital Object Enrichment	183.489	24,5 %
2. Total Digital Object Cleaning	257.492	34,3 %
3. Total Authority Record Enrichment	258.228	34,4 %
4. Total Authority Record Cleaning	50.695	6,8 %
Total amount of enriched and cleaned records:	749.904	100 %

Table 1: Total Numbers Cataloguing Activities of EFG Partners

The establishment of EFG-relevant relationships between object and authority records was an important enrichment activity performed by the archives. In the EFG Portal, digital objects are embedded into a context which means that they are displayed together with their film titles and/or corporate entities and person names. The connection between digital objects and film works, persons or corporate bodies also ensures that objects can be found via titles or names through the EFG search. A concrete example of this kind of work is the cataloguing of persons depicted on film stills. In total, the archives enriched around 269.000 object records with relationships to persons, film works or corporate entities (97,7 %) or they added a film title, person name or corporate entity name into the digital object record (1,7 %). In order not to report the same kind of enrichment activity twice, the number of back-links the archives established from the respective authority record to the object record was excluded from table 2 but can be found in *Annex III*. With 208.697 newly established references linking digital object records to person and film work records, WP 3 has exceeded the originally envisaged 200.000 digital items that should be enriched until the end of the project. Thus, EFG reached its success indicator “Number of Items enriched” listed in the Description of Work.

Activity	Number	Percent
1. Total Object Records related to Authority Records (Enrichment)	262.641	97,7 %
2. Total Names and Film Titles Inserted into Object Records (Enrichment)	6.296	2,3 %
Total amount of object records enriched with names and titles:	268.937	100 %

Table 2: Total Numbers Establishment of Relations / Inserted Names and Titles

As shown in table 3, partners indicated the respective enrichment or cleaning method (hand-edited, automatically, both) for around 724.000 records (see Annex III, questions 1 – 4). Most of these records were enriched or cleaned semi-automatically (37,2%). An example for semi-automatic enrichment is the use of the EFG Authority File Manager tool to clean doublets of Persons locally. 37,2 percent of the records were enriched or cleaned manually by a cataloguer while around 23,1 percent were processed completely automatically. This leads to the conclusion that the partners introduced a semi- or fully-automated procedure whenever possible in order to clean or enrich their data in the most efficient way for EFG.

Hand-edited	Number	Percent
1. Digital Object Enrichment	107.484	
2. Digital Object Cleaning	879	
3. Authority File Enrichment	128.883	
4. Authority File Cleaning	32.211	
Hand-edited total:	269.457	37,2 %
Automated		
1. Digital Object Enrichment	51.800	
2. Digital Object Cleaning	0	
3. Authority File Enrichment	115.580	
4. Authority File Cleaning	0	
Automated total:	167.380	23,1 %
Both		
1. Digital Object Enrichment	14.285	
2. Digital Object Cleaning	256.460	
3. Authority File Enrichment	966	
4. Authority File Cleaning	15784	
Both total:	287.495	39,7 %
Total amount of records with indicated enrichment or cleaning method	724.332	100 %

Table 3: Total Numbers Cleaning and Enrichment Methods

For more precise information about each partner's cleaning or enrichment methods, please refer to *Annex III*.

5 EFG's Vocabulary and Matching Work

This chapter describes EFG WP 3's activities to establish controlled vocabularies for the EFG database and to match the source values from the partner archives to a common set of terms. This data cleaning work refers to steps 3 and 6 of the EFG ingestion and metadata editing workflow (see figure 1). Overall objective of the vocabulary matching work in EFG was to harmonize the heterogeneous and multilingual source values aggregated in the common EFG database for a coherent display in the EFG and Europeana web portals.

5.1 EFG Controlled Vocabularies: Scope

EFG vocabularies are keyword lists which were established for end users of the EFG and Europeana portals, not for professional needs of film archive cataloguers and filmographers. In order to express the film archive data consistently and to contextualise them in the common EFG database, which is based on an entity-relationship model with subject-predicate-object triples, the EFG WP3 members compiled two kinds of vocabularies:

- Value lists for attributes
- Value lists for semantic relationships

The EFG vocabulary lists are publicly available in the “Guidelines and Standards” section of the project website and were translated by the EFG partner archives into 13 European languages: www.europeanfilmgateway.eu/guidelines_and_standards.php.

The EFG vocabularies focus on formal aspects to describe film archive material and filmographic data, such as persons and film works. Harmonisation of subject terms was not considered in EFG due to fact that most film archives do not index their resources at such depth. Also, if archives index subjects they do not use standardized rules or vocabularies. Thus, subject indexing in EFG would be an enrichment for which a thematic concept and a common indexing policy would be necessary. The development of such concept was not in scope of the EFG project. The list below illustrates what aspects the EFG vocabularies cover:

- Formal aspects to describe film works and film material (e.g. form/category, format, country, region, language)
- Formal aspects to describe non-film material (e.g. document type, format, country, region, language)
- Types for different attributes of data elements (date types, name types, title types, activity types)
- Types for semantic relationships (cast & credits)

Whenever possible, existing vocabularies and standards were used to compile the EFG vocabularies. The bullet points hereunder list the vocabulary sources and standards which were used for the vocabulary and matching work in EFG:

- FIAF Glossary of Filmographic Terms, 2008 (for film-related activities and other film-specific terms)
- ISO Country Codes 3166.1 (for countries currently in existence)
- AFNOR codes (for historical countries)
- Marc Geographic Area Codes (for large geographic regions)
- ISO Language Codes 639.1, 639.2
- Marc Relator Codes (for non-film related activities)

- EBU P/META 2.0 Concept Schemes (for language usage types)
- EAC Beta (for name and date types)
- IANA MIME Media Types (for media types)
- DCMI Metadata Terms (digital object types)
- Value lists of EFG partner archives (for document types et. al.)

5.2 Value Lists

The EFG value lists were managed by the WP-3 leader team in excel files and updates were regularly exchanged with ISTI-CNR who ingested them into the EFG Information Space. In total, WP 3 established:

- **38 value lists with 1.700 terms (preferred terms)**

The figure hereunder illustrates an example how the term “Documentary” is stored in the value list “Form”. The column “EFG Term” contains the preferred term to which the source values from the archives are matched or harmonized. The column “Display Term” lists the preferred term in all 13 languages specified by a language attribute (column “Language”). The terms were translated by the partner archives to allow that the harmonized values can be displayed in several languages in the EFG Portal. However, implementing a multilingual metadata display was considered as too work intensive by WPs 2 and 4. Since other high priority tasks needed to be tackled first, it was decided not to realise the multilingual metadata display in the project. Apart from language equivalences, the value lists also allow to distinguish genders: female (e.g. “Actress”) or male (e.g. “Actor”) or unknown (e.g. “Documentary”).

EFG Term	Vocabulary Name	Display Term	Language	Gender
Documentary	Form	Documentary	en	U
Documentary	Form	Dokumentar	da	U
Documentary	Form	Documentaire	fr	U
Documentary	Form	Dokumentarfilm	de	U
Documentary	Form	Dokumentti	fi	U
Documentary	Form	Ντοκιμαντέρ	gr	U
Documentary	Form	Dokumentarfilm	no	U
Documentary	Form	Dokumentární film	cz	U
Documentary	Form	Documentário	pt	U
Documentary	Form	Dokumentinis filmas	lt	U
Documentary	Form	Documentario	it	U
Documentary	Form	Documentaire	nl	U
Documentary	Form	Dokumentumfilm	hu	U

Figure 2: EFG Term “Documentary” with language equivalences in value list “Form”

5.4 Semantic Relationship Vocabulary

In order to contextualise the collected data in the common EFG database and hence in the portal, WP 3 established:

- **7 lists with 210 semantic relationship terms (preferred terms)**

Figure 3 displays how the source values are harmonized to the EFG term “Director”, according to the defined semantic relationship terms. These are EFG Terms that describe the semantic relation between two entities, for instance between a Person and a Film Work (in EFG Schema called “AVCreation”). In this example a Film work (Domain) is related to a Person (Range) through the semantic relationship “Director” (EFG Term). The semantic relationship vocabularies contain also equivalences of semantic relationship terms in different languages and genders.

EFG Term	Vocabulary Name	Domain	Display Term	Display Term Inverse	Range	Language	Gender
Director	AgentRelation	AVCreation	Regisseur	Regisseur	Person	nl	U
Director	AgentRelation	AVCreation	Režisér	Režisér	Person	cz	M
Director	AgentRelation	AVCreation	Režisérka	Režisérka	Person	cz	F
Director	AgentRelation	AVCreation	Rendezí	Rendezí	Person	hu	U
Director	AgentRelation	AVCreation	Regissor	Regissor	Person	no	U
Director	AgentRelation	AVCreation	Instruktør	Instruktør	Person	da	U
Director	AgentRelation	AVCreation	Ohjaaja	Ohjaaja	Person	fi	U
Director	AgentRelation	AVCreation	Realizador	Realizador	Person	pt	M
Director	AgentRelation	AVCreation	Realizadora	Realizadora	Person	pt	F
Director	AgentRelation	AVCreation	Režisierius	Režisierius	Person	lt	M
Director	AgentRelation	AVCreation	Režisier	Režisier	Person	lt	F
Director	AgentRelation	AVCreation	Réalisateur	Réalisateur	Person	fr	M
Director	AgentRelation	AVCreation	Réalisatrice	Réalisatrice	Person	fr	F
Director	AgentRelation	AVCreation	Regisseur	Regisseur	Person	de	M
Director	AgentRelation	AVCreation	Regisseurin	Regisseurin	Person	de	F
Director	AgentRelation	AVCreation	Σκηνοθέτης	Σκηνοθέτης	Person	gr	U

Figure 3: Semantic Relationship Vocabulary, Example “Director”

The figure below illustrates how EFG semantic relationships and controlled terms help to contextualise the filmographic data and digital archival objects in the common EFG database. The red frames highlight the relationship between an image and the person that is depicted on it. This was one of the most important enrichment activities the film archives performed as part of their cataloguing work for EFG (for further information please refer to chapter 6.2 *A Cataloguing Example*). By implementing these subject – predicate – object triples in an entity relationship

model, EFG Data can be exploited for RDF-implementations and are semantically interoperable for other uses.

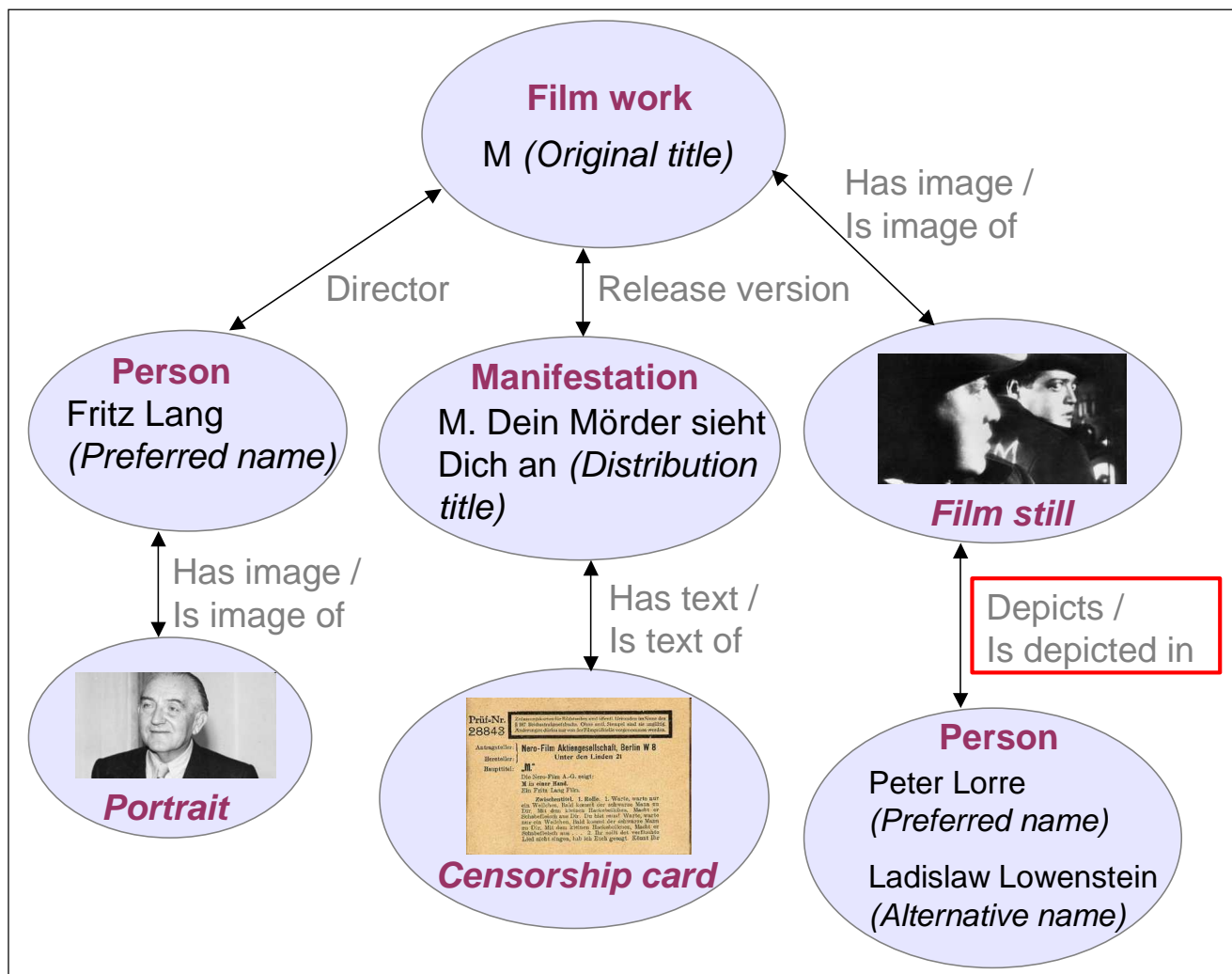


Figure 4: Contextualisation of Film archive Data in EFG through Semantic Relationships

5.3 Vocabulary Matching Experience

In order to harmonise the source values to the EFG vocabulary terms the WP3-leader team guided the partner archives to complete a matching table for their data. In total the partners established:

- **16 tables with 11.000 matched source values**

Figure 5 below illustrates an example on how the values from the archive EYE were harmonized to the controlled terms of the EFG vocabulary list “Colour”. The Dutch value is indicated in the column

“Source Value”; the controlled term to which this respective source value was matched is listed in the column “EFG Term”.

Data Provider	Source Value	Language	EFG Term	Vocabulary Name
EYE	Zwart-wit	nl	Black & White	Colour
EYE	Kleur	nl	Colour	Colour
EYE	Onbekend	nl	n/a	Colour
EYE	Tinting	nl	Tinted / Toned / Hand coloured	Colour
EYE	Toning	nl	Tinted / Toned / Hand coloured	Colour
EYE	Inkleuring	nl	Tinted / Toned / Hand coloured	Colour

Figure 5: An EFG Partner’s Matching Table for “Colour” Terms

A major challenge was that 60% of the source values were contributed as free-text strings to EFG. This increased the number of source values and hence the necessary matching work, especially when updates of contributions were ingested into the common EFG database. As no local vocabularies or specific rules were applied for the respective free-text fields, further terms were catalogued locally for the same concept in the meantime. For example, DFI submitted the following uncontrolled source values which all refer to the same concept and were matched to the single EFG term “Assistant camera operator”:

- *Fotografassistent*
- *2. Ass. Cameraman*
- *2. foto.ass*
- *3. Ass. Cameraman*
- *Assistant camera*
- *Assistant Cinematographer*
- *Assisterende fotograf*
- *B.foto*
- *Camera assistant*
- *C.foto, Danmark*
- *First assistant camera*
- *Foto praktikant*
- *foto.ass*
- *Foto.ass i Nicaragua*
- *Fotograf . motorcykel*
- *Fotograf 2. assistent*

- *Fotografass.*
- *Fotografisk assistance*
- *Kameraass.*
- *Kameraassistent*
- *Multicam technician*
- *Second assistant camera*
- *Suppl. kamera*
- *B.foto, 2. unit*
- *B.Fotograf*
- *Kamera, 2.unit*

Only 40% of the contributed terms were defined by local vocabularies which needed to be matched only one time to the EFG vocabulary. After the source values were converted to the EFG vocabulary terms in the EFG database, the archives verified their matching tables with the EFG Vocabulary Checker tool (see figure 1, step 6). The figure hereunder displays an EFG XML record in the Vocabulary Checker tool which contains an invalid value that was not included or not included correctly in the partner’s matching table:



EFG european film gateway

EFG Rotten Checker
Welcome franca.debole | EFG Tools | Logout

Last Update: 2011-01-24T13:10:49+01:00

Record XML

Invalid fields are **highlighted**

```

<efg:efgEntity>
  <efg:award>
    <efg:identifier scheme="CP_CATEGORY_ID"> DIF_award_0019FB69CE854B46BCC7D2764E442FDA </efg:identifier>
    <efg:recordSource>
      <efg:sourceID> 0019FB69CE854B46BCC7D2764E442FDA </efg:sourceID>
      <efg:provider schemeID="Institution acronym" id="DIF">Deutsches Filminstitut - DIF </efg:provider>
    </efg:recordSource>
    <efg:name> Prädikat: wertvoll </efg:name>
    <efg:sponsor> FBW </efg:sponsor>
    <efg:date> 2004-08-01 </efg:date>
    <efg:relAvManifestation>
      <efg:identifier scheme="CP_CATEGORY_ID"> DIF_avCreation_2BAB82FFFFF4466DBD3FB5178DFA1D5D </efg:identifier>
      <efg:title> Greenhorn </efg:title>
      <efg:type> film </efg:type>
    </efg:relAvManifestation>
  </efg:award>
</efg:efgEntity>
    
```

This value is invalid

Allowed values for AVManifestationRelation vocabulary are:

- Alternative version
- Archive version
- Censored version
- n/a
- Original version
- Other version
- Partial version
- Release version
- Short version
- TV adaptation

[Back]

Powered by D-NET

Co-funded by the Community programme eContentplus

EFG is solely responsible for the content of this site, which does not represent the opinion of the Community. The Community is not responsible for any use that may be made of the information contained on this site.

Figure 6: Verification of an EFG Record with Vocabulary Checker Tool

Figures 7 and 8 below illustrate the results of the vocabulary work in the EFG Portal. The homogenized metadata are used for the faceted search which helps users to get faster access to the material they are searching for (filters: Provider, Language, Media):

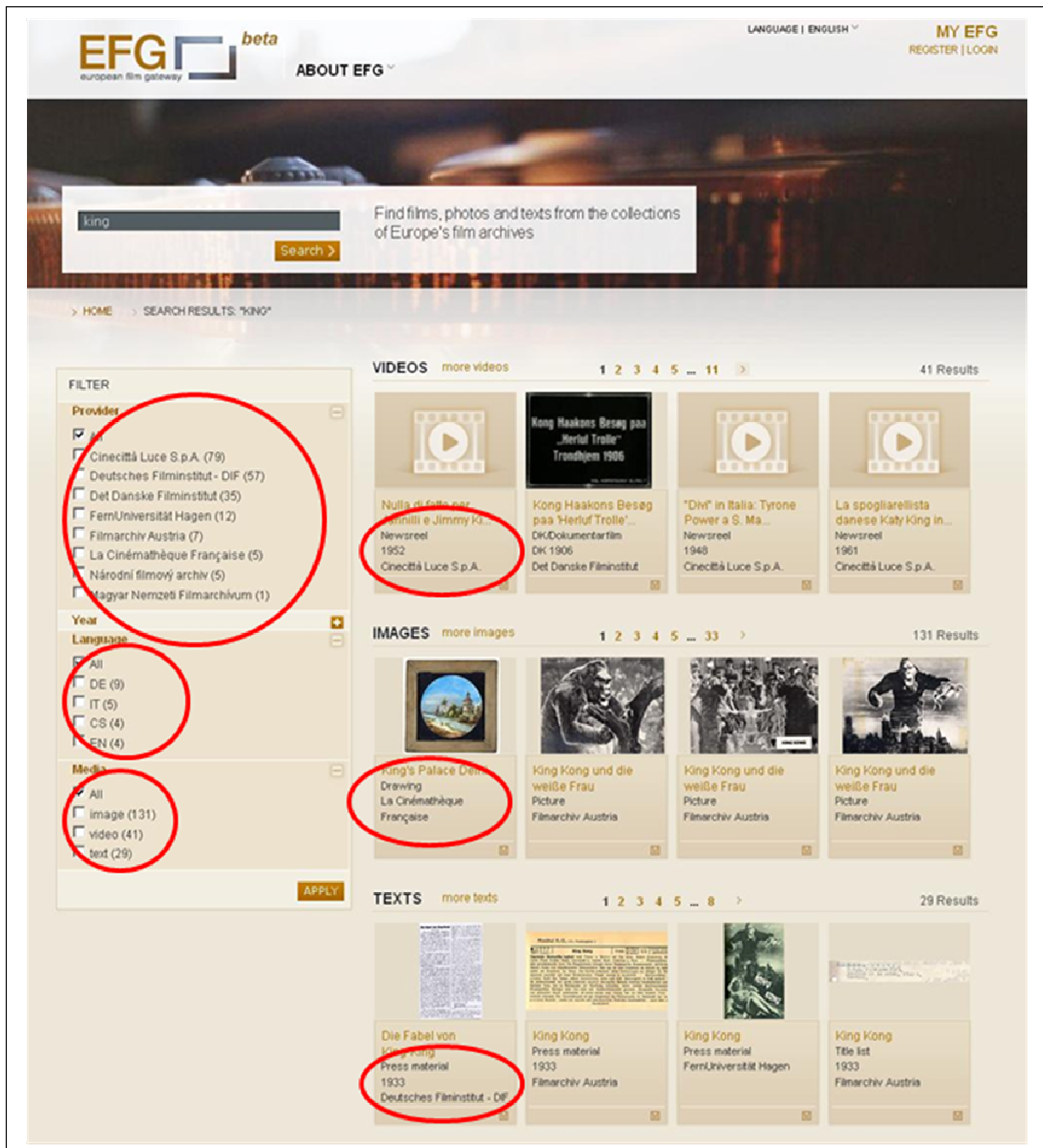


Figure 7: Controlled Vocabularies in EFG Portal's Result Page (Facetted Search, Thumbnail Display)

In the detailed page of a search result, respective metadata are expressed consistently in English language so that end users can understand the information context of the retrieved material:

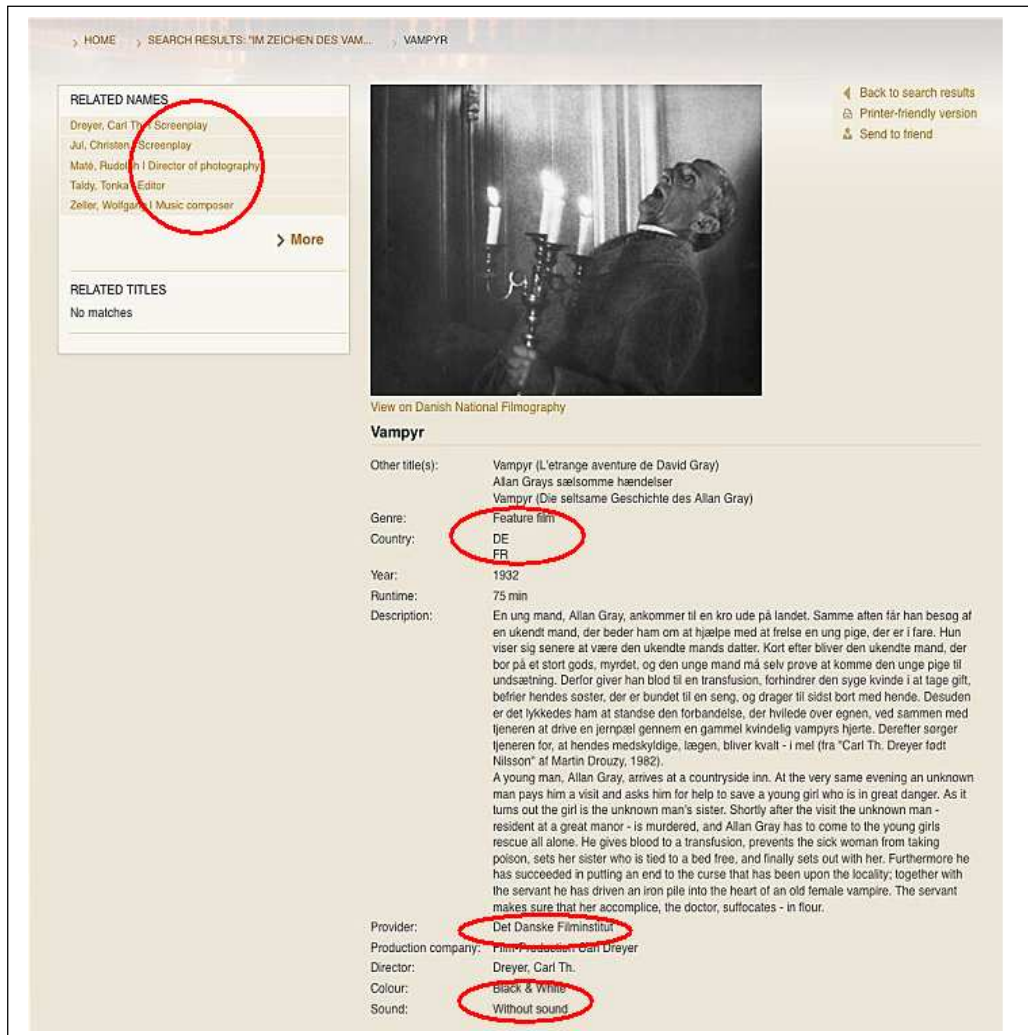


Figure 8: Controlled Vocabularies in the Detail View of a Result

5.5 Achievements and Lessons Learned

In the context of the EFG project, WP 3 has accomplished the following achievements regarding vocabulary matching work:

- A consistent display of film-relevant metadata in the EFG and Europeana web portals. Examples for EFG Portal are illustrated in figures 7 and 8 on the previous pages.
- Harmonisation of heterogeneous and multilingual source values in a common database of digital film-archival material and filmographic information. Special focus was placed on the contextualisation of film archive data through semantic relationships (e.g. relation between person and film work by shared cast & credits vocabulary).
- Groundwork for a common European registry of persons, film works and corporate entities. Through harmonized EFG XML records, filmographic metadata can be compared

automatically because the same things are related to the same term concepts (for instance: the activity “Actor” can only be identified as equal when the other record also contains the reference to the concept “Actor” and not to “Akteur” (fr) or “Darsteller” (ge)). This procedure is necessary to establish authority files efficiently. More information about this topic can be found in chapter 7.

Not only the EFG project, also the partner archives benefited from the vocabulary matching experience. A remarkable result from the archives’ matching work is that some of them introduced controlled vocabularies for database fields which were formerly managed as free-text. For instance, DFI and DIF are now using an according vocabulary for persons’ functions and activities. These are only two examples, which demonstrate what impact the EFG project had on the partners’ local cataloguing practises.

The establishment of the EFG multilingual vocabularies has generated great interest by other parties from the film archive domain beyond the EFG consortium. The vocabularies, which are available in 13 European languages, were published on the project website so that they can freely be used by other interested parties. In the third project year, the EFG team continued aligning its WP 3 vocabulary work with those of film archive standardisation initiatives. The FIAF Cataloguing Commission re-uses EFG’s multilingual filmographic vocabularies for the translation of the FIAF Glossary of Filmographic Terms².

Furthermore, EFG WP 3 co-operated with the European Committee of Standardization (CEN/TC 372). This initiative developed a metadata standard for cinematographic works, EN 15907 which was designed as an interoperability specification addressing the multitude of databases that exist in audiovisual heritage institutions throughout Europe³. Results of EFG’s vocabulary and data harmonization work were presented on the CEN TC 372 Workshop on 13 – 14 April in Copenhagen which was dedicated to disseminate the new standard⁴. This workshop was the second in a series of four dissemination events sponsored by the Enterprise and Industry Directorate of the European Commission. Other workshops took place in Rome (October 2010), Prague (May 2011) and Paris (June 2011). In the near future, the vocabularies also will be provided on the wiki filmstandards.org in order to disseminate them broadly within the film archive

²Originally released as: *Fédération Internationale des Archives du Film*, Jon Gartenberg (ed.), *Glossary of Filmographic Terms*, Munich 1989. Updated English version by Zoran Sinobad available at <http://www.fiafnet.org/publications/Glossary%20of%20Filmographic%20Terms%20%28English%20Version%20292008%20revision.pdf>

³ URL to the standard: http://filmstandards.org/fsc/index.php/EN_15907

⁴ More information on filmstandards.org:
http://filmstandards.org/fsc/index.php/CEN_TC_372_Workshop_Series#Copenhagen.2C_13-15_April_2011.
EFG’s presentation is publicly available on its project web site:
http://www.efgproject.eu/downloads/CEN_Copenhagen_0110511.pdf

community and to connect EFG's work with the FIAF and CEN standardisation initiatives. Through this wiki, filmographers and cataloguers can verify in how far the EFG vocabularies can be used in the context of specific film archive indexing requirements.

During the EFG project it turned out that the vocabulary matching process is very resource-intensive in terms of managing, updating and maintaining the vocabulary files. For the future data cleaning work, EFG recommends to apply a vocabulary management tool instead of excel files. The WP3 leader team was in contact with other initiatives which are using such tools to clean their data (for instance: MIMO – Music Instruments Museum Online, digiCULT – Digital Culture). A vocabulary tool would offer possibilities to establish uniquely addressable concepts for term control, to link EFG vocabularies with existing external vocabularies and to properly manage synonyms, homonyms and scope notes. However, implementing an according tool into EFG's data cleaning workflow was not accomplishable during the time of the project.

Furthermore, the EFG vocabularies should be properly exploited and expanded for the portal in order to satisfy end user needs. They should be available in all languages and not only in English. The vocabularies could enhance the information retrieval by using them for advanced search and browsing options. For instance, user tests carried out by WP 1 ("User needs and system requirements") showed that users wish to browse for media-specific categories (for instance: "Newsreels", "Posters", etc.) which could potentially be supported by the established vocabularies.

6 Cataloguing on EFG level by using the Metadata Editor

After the contributions were ingested into the EFG database, the archival users can now perform several cataloguing activities before their metadata are displayed through the EFG Portal (see figure 1, step 7).

6.1 Achievements: Cataloguing with the Metadata Editor

The Metadata Editor is a cataloguing tool, developed by ISTI-CNR, which allows to add, edit and delete records, as well as to establish relationships between authority records and digital objects directly in the EFG database. Several WP 3 workshops were held between the EFG partner archives, DFI and ISTI-CNR in order to finalise the tool according to the requirements of the

cataloguers⁵. Its functionalities are described in depth in Deliverables 3.1 “Report on type and quantity of archival resources tagged” and 4.6 “Report on EFG service operation and promotion activities”, and therefore not repeated here.

During the past three years, the WP3-leader team encouraged the partner archives to enrich as much of their metadata as possible in their local cataloguing systems. Through this procedure the film archives could benefit most from their cataloguing work for EFG. Only if local cataloguing was not possible, a partner archive performed this work directly in the EFG database with the help of the Metadata Editor Tool. Cataloguing work with the Metadata Editor started after the finalisation of the tool in May 2011. At this time of the project, most of the partner archives had already accomplished their cataloguing work for the EFG Portal locally. Thus, the tool was primarily used to perform final adjustments to improve the quality of the already ingested metadata records. The bullet points hereunder list the metadata editing activities that were carried out by the partner archives:

- Four partners used the MET to add or correct links to their digital objects (i.e. IsShownBy link) or to previews (i.e. Thumbnail link). For instance, the Norwegian Library corrected the thumbnail links in the EFG records for its complete contribution of 204 films.
- Other partners deleted digital object records that should not have been published through the EFG Portal. For instance some partners needed to delete digital object records due to copyright reasons (for instance in the case of Národní filmový archiv), or records needed to be deleted that were delivered by mistake by the provider (for instance in the case of Magyar Nemzeti Filmarchívum).
- Establishment of relationships between authority data (persons, corporate bodies, film works) and digital object records was the third MET enrichment activity. An example how Det Danske Filminstitut performed this task with the Metadata Editor is given below in this chapter.

Five partners modified in total around 1.000 metadata records with the Metadata Editor tool during the duration of the project. The WP3-leader team guided the partners with the help of a special manual established for this purpose [EFGMET]. The Metadata Editor is furthermore an important tool for integrating contributions of new providers into the EFG Portal after the end of the project. Archives contributing only a limited number of digital objects can use the tool and create the respective records directly with the tool. In these cases the delivery and ingestion of XML exports into the EFG database is not necessary. The first archive creating metadata records for its film

⁵“Hands on workshop EFG backend tools” (Pisa, September 2010) – minutes available on EFG project members web site: http://www.efgproject.eu/members/members-wp_3_Sep10_workshop.php, Workshop “Data Cleaning and Enrichment for EFG” at EFG Plenary (Frankfurt, Sept/Oct 2010) – minutes available on EFG project members web site: http://www.efgproject.eu/members/members-frankfurt2010_workgroup_sessions.php

content directly in EFG containing the respective links to the digital objects and previews was the Swedish Film Institute.

6.2 A Cataloguing Example

The following paragraphs give a concrete example of how an archive enriched its digital object records with further EFG-relevant metadata with the help of the Metadata Editor tool.

Figure 9 to the right shows a film still from the Det Danske Filminstitut depicting the actor “Ove Sprogøe” as “Egon” in the feature film “Olsen Banden over alle bjerge”. In the metadata Ove Sprogøe’s name is not indexed. In order to make this still retrievable when users perform searches for “Ove Sprogøe” in the EFG Portal the Det Danske Filminstitut image record needed to be enriched with this person. After reviewing the priorities in its EFG cataloguing plan, DFI’s cataloguing staff decided to enrich image records with persons for important films of the Danish Filmography. The enrichment was done by connecting the image record with the person authority record in the EFG database.



Figure 9: Film Still “Olsen Banden over alle bjerge”

As a first step the DFI cataloguer searched for the respective film still via the Metadata Editor tool. The film still is listed as a result by the EFG Content Checker tool (to which the Metadata Editor tool is connected) as shown at the top of Figure 10 below. By pressing the button „Edit this record“ the NonAVCreation record (film still) is displayed in an entry mask of the Metadata Editor Tool which gives the possibility to establish relationships to further persons (Figure 10, at the bottom). Through the search field “Person name” in the film still record the cataloguer found the Person record for “Ove Sprogøe” (as can be seen in Figure 11) and established the relationship between both records. Figure 11 shows all data the DFI catalogued locally for “Ove Sprogøe”. All mandatory EFG fields were already indexed in the Danish Filmography.



Figure 10: NonAVCreation Record (Film Still) in EFG Metadata Editor

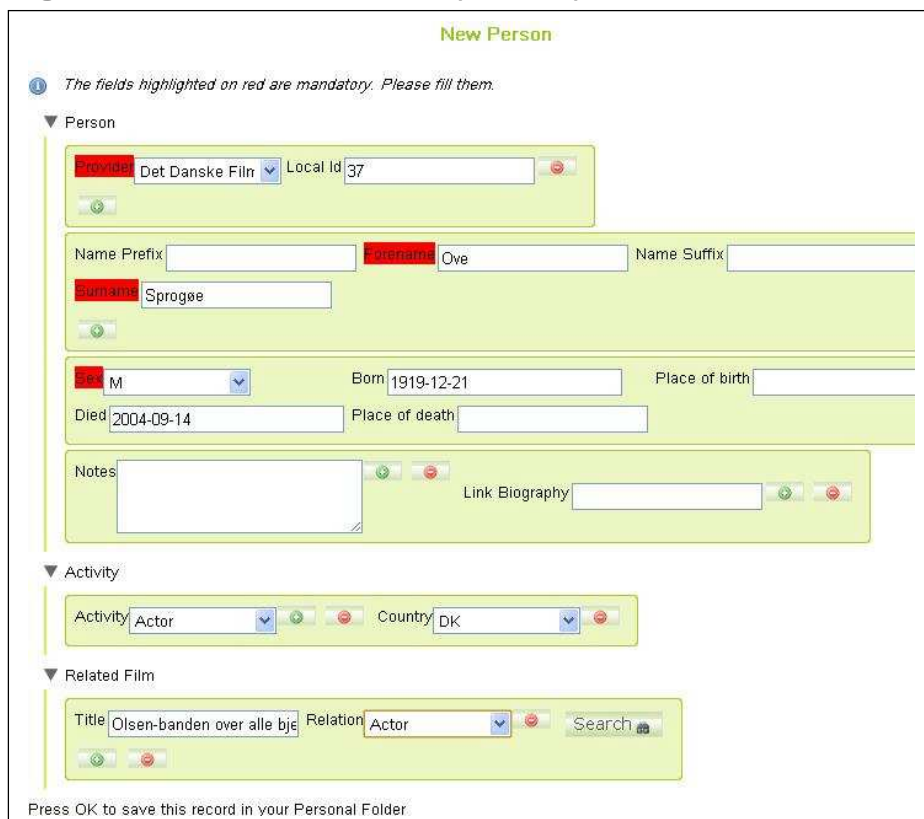


Figure 11: Person Record for “Ove Sprogø” in Metadata Editor

7 EFG's Approach to Authority File Building

A second aim for WP 3 was to establish a common European filmography in the common EFG Information Space of aggregated data as stated in the EFG Description of Work:

*“The common filmography will serve as an authority file of European **film works, persons and corporate bodies**, providing the highest possible precision in information and content retrieval”.*

This aim should be reached by establishing uniquely identifiable authority records in the common EFG database. This chapter reports on the approach followed by WPs 3 and 2 to establish authority files in EFG. These activities refer to step 7 of the EFG ingestion and metadata editing workflow (see figure 1).

7.1 Challenges

As a first step it was necessary to ensure that all kinds of filmographic information held in the film archive databases could be integrated into the common EFG database. For this reason, a complex metadata schema was defined for EFG by WP 2 which is based on the FRBR oriented Cinematographic Works Standard (EN:15907)⁶. Mapping and converting the heterogeneous source data into the common metadata schema was mainly carried out in the framework of WP 2 (see figure 1, steps 2-5). Since the film institutions do not use standardized metadata structures for their exports, and since the archives usually sent different export structures for each kind of data (also for filmographic data), the EFG ingestion process was very resource-intensive but could successfully be accomplished by WP 2. In total, more than 60 different data exports from 16 film institutions were integrated into the common EFG database.

Second challenge was that information about the same person or film work came from different data sources due to numerous European co-productions. Thus, the ingestion into the common EFG database first of all resulted into a considerable amount of duplicates of person and film work entities. It was decided to focus on persons and film works during the definition of the procedure to create authority records, as corporate entity duplicates were not considered quite as critical for EFG's common filmographic database. EFG's aim was that all duplicate information about the same film or person were merged under one single authority record by copying information from one record (“looser record”) into the other one (“winner record”). The winner record was to remain

⁶For more information about the EFG schema please refer to the documents under “EFG Metadata Schema & Vocabularies” in the “Guidelines & Standards” page of the EFG project website:
http://www.efgproject.eu/guidelines_and_standards.php

in the EFG database. This procedure has been used by librarians for a long time, and EFG is the first pilot project to make strong efforts to introduce this concept to the film archive domain. Figure 12 below illustrates an example how information pertaining to the same film work "Metropolis", which was contributed by two film institutions (DFI & DIF), can be expressed in the common EFG schema. In this fiction film example, information from both sources were merged into one single record.

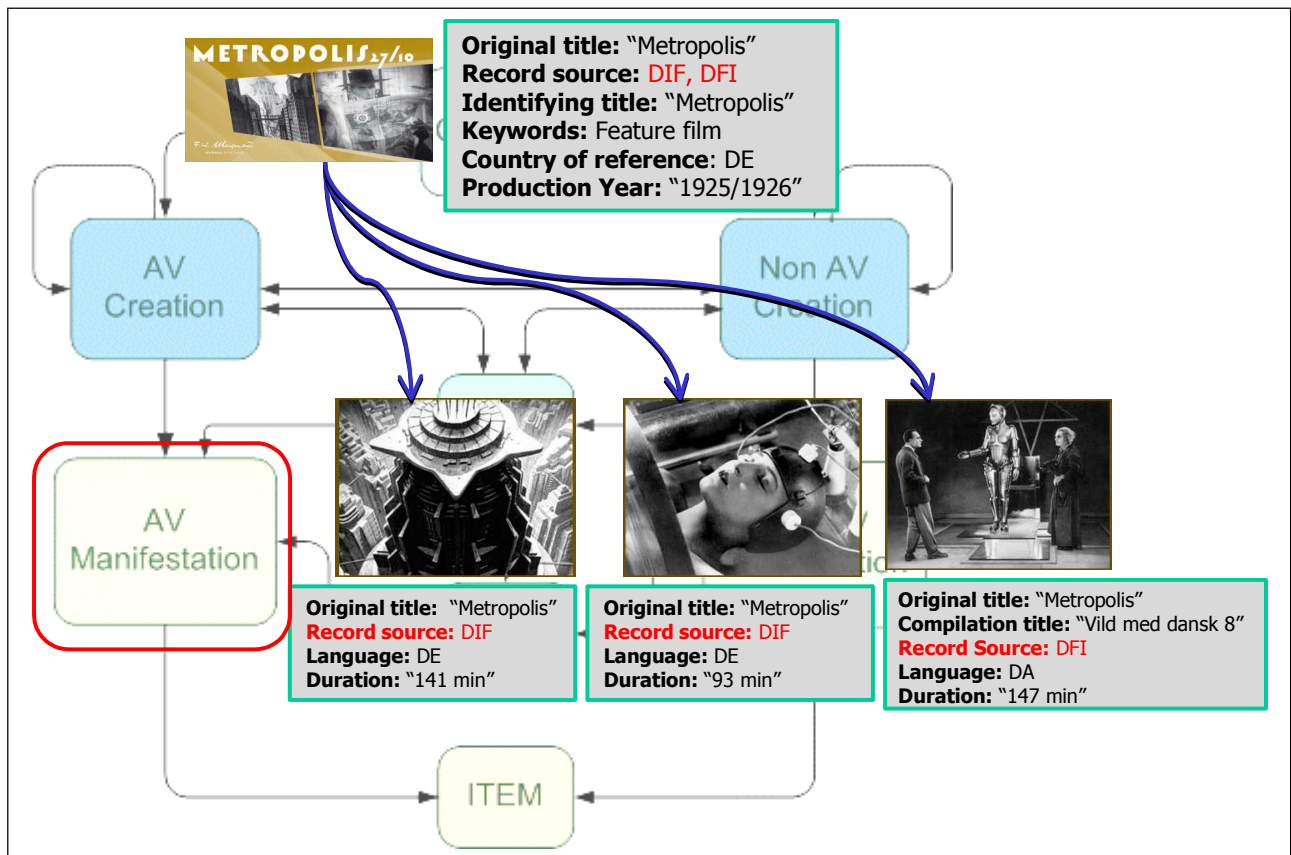


Figure 12: Authority Record in EFG Metadata Schema

7.2 Doublet Cleaning with the Authority File Manager

In order to support the cataloguers from the film institutions to identify records describing the same person or film work in the EFG database, ISTI-CNR developed a special web tool: the Authority File Manager (AFM). The tool was put into practice in October 2010, after feedback was requested from the partner archives in a series of workshops carried out between ISTI-CNR, DFI and WP 3 members, which lead to final adjustments of the tool. The functionalities of the Authority File Manager are already in depth described in deliverables 3.1 and 4.6, and therefore not repeated

here. The following paragraphs focus on how the tool was used in EFG by the archival cataloguers.

During the duration of the project, the tool was used to identify doublets in the archives' local databases. This work was guided by the WP3-leader team with the help of a manual which was established for this purpose: The Authority File Manager manual [EFGAFM]. It turned out that the Authority File Manager detected a considerable amount of local duplicates, so a cleaning round within the archives' local data was considered useful to improve the data quality in EFG. After the ingestion of its contribution into the EFG pre-production information space (ill. 1 step 5), the respective archive accessed the Authority File Manager to view all possible duplicates detected for its contribution in a result list as illustrated in the figure hereunder (more information in D3.1 p. 54):



The screenshot shows the 'EFG - Authority File Management' interface for 'Persons'. It includes navigation buttons like 'Back to Dashboard' and 'Help', and status indicators for 'merged (2)' and 'ignored (1)'. A 'Duplicates: 14715' count is shown. A pagination bar indicates 'Pages: << < [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >>'. The main content is a table with columns for Score, Name, Provider, Name, and Provider.

Score	Name	Provider	Name	Provider
100.0%	Jean Claude Petit.	portal DFI	Jean-Claude Petit	portal DFI
100.0%	Stefania D'Amario	portal DFI	Maria Stefania d'Amario	portal DFI
99.2%	Max Gülstorff	portal EYE	Max Walter Gülstorff	portal DIF
99.2%	David Friedmann	portal DIF	Erich-David Friedman	portal DIF
99.1%	Duta Skirtladze	portal DIF	Demetre >Duta< Skhirtladze	portal DIF
99.1%	Hans Schwarze	portal EYE	Hans-Joachim Schwarz	portal DIF
99.1%	Hans Schwarze	portal EYE	Hans-Heinz Schwarz	portal DIF
99.1%	Hans Schwarze	portal EYE	Hans-Dieter Schwarz	portal DIF
99.1%	Hans Schwarze	portal EYE	Hans Dieter Schwarz	portal DFI
99.1%	Claus Rathjens	portal DIF	Claus-Peter Rathjen	portal DIF
99.1%	Eva Maria Meineke	portal DIF	Eva-Maria Meinecke	portal DIF
99.1%	J. J. Johnston	portal DFI	J.J. Johnson	portal DFI
99.1%	Hans Hennings	portal EYE	Hans-Peter Henning	portal DIF
99.1%	Hans Geissler	portal DIF	Hans-Joachim Geisler	portal DIF
99.1%	Peter Kirschner	portal DIF	Hans-Peter Kirohner	portal DIF
99.0%	William Stranz	portal DIF	William von Strantz	portal DIF
99.0%	Theresa Scholze	portal DIF	Theresa-Sophie Scholz	portal DIF
99.0%	Steven Posters	portal DFI	Steven B. Poster	portal DFI
99.0%	Heinz Konrads	portal DIF	Karl-Heinz Konrad	portal DIF
99.0%	t Hermann	portal DFI	Norman T. Herman	portal DFI
99.0%	Philippe Gérardi	portal DFI	Philippe - Gérard	portal DFI

Figure 13: Authority File Manager

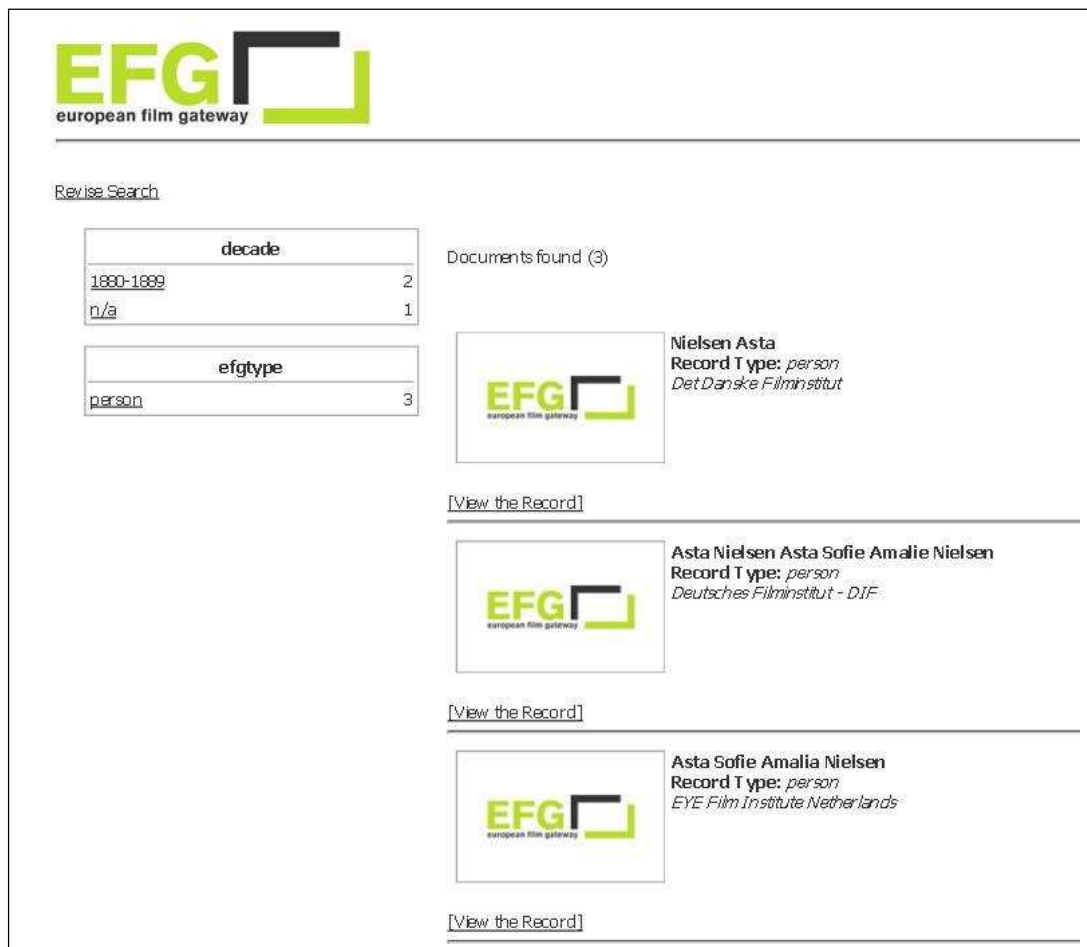
Thereafter, archival cataloguers verified which of the potential duplicates are actual ones with the help of filmographic resources (for instance: national filmographic databases, The Internet Movie

Database) and performed the corrections in their local databases. Thus, real duplicates were merged into authority records at the local level inside the archives' cataloguing systems. After finishing work on the list of duplicates, the providers delivered their cleaned records to ISTI-CNR, who re-ingested them into the EFG pre-production information space (either through new XML exports or OAI-PMH harvesting). The work done by the individual archives to clean authority records locally is further described in *Annex III: Evaluation Data Cleaning and Enrichment in Source Databases*. It must be mentioned that only film archives who manage person or film work authority records locally were able to perform this task.

Thereafter, it was planned that archival cataloguers should identify and merge their own records with those coming from the other archives into single EFG records. An example for the expected result is illustrated in figure 12 above. However, the work of creating EFG authority records could not be accomplished by the archives in the time period allotted to the project for the following reasons:

- 1) The Authority File Manager was released at a late stage of the project (October 2010).
- 2) The EFG system needed too much capacity to run the detection of duplicates among all providers' authority data and to allow for the merge directly in the EFG database (total numbers: 140.000 film works, 251.000 persons). The risk that the Match&Merge procedure of the AFM conflicted with other ingestion and metadata editing activities was considered as too high (please see also Deliverable 4.6, chapter 3.3.4 for details).
- 3) In addition, after the release, the procedure to compare archives' data automatically needed further refinements because the quality of the film work and person information contributed by the archives was very heterogeneous in terms of:
 - **Cataloguing Practices:** EFG archives do not use common cataloguing rules, no partner references its person or film work data to an external authority file
 - **Metadata Structures:** Archives do not use no common metadata formats to deliver this kind of information to EFG
 - **Authority Records:** Not all partners manage them, but instead most deliver only names and titles: 14 partners delivered authority records for film works, 8 for persons, 5 for corporate entities.

The WP3-leader team decided to focus on improving the quality of authority records in the local databases for the named reasons. Thus, the work to establish authority records directly in the EFG database still remains, as can be seen in the figure below (person duplicates for "Asta Nielsen" contributed by three providers):




The screenshot shows the EFG search interface. On the left, there are two filter tables:


decade	
1880-1889	2
n/a	1

efgtype	
person	3


Documents found (3)

Record 1:  **Nielsen Asta**
Record Type: *person*
Det Danske Filminstitut

[\[View the Record\]](#)

Record 2:  **Asta Nielsen Asta Sofie Amalie Nielsen**
Record Type: *person*
Deutsches Filminstitut - DIF

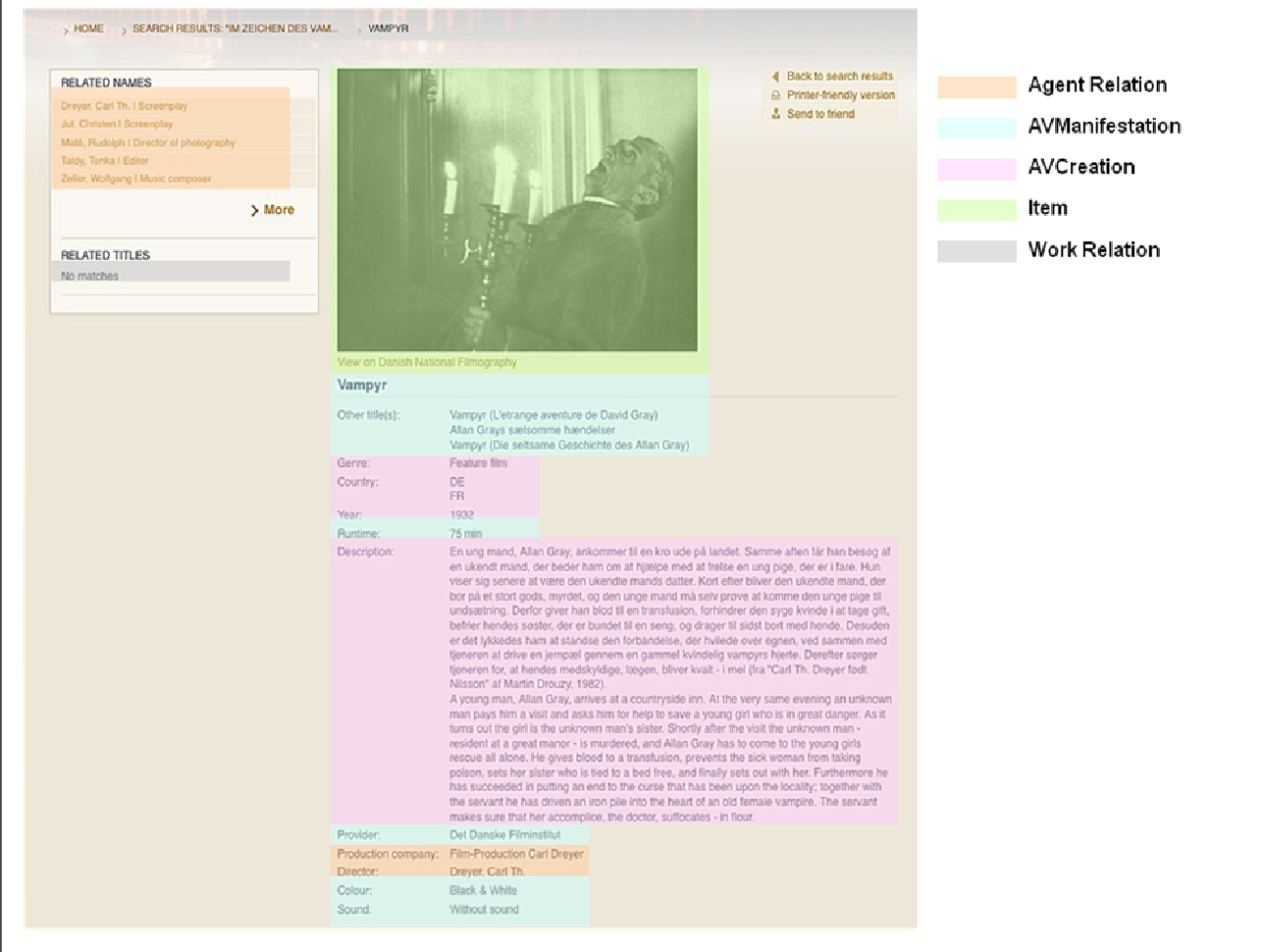
[\[View the Record\]](#)

Record 3:  **Asta Sofie Amalia Nielsen**
Record Type: *person*
EYE Film Institute Netherlands

[\[View the Record\]](#)

Figure 14: Doublets in EFG Database (viewed via Content Checker Tool)

Main consequence of the lacking EFG authority records was that less filmographic metadata could be displayed and exploited for the EFG Portal as initially planned (see D3.1 for details). This affected mainly the Person and Film Pages which could not be implemented by WP 4. However, this was considered as not so severe since the EFG Portal is not a filmography but mainly a search engine for digital material held in Europe’s film archives. Nevertheless, the EFG Portal shows filmographic data (person and corporate entity names, film titles which are directly related to digital objects) which help users to contextualise the retrieved digital material. An example for how filmographic data are displayed in the EFG Portal is given in figure 15 below. The colours indicate from which parts (entities) of the EFG schema the respective metadata come.



> HOME > SEARCH RESULTS: "IM ZEICHEN DES VAM... > VAMPYR

RELATED NAMES
 Dreyer, Carl Th. | Screenplay
 Jul, Christen | Screenplay
 Malt, Rudolph | Director of photography
 Taldy, Tonka | Editor
 Zeller, Wolfgang | Music composer
 > More

RELATED TITLES
 No matches

View on Danish National Filmography

Vampyr
 Other title(s): Vampyr (L'étrange aventure de David Gray)
 Allan Grays sælsomme hændelser
 Vampyr (Die seltsame Geschichte des Allan Gray)

Genre: Feature film
 Country: DE
 FR
 Year: 1932
 Runtime: 75 min
 Description: En ung mand, Allan Gray, ankommer til en kro ude på landet. Samme aften får han besøg af en ukendt mand, der beder ham om at hjælpe med at frelse en ung pige, der er i fare. Hun viser sig senere at være den ukendte mands datter. Kort efter bliver den ukendte mand, der bor på et stort gods, myrdet, og den unge mand må selv prøve at komme den unge pige til undsætning. Derfor giver han blod til en transfusion, forhindrer den syge kvinde i at tage gift, befrier hendes søster, der er bundet til en seng, og drager til sidst bort med hende. Desuden er det lykkedes ham at standse den forbandelse, der hvilede over egnen, ved sammen med tjeneren at drive en jernpæl gennem en gammel kvindelig vampyrs hjerte. Derefter sørger tjeneren for, at hendes medskyldige, lægen, bliver kvalt - i mel (fra "Carl Th. Dreyer fodt: Nilsson" af Martin Drouzy, 1982).
 A young man, Allan Gray, arrives at a countryside inn. At the very same evening an unknown man pays him a visit and asks him for help to save a young girl who is in great danger. As it turns out the girl is the unknown man's sister. Shortly after the visit the unknown man - resident at a great manor - is murdered, and Allan Gray has to come to the young girls rescue all alone. He gives blood to a transfusion, prevents the sick woman from taking poison, sets her sister who is tied to a bed free, and finally sets out with her. Furthermore he has succeeded in putting an end to the curse that has been upon the locality; together with the servant he has driven an iron pile into the heart of an old female vampire. The servant makes sure that her accomplice, the doctor, suffocates - in flour.

Provider: Det Danske Filminstitut
 Production company: Film-Production Carl Dreyer
 Director: Dreyer, Carl Th.
 Colour: Black & White
 Sound: Without sound

Legend:
 Agent Relation (orange)
 AVManifestation (light blue)
 AVCreation (pink)
 Item (light green)
 Work Relation (grey)

Figure 15: Filmographic Information in EFG Portal (Detail Page)

7.3 Achievements and Lessons Learned

Even though it was not possible to establish reliable EFG authority records in the time of the project, EFG WPs 2 and mainly 3 made important steps in this direction and have accomplished the following achievements in this matter:

- Set up and tested workflows for how authority records can be created in a common database with the help of semantic technology tools developed by EFG
- Increased quality of person and film work records in archives' **local databases** (partners use the AFM tool to clean their doublets)
- Increased quality of name and title records in the **EFG database** (normalisation, represented data in a uniform format)
- Laid the groundwork for a common European filmography

- Created an understanding of what data quality means in the scope of authority file building and maintenance within the film archive community and beyond

The lessons learned by EFG Work Packages 3 and 2 regarding authority files can be summarized as follows:

- Maintenance of a common registry of filmographic authority records cannot be assured by EFG. WP 3 recommends to follow-up on this task in the framework of other initiatives.
- Projects like EFG organized as best practice networks can only come up with experiences and recommendations how authority records can be created and managed in the film archive domain.
- Matching & merging heterogeneous data from 16 different data sources was not first priority in this pilot project. The AFM tool needs to be further refined for this purpose.
- Further work needs to be invested to create unique references for persons and film works and corporate entities managed by film archives and by EFG.

As already mentioned earlier, EFG Portal user tests revealed that users desire browsing functionalities, as well as the possibility to do controlled searches for person names and film titles. Thus, EFG WP 3 considers the further establishment of EFG authority records and the development of respective search functionalities in the portal as relevant tasks for the future.

Another promising future approach for the film archives may be cooperations with national libraries in order to create reliable authority records for persons and corporate entities. A first pilot project between Deutsches Filminstitut and Deutsche Nationalbibliothek (German National Library) is scheduled to start in autumn 2011. Person records from the Deutsches Filminstitut are going to be linked with those from the German National Library with the help of the linked open data technology. Main aim of this project is to make person authority records sustainable by introducing the GND (Gemeinsame Normdatei) as a common reference for person records in the film archive domain. This procedure, however, requires that film archives manage their person and corporate entity data in authority records. The creation of a common registry of film works still remain a future task for the film archive community. It could be built upon the aggregated film work data in EFG, but for this aim the archives would need to invest extensive further work in applying film archive cataloguing rules and metadata standards. Standardisation initiatives working in these fields and with which EFG WP 3 collaborated are:

- FIAF Cataloguing Commission (for common cataloguing rules and vocabularies)
- CEN/TC 372 (for common metadata structures).

8 EFG Data Quality Workshop

Deliverable 3.1 “Report on Type and Quantity of Archival Resources Tagged”, which was finalised in September 2010, generated great interest by the film archive and Europeana communities. The WP 3 members therefore decided to organise a workshop in which the outcomes of EFG WP 3’s work could be shared and discussed with other communities. Milestone 3.7 “Quantitative and qualitative assessment of content tagged and filmographic authority records established” was redefined into a corresponding workshop with the title “EFG Workshop on Data Quality and Semantic Interoperability Issues in European Film Archives”. The workshop was carried out on 30 May in Frankfurt. It was also supported and disseminated by Europeana. Minutes and presentations are publicly available on the EFG project web site:

- http://www.efgproject.eu/Data_Quality_Workshop_30May11.php

Overall aim of the workshop was to sum up what EFG had achieved in the fields of cataloguing, authority files and controlled vocabularies as well as to share these experiences with relevant third parties. Around 50 representatives from European film institutions and other cultural heritage initiatives participated in the workshop, apart from the EFG WP3 members, the participants were representatives of:

- FIAF Cataloguing Commission
- Standardization group for cinematographic works standard (CEN/TC 372)
- Film (heritage) institutions and ACE members (e.g. Deutsche Kinemathek – Museum für Film und Fernsehen, Bundesarchiv-Filmarchiv, Checkpoint Media / Österreichisches Filmmuseum, Filmmuseum Düsseldorf, Cinémathèque Suisse, British Film Institute, Kinoteka na Makedonija, Film and Television University Potsdam-Babelsberg)
- Europeana and related projects (Europeana office, PrestoPrime, Swedish National Heritage Board)
- Libraries and digital library initiatives (e.g. Deutsche Nationalbibliothek, Bibliothèque nationale de France / The Virtual International Authority File)
- External information consultants involved in museums’ networks

The workshop consisted of four sessions: Cataloguing, Authority Files, Linked Open Data, Vocabularies. In each session also external speakers reported on their recent work, for example the revised FIAF cataloguing rules, Europeana’s data enrichment and linked open data activities, the VIAF - Virtual International Authority File as well as the planned co-operation between the German National Library and the Deutsches Filminstitut to establish common person authority files. In particular, the presentations and discussions focused on the following issues:

- Film archive vocabularies and vocabulary management
- Cataloguing in the film archives
- EFG’s approach to authority file building

- Linking film archive data to the semantic web
- Integrating film archive data into external authority files
- Standardization of cataloguing and vocabulary work within the film archive sector and using shared platforms (wikis)

Feedback from the workshops' participants revealed that it was appreciated as an important and necessary forum in which the different institutions and initiatives could discuss data quality and semantic interoperability issues across domains on the European level, also beyond the film archive sector. Or as a presenter put it in a nutshell: "Future is in the semantic web! Beyond libraries!".

A major benefit for the film archive community was that the workshop put EFG's WP 3 work into context with existing film standardisation initiatives. It is worth highlighting that the CEN standardisation group will publish EFG's vocabularies on its newly launched website filmstandards.org which serves as a common platform to discuss metadata issues among the members of the community. Also, the FIAF Cataloguing Commission showed great interest in EFG's cataloguing experiences, and valued the input for its work on revising the FIAF cataloguing rules. Furthermore, as already mentioned earlier in this deliverable, this initiative uses EFG's multilingual filmographic vocabularies for the translation of the FIAF Glossary of Filmographic Terms.

The workshop was also a great opportunity for the film archive participants to learn from the experiences of other projects working with similar issues like EFG. Since, as already mentioned in chapter 5.5 *Achievements and Lessons Learned*, a main challenge was to update and maintain the EFG vocabularies, the WP3-leader team invited an information consultant to present the vocabulary management tool xTree by means of selected film archive vocabularies. This web-based tool allows to manage thesauri and other vocabularies collaboratively. It was developed by the digiCULT project for the purpose to digitally record and publish inventories of museums. Before digiCULT, the museums faced the same problem like the film institutions that no common vocabulary tools were used within the different institutions. The workshop participants discussed how a corresponding tool could be introduced into the film archive sector. Even though such a tool could not be implemented in EFG, the film archive participants expressed that they would appreciate to be able follow up on this task in the framework of possible future projects.

9 Evaluation of EFG WP 3 Work by the Partners

This chapter sums up on partner archives' opinion on the work carried out for WP 3 and the relevance of it for their daily routine as cataloguers.

9.1. Questionnaire for Partner Archives

To conclude the WP3 work, the WP3-leader team sent out a questionnaire to the partner archives asking for feedback and experiences regarding their cataloguing work for the EFG project in June 2011. For the questionnaire sheet please refer to *Annex III: Evaluation Data Cleaning and Enrichment in Source Databases*. The overall feedback by the 13 partners which replied to the questionnaire was positive and about 80% of the EFG archives considered their WP 3 work as very relevant to their local cataloguing practices. Replies from the partners were treated confidentially and anonymously, also criticism was welcomed. Each partner submitted one questionnaire on behalf of the whole cataloguing team (which consisted of 1 up to 5 staff members). Following up is a summary of the replies for each of the four according question categories.

9.2 Results Question 1: Impact on Local Cataloguing

11 out of 13 partners confirmed that the quality of their local databases and cataloguing processes improved during the EFG project (question 1a). Most partners reported that their local cataloguing quality improved much (6) or even very much (2) while in three cases the quality improved a little and only in two cases not at all (q. 1c). The partners highlighted the following positive impacts from the EFG work on their local cataloguing practices (q. 1b):

- *Five partners reported an overall improvement of their cataloguing practices and data quality, notably because the EFG WP 3 work encouraged them to discuss problems concerning cataloguing rules and practices and to carry out respective cataloguing work more in-depth*
- *Three partners highlighted their data cleaning activities and reported that respective data became more coherent and homogeneous (particularly person and corporate entity names)*
- *Two partners underlined that EFG cataloguing guidelines and recommendations were in line with international standards and thus supported standardization of local metadata*
- *Major benefit for one partner was the possibility to get in contact with other film archive institutions from the EFG network in order to share experiences about indexing and cataloguing issues*

Most partners (11) indicated that the EFG cataloguing work was in line with their institutional activities (q. 1d). The partner archives also faced difficulties when cataloguing their data for EFG, which were in particular (q. 1e):

- *Four partners reported IT problems, for instance: aging/outdated technical equipment, lacking IT knowledge in the local EFG team, technical difficulties to export data or establish OAI-PMH interfaces for EFG*
- *Three partners indicated that cataloguing for EFG was very time consuming. Either the timetable was too strict for in-depth cataloguing or specific cataloguing tasks could not entirely be accomplished because the workload was unrealistically high*
- *One partner underlines unexpected rights clearance problems for the material which was originally slated for EFG*
- *Understanding the mappings and the transmitting of data from the local to the EFG database was mentioned by two partners to have been a challenge*

Most of these problems were solved in constructive ways and by help of the EFG project co-ordinators as well as the WP3-leader team. In particular, they found the following solutions to overcome the above mentioned problems:

- *Hiring or subcontracting qualified IT specialists (4 partners)*
- *Asking for the help and guidance of EFG coordinators and WP 3 leaders (3 partners)*
- *Concentrating on the most important cataloguing tasks, for instance focusing on the most relevant films regarding adding person names of the persons depicted on still images. (2 partners)*
- *Cataloguing records and material according to EFG and standardized guidelines (2 partners)*
- *Building a local EFG team according to competences, and training staff members (1 partner)*
- *Transferring data manually through the filmarchives-online system to EFG (1 partner)*
- *Spending more resources on rights clearing, and clearing more films (1 partner).*

All 8 partners which replied to question 1f indicated that database fields in which they index filmographic information (such as person names and film titles, film work specific data) improved most significantly from the EFG cataloguing work.

9.3 Results Question 2: Partners' Feedback for WP3-leader team

In this section the partners were asked to give their opinion on the guidance given by the WP3-leader team. As the table below shows, most partners used the WP 3 documents and guidelines and found them useful for their EFG cataloguing work (q. 2a). The partners considered the data

enrichment and cleaning guidelines, the EFG vocabularies as well as the matching manual as the as especially useful. It is worth mentioning that the manuals for the Authority File Manager and the Metadata Editor haven't been used by all partners even though they find them useful. The reason is that the tools as well as the manuals were only finalized late in the project as explained earlier in this deliverable.

Total overview 2a	<i>I used it and found it useful.</i>	<i>I used it but found it not useful.</i>	<i>I have not yet used it but find it interesting.</i>	<i>I have not yet used it and find it not interesting.</i>	<i>I can't tell.</i>	<i>Total</i>
Cataloguing plan	8	1	2	1	1	13
Matching manual	11	1	1	0	0	13
Guidelines for data cleaning and enrichment	12	0	1	0	0	13
EFG vocabularies	12	1	0	0	0	13
WP 3 reports (Milestones & Deliverables)	8	2	2	0	1	13
Data Provider Handbook	6	0	3	0	4	13
Authority File Manager manual	4	0	5	0	4	13
Metadata Editor Manual	5	0	6	0	2	13
Evaluation sheet to monitor your cataloguing work for EFG	7	2	1	0	3	13
Delivery schedule for your EFG data contributions	9	0	1	0	3	13
Total	82	7	22	1	18	130

Table 4: Partners' Answers for Question 2a (Usefulness of WP3 Guidelines and Documents)

Also, the WP 3 workshops and meetings were considered as useful or even very useful by most partners to support their cataloguing and data submission work (q. 2b). The relatively high number of indications in the column "I can't tell" means that the person who replied to the questionnaire did not attend the respective meeting or workshop. Most partners indicated that they specially benefited from the first joint WPs 2&3 meeting in Pisa. In the beginning of the project these two WPs co-operated on establishing semantic and syntactic interoperability rules for EFG which were documented in milestone 3.2 [EFGM322010]. The workshop was used to present the results of the survey on the archives' local databases to which 9 of 14 content providers had answered in October 2009.

Also, the “WG 3 Workshop on Cataloguing Rules and Vocabularies” in May 2009 in Copenhagen was considered as very useful for most partners’ WP 3 work. With the results of the survey and the discussions at the workshop, WP 3 members decided not to apply common cataloguing rules to enrich and harmonise their data for EFG purposes. Therefore, it was decided to establish controlled vocabularies that local database terms could be matched to and that would allow for a consistent display of data in the EFG Portal. So, after the workshop in Copenhagen the main activity in WP3 became the development of EFG vocabularies. The relevance of this workshop reflects also the results from the previous question that partners considered the established vocabularies as one of the most useful WP 3 outcomes for their cataloguing work.

At a stage of the project in which prototypes of the EFG backend tools (Authority File Manager, Metadata Editor, Content Checker) became operational in September 2010, the WP3-leader team organised a series of workshops in order to request feedback on the tools’ usability from the partner archives and to carry out practical exercises (hands-on workshop in Pisa, WP3 session at EFG Plenary in Frankfurt). Even though only five partners worked with the tools after their finalisation in May 2011, 11 partners were satisfied with the workshops which is mainly due to the fact that the discussions raised the partners’ understanding of what needs to be done and what can be done in order to establish reliable authority files within the film archive domain.

The open Workshop "Data Quality and Semantic Interoperability Issues in European Film Archives" carried out by DFI and DIF in Frankfurt served as an important platform to bring forward the discussion on standardised cataloguing work and vocabularies. This workshop was also attended by representatives from other digital heritage communities and all participating WP 3 members considered the workshop as an important forum to discuss and share the experiences from their EFG cataloguing work with other initiatives.

Total overview 2b	<i>I participated and found it very useful.</i>	<i>I participated and found it useful.</i>	<i>I participated but found it not very useful.</i>	<i>I participated but found it not useful at all.</i>	<i>I can't tell</i>	<i>Total</i>
EFG WG 2&3 Kick-off meeting 17-18 Nov 2008, ISTI-CNR, Pisa	6	2	0	0	5	13
WG 3 Workshop on Cataloguing Rules and Vocabularies 11-12 May 2009, DFI, Copenhagen	5	3	1	0	4	13
Hands-on workshop EFG backend tools 14 - 15 September 2010, CNR-ISTI, Pisa	2	1	1	0	9	13

Workshop session "Data enrichment and cleaning for EFG" at EFG Plenary 30.9. – 1.10., Frankfurt	5	3	1	0	4	13
Workshop "Data Quality and Semantic Interoperability Issues in European Film Archives" 30 May 2011, DIF, Frankfurt/Main	3	2	0	0	8	13
Total	21	11	3	0	30	65

Table 5: Partners' answers for question 2b (usefulness of WP3 meetings and workshops)

12 out of 13 partners were satisfied with the level of information and communication by the WP3-leader team regarding cataloguing issues (q. 2c). However, from three partners' point of view the following issues could have been addressed better during the project (q. 2d):

- *Further work should have been invested to ensure that metadata displayed in the EFG Portal are more standard-compliant.*
- *One partner suggested that there should have been more time for practical exercises during the WP 3 workshop at the third EFG Plenary Board meeting, and that there should have been more practical "hands-on" workshops overall.*
- *One partner is of the opinion that the WP 3 and 2 leader teams should have been more familiar with the specific local conditions and information environment. Also, delays of the data integration into EFG and of the Metadata Editor Tool were critical because the necessary cataloguing activities on the EFG level had to be postponed. However, with the help of the WP3-leader team the according cataloguing activities were accomplished in time.*

9.4 Results Question 3: Long-Term Impact on Local Cataloguing

The partners gave the following replies to the question regarding which of the cataloguing changes they introduced during the project duration will have a long term impact on their local cataloguing practices (q. 3a):

- *Use recommended standards and EFG cataloguing guidelines for future cataloguing work (4 partners)*
- *Continue to catalogue information in newly established fields (3 partners)*
- *Continue to use new controlled vocabularies (3 partners)*

- For three partners the EFG project revealed the need for new cataloguing rules and practices. They either changed them during the project time or will introduce them after the release of new databases (3 partners)
- Exploit normalized data structures and enriched information (e.g. gender) in other contexts (1 partner)
- New collection management for streaming videos (1 partner)

These replies also conform to those regarding the next question, to which most partners replied that they will use the main WP 3 outcomes (especially: EFG vocabularies, EFG data cleaning and enrichment guidelines) also in future for their local cataloguing work. Other partners which have not used the according documents or tools still find them interesting for future uses. The lack of replies for the vocabulary checker is due to the fact that this tool was mainly used by the WP3-leader team to discover invalid terms within the contributed data. These terms were extracted from the tool and included into the matching tables which were then corrected by the respective partners. However, as partners had the possibility to access the tool via the EFG data backend this answer option was included into the questionnaire.

Total overview 3b	<i>I used it during the project and will also use it in future.</i>	<i>I used it during the project but will not use it anymore in future.</i>	<i>I have not yet used it but find it interesting.</i>	<i>I have not yet used it and find it not interesting.</i>	<i>I can't tell.</i>	Total
EFG vocabularies	8	1	2	0	2	13
Data cleaning & enrichment guidelines	6	2	2	0	3	13
Metadata Editor	1	3	3	1	5	13
Authority File Manager	2	2	5	0	4	13
Vocabulary checker						0
Recommended external cataloguing rules	9	0	2	0	2	13
Recommended external vocabularies	8	0	3	0	2	13
Total	34	8	17	1	18	78

Table 6: Partners' answers for question 3b (long term use of WP 3 outcomes)

9.5 Results Question 4: Personal Feedback from Partners

In the last section, the partners were asked for their personal feedback on the EFG WP 3 work. All partners stated that they especially enjoyed working in a Best Practice Network, in particular (q 4.1):

- *Collaborative work on common vocabularies and cataloguing issues*
- *Possibility to meet cataloguers and archivists from other European film institutions in order to discuss cataloguing problems and to learn from their experiences*
- *Possibility to introduce and to review new cataloguing practices as well as to catalogue collections to a very high degree of consistence and depth which is not possible in the archives' day-to-day practices*
- *The effort to bring European film institutions together and to become more homogenous in data treatment in order to make the archival material publicly available*
- *Professional guidance and collaboration with EFG project coordinators and WP3-leader team*

Partners stated that problems with regard to the Metadata Editor tool and integration of their data into EFG were solved with the help of the WP3-leader team, so that negative impacts on the speed of cataloguing activities were not severe.

The overall conclusion about the EFG WP 3 work was very positive (q 4.2). As members of a Best Practise Network the partners indicated that they extended their knowledge about cataloguing and metadata quality issues to a high degree and that they gathered useful experiences on how to make their metadata interoperable and sustainable.

10 Conclusions – Lessons Learned

The overall objective of WP 3 was to structure and harmonize the heterogeneous metadata contributed by 16 European film archives into the common EFG database. Main challenges were that film archives do not use common standards for metadata structures, cataloguing rules and vocabularies. Furthermore, WP 3 had to deal with the issue that the same person or film work can be contributed by different archival databases to EFG.

In order to tackle these challenges, WP 3 closely collaborated with other WPs, especially with WP 2 (“Technical Interoperability and Access”) regarding metadata ingestion and editing in the common EFG Information Space, WP 2 (“User Needs and Service Requirements”) for defining

metadata based functionalities wished by users and WP 4 (“Service Implementation and Operation, Web Platform”) to implement according functionalities in the EFG Portal.

In the beginning of the project, EFG’s aim was to display a more comprehensive set of metadata in the Portal. Especially filmographic information from the participating film institutions should have been properly exploited in order to contextualise the retrieved digital material more thoroughly in the portal. However, due to the high amount of records delivered by the partner archives, it turned out that it was not feasible for WP 3 to ensure that these very heterogeneous metadata are available in the required quality and to carry out the necessary cataloguing activities (including: doublet cleaning with the Authority File Manager, enrichment with the Metadata Editor, verification of the ingested records with the Content Checker). The common EFG Information Space contains more than 1 million records and six of the 16 partners delivered significantly more than 10.000 records. Therefore, WP 3 changed its cataloguing strategy and focused on providing the metadata available in the portal in the best possible quality from the second project year on.

The WP3-leader team co-ordinated this cataloguing process with the help of the more general “EFG data enrichment and guidelines” as a first step, and established individual cataloguing plans for each partner at the end of the second year. These plans turned out to be an adequate method to discuss with partners whether or in how far they could accomplish specific cataloguing tasks for EFG. For instance, the Danish Film Institute would ideally have indexed the names of the Persons depicted on their images for EFG. Since the workload would have been unrealistically high to index the names for all its 40.000 film stills, priority was given to the most important films of the Danish National Filmography. Similarly, cataloguing plans for all other content contributors were established taking into account their individual cataloguing practices. The plans gave a better insight into the local cataloguing decisions of the partners, and optimized the efforts invested by the staff.

Film archives were encouraged to carry out as much of their cataloguing activities for EFG locally. This had successful and sustainable impacts on the cataloguing quality and practises in the individual institutions, confirmed by 11 out of 13 partners through their WP 3 feedback questionnaires. Against the background that film archives do not apply common cataloguing rules and vocabularies, it is very positive that most of the partners found the main WP 3 outcomes (especially: EFG vocabularies, EFG data cleaning & enrichment guidelines) useful and will use them for their future cataloguing work. Through this and through WP3’s networking with film standardization initiatives (FIAF Cataloguing Commission, CEN/TC 372) and the film archive community, EFG actively participated in the process of making cataloguing practices within film institutions across Europe more homogenous and sustainable.

Even though EFG was not able to establish reliable authority files in the EFG database, the project laid solid groundwork for a European registry of film works, persons and corporate entities, as an extensive part of the EFG database. The metadata contributed by the 16 archives were so heterogeneous that more work than expected needed to be invested to harmonize them in EFG, which was successfully accomplished. The creation of authority files was tested, but it turned out that it demanded too much work to clean doublets in the EFG database during the project time. Therefore, partners cleaned their authority data locally with the help of the EFG Authority File Manager tool. WP 3 presented its experiences and recommendations on how authority records can be created and managed in the film archive domain to the film archive, Europeana and other cultural heritage communities. On the "Open WP 3 Workshop on Data Quality and Sustainability Issues in European Film Archives" on 30 May in Frankfurt the participants agreed that a direction in which film archives should move in the future is to integrate their person and corporate identity data into authority files of national libraries, with the help of linked open data technology. This is envisaged in a first pilot project between Deutsches Film Institut and Deutsche Nationalbibliothek which will built upon EFG's experiences with the Authority File Manager tool. WP 3 recommends to follow up on the task of creating common authority files in the film archive domain in the future.

The vocabulary matching process was very resource-intensive in terms of managing, updating and maintaining the vocabulary files. For future data cleaning work, EFG recommends to apply a vocabulary management tool. Nevertheless, the vocabulary matching process carried out during the project was successful and the partners' multilingual and heterogeneous metadata values were cleaned by using the established comprehensive EFG vocabularies. Through this work metadata are displayed coherently in the EFG and Europeana web portals.

To sum up, it can be said that WP 3 was very successful with respect to the challenges and obstacles it had to deal with. EFG can refer to highly relevant achievements in the fields of cataloguing, vocabularies, authority file building and networking. The WP made very important steps towards harmonizing the heterogeneous metadata held in European film archives' databases. The partners highly benefited from their participation in the Best Practise Network. The discussions, workshops and meetings challenged the partners' cataloguing routines and gave a common understanding of what needs to be done to standardize vocabularies and improve authority data. The workshops gave the opportunity to share experiences with the film archive and other digital heritage communities. The outcomes and experiences of WP 3 had valuable impacts on the film archives and beyond the EFG community. A notable outcome of knowledge sharing is that the EFG vocabulary will be uploaded on the filmstandards.org website, free to be used by relevant third parties outside the EFG community. Furthermore, the FIAF Cataloguing Commission showed great interest in EFG's cataloguing experiences as a reference for its work on revising the FIAF cataloguing rules. This initiative uses EFG's multilingual filmographic vocabularies for the translation of the FIAF Glossary of Filmographic Terms.

11 References (WP 3 Guidelines and Reports)

[EFGAFM]

Manual for the EFG Authority File Manager [AFM]

A manual including the basic instructions on how to use the Authority File Manager [AFM] which is a web tool used by the EFG archives to clean duplicates of person and film work records in their local cataloguing systems. URL to tool's public demo version on EuropeanaThoughtLab: http://www.europeana.eu/portal/thoughtlab_improvingmetadata.html

[EFGD242011]

Deliverable 2.4 „Report on inclusion of archives´ repositories“. Internal WP 2 report. March 2011. Report on data integration and metadata activities.

[EFGD312010]

[Deliverable 3.1 “Report on type and quantity of archival resources tagged”](#). Francesca Schulze, Uffe Smed, Pernille Schütz. September 2010.

This deliverable is a comprehensive report on the work carried out within WP3 during the first two project years. This document is publicly available on the EFG project web site: <http://www.efgproject.eu/outcomes.php>

[EFGD462011]

Deliverable 4.6 „Report on EFG Service Operation and Promotion Activities“. Internal WP 4 Report. September 2011.

[EFGDH2011]

[EFG Data Provider Handbook](#).

This handbook provides all necessary information for film institutions that wish to contribute their data to EFG. The “EFG Data Cleaning and Enrichment Guidelines” can be found in chapter 4 “Preparing Data for EFG” of this handbook. The handbook is publicly available on EFG's guidelines & standards web site: http://www.efgproject.eu/guidelines_and_standards.php

[FIAF2008]

Fédération Internationale des Archives du Film, Jon Gartenberg (ed.), Glossary of Filmographic Terms, Munich 1989. Updated English version from 2008 by Zoran Sinobad available at <http://www.fiafnet.org/publications/Glossary%20of%20Filmographic%20Terms%20%28English%20Version%292008%20revision.pdf>

[EFGMET]

Manual for the EFG Metadata Editor [MET]

A manual with the basic instructions on how to use the Metadata Editor [MET] which is the EFG web tool used to add, edit, delete and enrich metadata records directly in the EFG database. URL to the tool's public demo version on EuropeanaThoughtLab: http://www.europeana.eu/portal/thoughtlab_improvingmetadata.html

[EFGM332009]

Milestone 3.3 "Best practices for Filmographic Editing and Authority File Administration": Internal WP 3 report. It contains the results of a cataloguing survey among EFG partner archives in April 2009.

[EFGM322010]

Syntactic and semantic interoperability rules for EFG. Refers to Milestone 3.2". Internal WP 3 report. Establishment of minimum rules for content enrichment, based on analysis of current and best practises".

[EFGSchema]

Information about the EFG schema can be found under "EFG Metadata Schema & Vocabularies" in the "Guidelines & Standards" page of the EFG project website: http://www.efgproject.eu/guidelines_and_standards.php

[EFGVOC1]

[EFG vocabularies I: Value lists and types for EFG data elements](#)

This excel sheet contains the value lists and types defined for elements of the EFG metadata schema. The vocabulary file is publicly available on EFG's guidelines & standards web site: http://www.efgproject.eu/guidelines_and_standards.php

[EFGVOC2]

[EFG vocabularies II: Types for semantic relationships](#)

This excel sheet contains the types defined for semantic relationships of the EFG metadata schema. The vocabulary file is publicly available on EFG's guidelines & standards web site: http://www.efgproject.eu/guidelines_and_standards.php

Annex I: An EFG Partner's Cataloguing plan

EFG Cataloguing Plan for Year 3 (SEP 2010 – AUG 2011)

Content Provider: Det Danske Filminstitutet - DFI

Date: 2010-09-24

Preliminary note: This cataloguing plan lists data enrichment and cleaning activities your institution should carry out within Year 3 to enhance the quality of the common EFG database. It focuses on the activities which are most important for the EFG web portal and does not claim to be complete. If you do any other kinds of cataloguing than listed in this plan, please be so kind as to inform us about this:

- DIF: Francesca Schulze (XXX)
- DFI: Uffe Smed (XXX)

1. Contributions according to DoW vs. real

Collection / Data set	Type	Numbers DoW	Document type	Numbers real
Danish films	Video	20	Films	20
Danish films or clips	Video	1.150	Films	1.150
Danish silent films, selected names, colour stills	Image	53.000	Stills + Portraits	60.000
Posters	Image	1.000	Posters	2.500
Programs	Text	0	Programs	989
	TOTAL	55.170 Items		64.659 Items

According to its delivery schedule DFI will deliver 9.300 digital items more than indicated in the content list of the DoW. This means that DFI delivers 17,2% more than the original total. The amount of the video material (1.170 hours runtime) will stay the same. To date, DFI delivered around 51.500 items (42.500 images, 200 videos, 8.800 text documents). **Please let us know, if the figures provided above are correct.**

Questions:

- a) How many single video items are the indicated 1.170 hours?
- b) Is it possible to provide us with default entries for the following fields for each of your digital collections: collection, provenance, rights holder, rights holder URL?

2. Data enrichment for EFG

In order to have the best data quality possible for EFG, DFI should invest further time in enriching the data they deliver. If DFI cannot do this locally, there is the possibility to carry out this task directly in the EFG database with the help of the [Metadata Editor Tool \(MET\)](#) which will be finally released in May 2011. We prefer you to do the enrichment locally since your own catalogue will benefit from this. The list hereunder informs you, which kind of data we would like you to enrich and where this should be done. **These activities are probably not in line with your cataloguing priorities. So, please give us your feedback what you will be able to carry out within Year 3.**

Collection(s) / Data set(s)	Activity	Where?	When?	Priority
Image / Film stills	Add relations for <i>Depicted persons</i> . Start with the most film-relevant persons.	MET	Start 1 Nov	High
Image / Posters	Add <i>Creation date</i> and <i>Language</i> (used on the poster)	Locally/MET?	TBD	High
Image / Posters	<i>Description / Title, Keywords</i>	Locally	TBD	Medium
Image / Posters	Add relations for the <i>Depicted persons</i> . Start with the most film-relevant persons.	MET	Start 1 Nov	High
Image / Posters	Index the <i>Creator (Poster designer)</i> or establish the relationship to the <i>Creator</i>	Locally/MET?	TBD	High
Image / Portraits	Add relations for the <i>Depicted persons</i> . Start with the most film-relevant persons.	MET	Start 1 Nov	High
Image / Portraits	Index the <i>Creator (Photographer)</i> or establish the	MET	TBD	High

	relationship to the <i>Creator</i>			
Video / Film clips	Add <i>Language</i> (of the film clips you are contributing) and optionally <i>LanguageUsage</i> (e.g. “Spoken language”, “Subtitle”, etc.)	Locally/ MET?	TBD	High
Persons	Continue to add biographical data for selected film-relevant persons. First priority to those which have a relation to a digital object (<i>Date of birth / death</i> , <i>Place of birth / death</i>)	Locally	TBD	Medium

3. Data cleaning for EFG

Also, data cleaning will be a task that has to be carried out after your data were ingested into the EFG database. In order to avoid doublets, the EFG partners will have to merge person and film records that have been identified by the system as possible doublets. This will be done with the help of the [Authority File Manager \(AFM\)](#) tool. This work, which is scheduled to start in October 2010, will concern your person names as well as your film titles.

Collection(s) / Data set(s)	Activity	Where?	When?	Priority
Film works / Danish Films	Use the duplicate list of the AFM to check whether you have doublets of film titles and merge these.	Locally	4 – 22 Oct	High
Persons	Use the duplicate list of the AFM to check whether you have doublets of film titles and merge these.	Locally	4 – 22 Oct	High
Film works / Danish Films	Check potential doublets of your film titles with those coming from other institutions. Merge doublets or mark non-identical persons as “checked” in the AFM.	Locally / AFM	1 – 30 Nov	High

Persons	Check potential doublets of person records with those coming from other institutions. Merge doublets or mark non-identical persons as “checked” in the AFM.	AFM	1 – 30 Nov	High
Image / Stills	Clean the <i>specific types</i> which are wrong in some cases (e.g. DFI_NonAVCreation_20090701-154523-6 should be “still” instead of “tv serie)	Locally	TBD	High

- Since you clean your doublets locally please deliver new XML exports of your film works and persons *until 22-Oct* to DIF. These data will then be re-ingested into the EFG database in order to prepare step 2 of the EFG authority file cleaning in which you will merge your doublets with those from other institutions directly in the AFM.

4. Further WP 3 activities in year 3

- Update translations of EFG vocabularies (ongoing work)
- Update your vocabulary matching (if necessary)
- Implement an OAI-PMH interface that EFG can harvest your data automatically on a regular basis
- Provide your video thumbnails in higher size (250 px/height)
- **Select and make available further content for EFG. Proof if you can contribute your newly digitised resources (e.g. 1.100 documentaries).**

Annex II: EFG WP 3 Evaluation Questionnaire

What we have learned from the EFG WP 3 work

This questionnaire asks for your feedback and experiences on the cataloguing work you carried for EFG within WP 3. Please complete this questionnaire and send it back to Uffe Smed (XXX) and Francesca Schulze (XXX) until **1 July 2011**. Your data will be treated confidentially and anonymously. So, please do not hesitate to give us your personal and faithful feedback. Also, criticism is welcomed. The results from this survey will contribute to the final WP 3 report “D3.2 Final report on type and quantity of archival resources tagged” and will be presented at the final Plenary Board meeting on 18/19 August in Frankfurt.

Name:

Institution:

Date:

I answer this questionnaire on behalf of the EFG cataloguing team at my institution. *(If you want to fill out this questionnaire from your personal point of view do not click on this box and ignore the next sentence. It is possible that partners send several questionnaires from their institution. **You can double click on the boxes above in order find the opportunity to mark the boxes.**)*

The EFG cataloguing team consists of people. *(Please fill in a number here)*

1: What impact has the EFG project had on your local cataloguing work?

- a) Would you say that the cataloguing process(es) and/the quality of your archive’s database(s) has been improved during the time of the EFG project?

Yes No I can’t tell

- b) If you should mention something that has had a positive effect on your local cataloguing practice then what would you say?

Please write your answer here:

- c) To what degree do you think the quality of your database(s) has been improved?

very much much a little not at all

d) How far was your cataloguing work for EFG in line with your institutional activities?

very much much a little not at all

e) What difficulties did you face during this work and what did you do to overcome them (e.g. technical problems, teaching staff members, etc.)? *(Please expand this table if you want to list further activities)*

<i>What difficulty I faced...</i>	<i>What solution I applied...</i>

f) If you should mention a couple of fields in your database that have been improved then which ones would you highlight?

Please write your answer here:

2: Did you feel well guided by the WP3-leader team?

a) Did you find the various guidelines and documents useful for your WP 3 work?

	I used it and found it useful.	I used it but found it not useful.	I have not yet used it but found it interesting.	I have not yet used it and found it not interesting.	I can't tell.
Cataloguing plan	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Matching manual	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Guidelines for data cleaning and enrichment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
EFG vocabularies	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
WP 3 reports (Milestones & Deliverables)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Data Provider Handbook	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Authority File Manager manual	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Metadata Editor Manual	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Evaluation sheet to monitor your cataloguing work for EFG	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Delivery schedule for your EFG data contributions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

b) Did you find the meetings and workshops useful for your WP3 work?

	I participated and found it very useful.	I participated and found it useful.	I participated but found it not very useful.	I participated but found it not useful at all.	I can't tell.
EFG WG 2&3 Kick-off meeting 17-18 Nov 2008, ISTI-CNR, Pisa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WG 3 Workshop on Cataloguing Rules and Vocabularies 11-12 May 2009, DFI, Copenhagen	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Hands-on workshop EFG backend tools 14 - 15 September 2010, ISTI-CNR, Pisa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Workshop session "Data enrichment and cleaning for EFG" at EFG Plenary 30.9. – 1.10., Frankfurt	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Workshop "Data Quality and Semantic Interoperability Issues in European Film Archives" 30 May 2011, DIF, Frankfurt/Main	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

c) Was the level of information and communication from the WP3-leader team about cataloguing issues satisfactory?

- Yes No I can't tell

d) Is there anything you missed or could have been done better by the WP3-leader team from your point of view?

Please write your answer here:

3: Does the work you carried out for EFG have a long term impact on your local cataloguing practice?

a) Have you introduced changes during project, which you also apply afterwards (e.g. applying a new vocabulary, cataloguing new fields)?

Please write your answer here:

b) Will you use the main outcomes of the EFG project for your local cataloguing activities or other purposes?

	I used it during the project and will also use it in future.	I used it during the project but will not use it anymore in future.	I have not yet used it but find it interesting	I have not yet used it and find it not interesting.	I can't tell.
EFG vocabularies	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Data cleaning & enrichment guidelines	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Metadata Editor	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Authority File Manager	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Vocabulary checker	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Recommended external cataloguing rules	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Recommended external vocabularies	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

4: What is your personal conclusion from the experiences you made during the project?

- a) What did you like and what did you not like?

What I particularly liked:

Please write your answer here:

What I did not like so much:

Please write your answer here:

- b) What have you learned personally from the EFG WP 3 work?

Please write your answer here:

- c) What have you learned personally from the EFG WP 3 work? Is there anything you would like to mention about your EFG WP 3 work which was not covered by the questions above?

Please write your answer here:

Annex III: Evaluation Data Cleaning and Enrichment in Source Databases

1. DATA ENRICHMENT DIGITAL OBJECTS. Total overview of replies.

Question 1: Please enter in the table hereunder which kind of enrichment work you carried out for EFG on your digital collections from 1 Sep 2008 until 31 August 2011. These are the numbers from the beginning of the project until the date you are answering to this evaluation sheet.

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
CCB (Archive)	Corona	VIDEO	All fields newly created. Please take note: size, bit rate, standard were not filled because the work on digital items is not finished.	Creation: Creator, country of reference (geographic origin of the Audiovisual creation), production year, description (IT) Manifestation: Title, Language, Document Size, Duration, Dimension (mt.), Duration. Agents: Persons; Companies Events: Publication event (first projection) – Date Collection: Title of the Collection Format, Gauge, Aspect ratio, Sound, Colour	Both (Hand-edited and automatically)	XXX	
CCB (Archive)	Propaganda	VIDEO		Creation: Creator, country of reference (geographic origin of the Audiovisual creation), production year, description (IT) Manifestation: Title, Language, Document Size, Duration, Dimension (mt.), Duration Agents: Persons; Companies Events: Publication event (first projection) – Date Collection: Title of the Collection Format, Gauge, Aspect ratio, Sound, Colour, Size, Bit rate, Standard (Pal, NTSC)	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
CCB (Archive)	Other (Restored Film Works)	VIDEO	All fields newly created. Please take note: size, bit rate, standard were not filled because the work on digital items is not finished.	Creation: Creator, country of reference (geographic origin of the Audiovisual creation), production year, description (IT) Manifestation: Title, Language, Document Size, Duration, Dimension (mt.), Duration Agents: Persons; Companies Events: Publication event (first projection) – Date Collection: Title of the Collection Format, Gauge, Aspect ratio, Sound, Colour, Size, Bit rate, Standard (Pal, NTSC)	Hand-edited	XXX	
CCB (Library)	Angelo Novi Photos collection	IMAGE	Digital objects from this collection have only partially been catalogued until now. We catalogued around 1330 digital objects of around 25.000.	We worked on the fields "Related film title" and "Description", paying attention to the person names included in the description fields. Title, creator, persons, date, colour, rights holder.	Hand-edited	XXX	
CCB (Library)	Posters collection	IMAGE	The original posters (around 20000) are catalogued but only partially digitised	Title, creator of the film work, creator of the poster, persons, date, size	Hand-edited	XXX	
CCB (Library)	Censorship visa documents	TEXT	We are now able to export also the second part of the db (from 1944-1955) 18201 records.	We worked on the fiels "title" , "other titles" and "Verdict"	Hand-edited	XXX	
CCB TOTAL							XXX
DFI	Video clips	VIDEO	786 video clips are uploaded and linked via dfi.dk		Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
DFI	Poster and stills collection	IMAGE	Our Danish poster and stills collection is partial enriched with relevant metadata. One main still is selected for each film title. Erased doublets and replaced poor digitized objects.	Credits: Stills photographer, director. Title and production year. Poster artist	Automatically	XXX	
DFI	Workshop film stills	IMAGE	1200 digitized stills from the workshop collection are implemented in the stills database. One main still is selected for each film title. We have uploaded, catalogued and linked around 250 new films to dfi.dk since the last evaluation period.	The keyword "workshop" is added for each still. Credits: Stills photographer, director. Title and production year.	Hand-edited	XXX	
DFI	Top 100 Danish feature film	IMAGE	Names added for most relevant via our image database 'Fotorama'	Actor depicted	Hand-edited	XXX	
DFI	Danish feature films	IMAGE	Person names added for most relevant feature films via our image database 'Fotorama' from the period: 2011-1975	Actor depicted + keywords added	Hand-edited	XXX	
	Portraits	IMAGE	Uploaded, catalogued and linked to dfi.dk		Hand-edited	XXX	
DFI TOTAL							XXX

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
DIF	Artur Brauner collection	TEXT	Digital objects from this collection have only partially been catalogued in the past, so a full revision and further indexing of the digital objects was carried out taking into account the database fields listed in the column to the right.	creator, description (languages: GE, EN, FR), title, document language, document size, document type, rights holder, resolution, colour bitrate, file size.	Hand-edited	XXX	
DIF	Photo Collection	IMAGE	Photos have been partially catalogued in the past. Further indexing work was carried out for the database fields listed in the column to the right. Relationships between were established between the digital objects and the authority records (film work, person) of the filmographic database filmportal-zdb.	descriptions (languages: EN, FR), Specific type, Data from the field "Source" were split up into the ESE elements "Creator", "Publisher" and "Rights"	Hand-edited	XXX	
DIF	Special collection	TEXT	Creation of new object records		Hand-edited	XXX	
DIF	Special collection	IMAGE	Creation of new object records		Hand-edited	XXX	
DIF TOTAL							XXX
FAA	Saturn Film Collection	VIDEO	Cataloguing enrichment, that means watching all videos again to generate new and verify the existing metadata	Title, Country, Year, Production company, Domicile of the Production company, Runtime, Rights Holder, Genre, Description	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
FAA	Censorship Cards	TEXT	Digital objects from this collection have only partially been catalogued in the past, so a full revision and further indexing of the digital objects was carried out taking into account the database fields listed in the column to the right.	Date of Censorship decision, Certificate number, Agency, Verdict, Regional Scope	Both (Hand-edited and automatically)	XXX	
FAA	Paimann Collection	IMAGE	Digital objects from this collection have only partially been catalogued in the past, so a full revision and further indexing of the digital objects.		Both (Hand-edited and automatically)	XXX	
FAA	Steinwendner Collection	VIDEO	Cataloguing enrichment, that means watching all videos again to generate new and verify the existing metadata	Title, Country, Year, Production company, Director, Writer, Cinematography, Film Editing, Original Music, Domicile of the Production company, Runtime, Rights Holder, Genre, Description.	Hand-edited	XXX	
FAA	Austria Wochenschau	VIDEO	Cataloguing enrichment, which means watching all videos again to generate new and verify the existing metadata.	Title, Date, Runtime, Description, Rights Holder	Hand-edited	XXX	
FAA	News Collection: ÖBUT (Österreich in Bild und Ton)	VIDEO	Cataloguing enrichment, that means watching all videos again to generate new and verify the existing metadata	Title, Date, Runtime Description, Rights Holder.	Hand-edited	XXX	
FAA TOTAL							XXX

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
LCA	Lithuanian documentaries produced by the Lithuanian Film Studio in 1945 – 1960 (part from this collection)	VIDEO	Digital objects from this collection have only partially been catalogued in the past, so a full revision (starting from digitization and viewing of digital images), accumulation and processing of lacking data, and finally description of the digital objects was carried out taking into account the database fields listed in the column to the right. Translating (from Lithuanian into English) of description was performed as well.	Filling the database fields: content (description), document language, credit, and technical data: colour, format, aspect ratio, duration, sound, keyword, runtime.	Hand-edited	XXX	
LCA	The collection of the newsreels „Lithuanian Pioneer“ (part from this collection, produced in 1958-1960)	VIDEO	Digital objects from this collection have only partially been catalogued in the past, so a full revision (starting from digitization and viewing of digital images), accumulation, adding and processing of lacking data, and finally description of the digital objects was carried out taking into account the database fields listed in the column to the right. Translating (from Lithuanian into English) of description was performed as well.	Filling the database fields: content (description), document language, credit, and technical data: colour, format, aspect ratio, duration, sound, keyword, runtime.	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
LCA	Lithuanian feature films collection from the period of 1947-1960		Digital objects from this collection have only partially been catalogued in the past, so a full revision (starting from digitization and viewing of digital images), accumulation, adding and processing of lacking data, and finally description of the digital objects was carried out taking into account the database fields listed in the column to the right. Translating (from Lithuanian into English) of description was performed as well.	Filling the database fields: content (description), document language, credit, and technical data: colour, format, aspect ratio, duration, sound, keyword, runtime.	Hand-edited	XXX	
LCA	The collection of film sketches „Chronicle of German occupation in Lithuania during World War II“ – 41 records		Digital objects from this collection have only partially been catalogued in the past, so a full revision (starting from digitization and viewing of digital images), accumulation, adding and processing of lacking data, and finally description of the digital objects was carried out taking into account the database fields listed in the column to the right. Cataloguing descriptions in English language was performed as well.	Filling the database fields: content (description), document language, credit, and technical data: colour, format, aspect ratio, duration, sound, keyword, runtime.	Hand-edited	XXX	
LCA TOTAL							XXX

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
MNFA	Foto and Poster Dept. – Stills	IMAGE	Digital objects from this collection have only partially been catalogued in the past. In case we know the creator, the place and time of create or we have information (e.g. there was a report when the item got into the collection) it is recorded. There is information we can lay down easily: material, origin, techniques, size, colour, etc. These are all recorded.	ID number, description (language: HUN), title, format, document size, document type, rights holder, creator, colour, URL.	Hand-edited	XXX	
MNFA	Foto and Poster Dept. - Posters	IMAGE	Digital objects from this collection have only partially been catalogued in the past. In case we know the creator, the place and time of create or we have information (e.g. there was a report when the item got into the collection) it is recorded. There is information we can lay down easily: material, origin, techniques, size, colour, etc. These are all recorded.	ID number, description (language: HUN), title, format, document size, document type, rights holder, creator, colour, URL.	Hand-edited	XXX	
MNFA TOTAL							XXX

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
EYE	Kolvenbach	VIDEO	Category (fiction, non-fiction) checked of all records but 16 Unknown, to be checked.	Category (fiction, non-fiction) checked of all records but 16 Unknown, to be checked. The following fields of the collection are partially cleaned and enriched: Title, country, year, creator, company, category. Short description (Dutch). Titles, dates (production, premiere, release, censorship). Cast and crew. Keywords: genres, subjects, geographical, temporal, persons. Physical characteristics from original work.	Hand-edited	XXX	
EYE	Kolvenbach	VIDEO	Make a relation to mpegs and film work. 3300 already done before EFG started. Problematic relations, approximately 300 from start EFG in 2008, still problematic digital works in relation to catalogue, and many have to be checked.	Category (fiction, non-fiction) checked of all records.	Hand-edited	XXX	
EYE	Kolvenbach	VIDEO	In 2010 and 2011: New digital objects with other formats will be made and linked to the film work. New streaming files will be available from another digital source element..	The following fields of the collection are partially cleaned and enriched: Title, country, year, creator, company, and category.	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
EYE	Kolvenbach	VIDEO	Identifying works.	Short description (Dutch).	Hand-edited	XXX	
EYE	Kolvenbach	VIDEO	Clearing rights in progress. Additional cataloguing needed to be done. Public Domain records enriched and cleaned as a priority.	Titles, dates (production, premiere, release, censorship). Cast and crew. Keywords: genres, subjects, geographical, temporal persons. Physical characteristics from original work.	Hand-edited	XXX	
EYE		VIDEO	2010-03-31: all mpegs replaced from the Filmography module into the Copy module	New fields digital object: file name, url, file size, date of creation. (Additional fields to be developed.)	Hand-edited	XXX	
EYE TOTAL							XXX
NNB	*1) Småfilm/Commercials (*1) Due to work on the merger of our two databases, we did not have the time to accomplish the tasks we had planned to do this period. They will therefore be postponed to the next period)	VIDEO	Winter 2011: Constructed titles corrected according to cataloguing rules. Added rights holders: agent + role Added Title type Added more names. Autumn 1010: Added language and language role Autumn 1010: Added language and language role	Title, Title type, Language, Related film title, Related person, Country, Year, Rights, Director, Production, company, Domicile of production, company, Genre/category, Description, Keywords, Colour, Sound, Runtime	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
NNB	Ekofisk - "Films 1900-1935"	VIDEO	Partially catalogued in the past, checked (Winter 2011) the credits on the film and added more persons/agents. Added Rights holder (agent + role) Added more names. Checked titles.	Title, Title type, Language, Related film title, Related person, Country, Year, Rights, Director, Production, company, Domicile of production, company, Genre/category, Description, Keywords, Colour, Sound, Runtime	Hand-edited	XXX	
NNB TOTAL							XXX
TTE	Digitized fiction films	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, file synopsis	Both (Hand-edited and automatically)	XXX	
TTE	Digitized documentaries	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, film synopsis	Both	XXX	
TTE	Digitized newsreels	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, film synopsis	Both	XXX	
TTE	Photos	IMAGE	Cleaning of fields concerning technical aspects, checking relations of photos with film works	Film title	Both	XXX	
TTE	Programs	TEXT	checking relations of programs with film works	Film title	Both	XXX	
TTE TOTAL							XXX
CP	Silent Portuguese non-fiction 1895-1931	VIDEO	New catalogue records of the digital objects 'created' for the project	Manifestation: all fields concerning the digital format	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
CP	Silent Portuguese non-fiction 1895-1931: press clippings and other bibliographic material	TEXT	New catalogue records of the digital objects 'created' for the project	Document size, document type, resolution, colour bit rate, file size, rights holder	Hand-edited	XXX	
CP	Silent Portuguese non-fiction 1895-1931: graphic material (still image)	IMAGE	Revision and addition of data (fields listed in the column to the right) of the documents selected for the project in the local database	Rights holder, geographic scope	Hand-edited	XXX	
CP TOTAL							XXX
NFA	Collection of Czech Documentary Films	VIDEO	checking of complexity of cataloguing records, adding new records to NFA's filmographic database	genre, country, year, runtime, description, keywords, production company, director, director of photography, colour, sound, language	Hand-edited	XXX	
NFA	Collection of Czech Feature Films	VIDEO	checking of complexity of cataloguing records, adding new records to NFA's filmographic database	genre, country, year, runtime, description, keywords, production company, director, director of photography, actors, colour, sound, language	Hand-edited	XXX	
NFA	Collection of Czech Documentary Films	IMAGE	selection of images from film material and description	description	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
NFA	Collection of Czech Feature Films	IMAGE	checking of complexity of cataloguing records for photos	type, size, state of the picture, colour, author of the picture, actors and roles	Hand-edited	XXX	
NFA TOTAL							XXX
CF	Royal Polytechnic	IMAGE	Magic lantern slides	Database Laterna Magica enriched by 83 new slides	Hand-edited	XXX	
CF	Life Models	IMAGE	Magic lantern slides	Database Laterna Magica enriched by 36 new slides	Hand-edited	XXX	
CF	Lapierre	IMAGE	Magic lantern slides		Hand-edited	XXX	
CF	Will Day	TEXT	The cataloguing work is done with the database CINEDOC ouvrages wich include : authority person and authority film, and vocabulary from the thesaurus if it is relevant	Database CINEDOC Ouvrages enriched by 170 new documents	Hand-edited	XXX	
CF	Méliès	IMAGE	The cataloguing work is done with the database CINEDOC photo		Hand-edited	XXX	
CF	Méliès	IMAGE	The cataloguing work is done with the database CINEDOC drawings		Hand-edited	XXX	
CF	Triangle	IMAGE	The data enrichment with names of actors is done with CINEDOC photo		Hand-edited	XXX	
CF	Marey	IMAGE	The data enrichment with names of actors is done with CINEDOC photo		Hand-edited	XXX	
CF	Muybridge	IMAGE	The data enrichment with names of actors is done with CINEDOC photo		Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
CF	Books	TEXT	Digital objects have been catalogued		Hand-edited	XXX	
CF	Photographs, designs and posters	IMAGE	Digital objects have been catalogued		Hand-edited	XXX	
CF TOTAL							XXX
LUCE	11 different Photo collections	IMAGE	XXX photo records already fully catalogued, we exported all the records in XML compliant with the EFG schema				
LUCE	Newsreels fond "La Settimana Incom"	VIDEO	XXX newsreels records already fully catalogued. All the records in XML compliant with the EFG schema exported.				
LUCE	Documentaries	VIDEO	XXX documentaries and short films already fully catalogued. All the records in XML compliant with the EFG schema exported. Spell and consistency checking of all the mandatory EFG fields.				
LUCE	DIAL Photo collection	VIDEO	XXX photo records already fully catalogued. All the records in XML compliant with the EFG schema exported. Spell and consistency checking of all the mandatory EFG fields.				
LUCE TOTAL							XXX

Providing institution	Collection name	Object type	Description of enrichment work	Enriched data fields	How	Enriched records	Totals
KAVA	Finlandia-katsaus (no 1-700)	VIDEO	Enriching and editing the metadata. Hand editing. Resource about ½ persons from 2010 till 2011.	Person names, name types, synopsis.	Hand-edited	XXX	
KAVA TOTAL							XXX
Total all							183.489

2 DATA CLEANING DIGITAL OBJECTS Total overview of replies.

Question 2: Please enter in the table hereunder which kind of cleaning work you carried out for EFG on your digital collections from 1 September 2008 until 31 August 2011. These are the numbers from the beginning of the project until the date you are answering to this evaluation sheet. Data cleaning also includes activities like checking spelling mistakes and merging doublet digital object records.

Providing institution	Collection name	Object type	Description of cleaning work	Cleaned data fields	How	Cleaned records	Totals
CCB (Archive)	Corporations	TEXT	Cleaned and corrected names of the same corporations listed in slightly different ways (examples: "Ambrosio " and "S.A. Ambrosio"; "Pasquali" and "Pasquali e C."; in other cases some Person names of Producers were wrongly listed under Production Company – i.e. Corporations)	Corporations Name	Both (Hand-edited and automatically)	XXX	
	Persons	TEXT	Checking spelling mistakes	Persons	Both	XXX	

Providing institution	Collection name	Object type	Description of cleaning work	Cleaned data fields	How	Cleaned records	Totals
CCB TOTAL							XXX
DFI							
DFI TOTAL							XXX
DIF							
DIF TOTAL							XXX
FAA							
FAA TOTAL							XXX
LCA	Lithuanian documentaries produced by the Lithuanian Film Studio in 1945 – 1960 (part from this collection)	VIDEO	Checking for doublets and spelling mistakes.	Cleaned database fields: title type, genre, country of origin, content (description), document language, credit, technical data: colour, sound, keyword, runtime.	Hand-edited	XXX	
LCA	The collection of the newsreels „Lithuanian Pioneer“ (part from this collection, produced in 1958-1960)	VIDEO	Checking for doublets and spelling mistakes.	Cleaned database fields: title type, genre, country of origin, content (description), document language, credit, technical data: colour, sound, keyword, runtime.	Hand-edited	XXX	
LCA	Lithuanian feature films collection from the period of 1947-1960	VIDEO	Checking for doublets and spelling mistakes.	Cleaned database fields: title type, genre, country of origin, content (description), document language, credit, technical data: colour, sound, keyword, runtime	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of cleaning work	Cleaned data fields	How	Cleaned records	Totals
LCA	The collection of film sketches „Chronicle of German occupation in Lithuania during World War II“ – 41 records	VIDEO	Checking for doublets and spelling mistakes.	Cleaned database fields: title type, genre, country of origin, content (description), document language, credit, technical data: colour, sound, keyword, runtime.	Hand-edited	XXX	
LCA TOTAL							XXX
MNFA							
MNFA TOTAL							XXX
EYE	Kolvenbach	VIDEO	Mainly checking fields on wrong accessed data: Some examples: Unjust text in suffixes like (Van Gasteren). Missing initials when surname is known. A lot of data on the digital objects is not ingested in our current database.	Suffixes Connecting urls from Filmotech to the digital objects. Checking links and data in Diva manually.	Both (Hand-edited and automatically)	XXX	
EYE TOTAL							XXX
NNB	Småfilm	VIDEO	Cleaning of doublets and misspellings in our local database	Title, Title type, Language, Related film title, Related person (name + role), Year, Rights, Director, Production, company, Domicile of production, company, Genre/category, Description, Keywords, Colour, Sound, Runtime	Hand-edited	XXX	

Providing institution	Collection name	Object type	Description of cleaning work	Cleaned data fields	How	Cleaned records	Totals
NNB	Ekofisk	VIDEO	Cleaning of doublets and misspellings in our local database	Title, Title type, Language, Related film title, Related person, Country, Year, Rights, Director, Production, company, Domicile of production, company, Genre/category, Description, Keywords, Colour, Sound, Runtime	Hand-edited	XXX	
NNB TOTAL							XXX
TTE	Digitized fiction films	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, fil synopsis	Both (Hand-edited and automatically)	XXX	
TTE	Digitized documentaries	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, film synopsis	Both (Hand-edited and automatically)	XXX	
TTE	Digitized newsreels	VIDEO	Cleaning of fields concerning technical data in translations	Colour, aspect ratio, format, sound, film synopsis	Both (Hand-edited and automatically)	XXX	
TTE	Photos	IMAGE	Cleaning of fields concerning technical aspects, checking relations of photos with film works	Film title	Both (Hand-edited and automatically)	XXX	
TTE	Programs	IMAGE	checking relations of programs with film works	Film title	Both (Hand-edited and automatically)	XXX	
TTE TOTAL							XXX
CP	Silent Portuguese non-fiction 1895-1931	VIDEO	Addition of the CP logo (renewed) in the digital objects		Both (Hand-edited and automatically)	XXX	
CP TOTAL							XXX

Providing institution	Collection name	Object type	Description of cleaning work	Cleaned data fields	How	Cleaned records	Totals
NFA TOTAL							XXX
CF							
CF TOTAL							XXX
LUCE	11 different Photo collections	IMAGE	200.000 photo records already fully catalogued, we exported all the records in XML compliant with the EFG schema	We've done a spell and consistency checking of all the mandatory EFG fields	Both (Hand-edited and automatically)	XXX	
LUCE	Newsreels fond "La Settimana Incom"	VIDEO	13.838 newsreels records already fully catalogued. All the records in XML compliant with the EFG schema exported.	We've done a spell and consistency checking of all the mandatory EFG fields	Both (Hand-edited and automatically)	XXX	
LUCE	Documentaries	VIDEO	4.400 documentaries and short films already fully catalogued. All the records in XML compliant with the EFG schema exported. Spell and consistency checking of all the mandatory EFG fields.	We've done a spell and consistency checking of all the mandatory EFG fields	Both (Hand-edited and automatically)	XXX	
	DIAL Photo collection	VIDEO	30.000 photo records already fully catalogued. All the records in XML compliant with the EFG schema exported. Spell and consistency checking of all the mandatory EFG fields.	We've done a spell and consistency checking of all the mandatory EFG fields	Both (Hand-edited and automatically)	XXX	
LUCE TOTAL							XXX
KAVA	Film work and person	VIDEO	Checking and correcting names and synopsis. Resource about 1 person from 2010 till 2011	Type of activity, person name, rights holder	Hand-edited	XXX	
KAVA TOTAL							XXX
Total all							257.492

3 AUTHORITY RECORD ENRICHMENT. Total overview of replies.

Question 3: Please enter in the table hereunder which kind of enrichment work you carried out for EFG on your authority data until 1 September 2008 until 31 August 2011. These are the numbers from the beginning of the project until the date you are answering to this evaluation sheet. Please list your indications as precisely as possible (e.g. to break down the numbers for Film Works into genres or categories). If you enriched further authority data in the last evaluation period please feel free to extend the table. Please use one row for each record type and enter only one type (either “Person”, “Film work”, “Corporation” or “Event”) in the column “Kind of authority data”.

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
CCB (Archive)	Person/Film Work	Documentaries	Adding new authority records (30 Film Works and 22 Persons) in CCB AV db. Authority records concerning Corporations were already included in the CCB AV db.	All according fields to describe works reported in CCB validation List	Both (Hand-edited and automatically)	XXX	
CCB (Archive)	Person/Film Work/Corporation	Documentaries, Short Films	Adding new authority records (10 Film Works, 12 Persons and 3 Corporations) in CCB AV db.	All according fields to describe works reported in CCB validation List	Both (Hand-edited and automatically)	XXX	
CCB (Archive)	Person/Film Work/Corporation	Short Films (20), Documentaries (10), Interviews (10)		All according fields to describe works reported in CCB validation List	N/A	XXX	
CCB (Library)	Person/Film Work/Corporation	Photos	The 3 authority files have been revised following the main db of films	Director, Film work, Corporation	N/A	XXX	

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
CCB (Library)	Person/Film Work/Corporation	Documentaries	Adding new authority records (30 Film Works and 22 Persons) in CCB AV db. Authority records concerning Corporation were already included in the CCB AV db.	All according fields to describe works reported in CCB validation List	Both (Hand-edited and automatically)	XXX	
CCB (Library)	Person/Film Work/Corporation	Short Films	Adding new authority records (10 Film Works, 12 Persons and 3 Corporations) in CCB AV db.	All according fields to describe works reported in CCB validation List	Both (Hand-edited and automatically)	XXX	
CCB (Library)	Person/Film Work/Corporation	Photos collection	The 3 authority files have been revised following the main db of films	All according fields to describe works reported in CCB validation List	Hand-edited	XXX	
CCB (Library)	Person/Film Work/Corporation	Posters collection	The 3 authority files have been revised following the main db of films	Director, Film work, Corporation	N/A	XXX	
CCB (Library)	Person/Film Work/Corporation	Censorship visa documents / Visa Cards	An important revision was made concerning the fields "production" and "distribution", but also "year"	All according fields to describe works reported in CCB validation List	Hand-edited	XXX	

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
CCB (Library)	Corporation	Censorship visa documents / Visa Cards	Cleaned and corrected names of the same corporations listed in slightly different ways (examples: "Ambrosio" and "S.A. Ambrosio"; "Pasqual" and "Pasquali e C."; in other cases some Person names of Producers were wrongly listed under Production Company – i.e. Corporations)		N/A	XXX	
CCB Total							XXX
DFI	Film Work	200 Danish films from 1992-2002	English synopsis on 200 Danish films from 1992-2002, taken from paper printed material.	Synopsis engelsk (Synopsis english)	Hand-edited	XXX	
DFI	Film Work	788 Danish films from 1954-2002	Danish synopsis on 788 Danish films from 1954-2002	Synopsis	Hand-edited	XXX	
DFI	Film Work	200 titles catalogued	Retro cataloguing of Danish and Foreign films from card catalogue.	All relevant fields	Hand-edited	XXX	
DFI	Film Work	500 documentaries	Retro subject indexing of keywords to Danish documentaries	Emneord (Keyword)	Hand-edited	XXX	
DFI	Person	Persons enriched	Description of the persons' activities. Actor, director, etc. All names are enriched.	Funktion (Function)	Automatically	XXX	

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
DFI	Person	Nationalfilmografien	We have added gender/sex for persons. We have used an official namelist from 'Familiestyrelsen' which indicates whether a given person name is female or male. Our external database developer have then paired this list with the names in our database and created a new field in the backend of our database, so we now have the possibility to include gender in our XML for EFG.	"Køn" = Gender (Photographer)	Automatically	XXX	
DFI	Person		Credits to Danish posters	Ophav Name: "Navn"	Hand-edited	XXX	
DFI	Person		Credits to Danish stills	Still photographer: "Ophav"	Hand-edited	XXX	
DFI Total							XXX
DIF	Film Work	Films produced in Germany and co-productions	Adding new authority records to DIF's filmographic database filmportal-zdb.	All according fields to describe film works in the filmportal-zdb database.	Hand-edited	XXX	
DIF	Person		As above	As above	Hand-edited	XXX	
DIF	Corporation		As above	As above	Hand-edited	XXX	
DIF	Event		As above	As above	Hand-edited	XXX	
DIF Total							XXX

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
FAA			No authority data in the archive and no cataloguing work concerning this topic.		N/A	XXX	
FAA Total							XXX
LCA	Person		Revised, corrected and unified orthography of person names in order to prepare a proper display of names in the EFG Portal.		N/A	XXX	
LCA Total							XXX
MNFA	Film Work		Adding new records to MNFA's filmographtis database.	All according fields to describe film works in the in-house database	Hand-edited	XXX	
MNFA Total							XXX
EYE	Person	Kolvenbach	New authority records in Diva. Cleaning data. Enriching data. Cleaned Persons who were stored in the table Corporations. In Excel, actual change during conversions to the new system.	Partially enriched: full name, parts of name, biography (biographies in Dutch), dates of birth and death, relation work and function.	Hand-edited	XXX	
EYE	Corporation	Kolvenbach	New authority records in Diva. Cleaning data. Enriching data. Cleaned Persons who were stored in the table Corporations.		Hand-edited	XXX	

EYE	Film Work	Kolvenbach	<p>It appears during cleaning and enriching that parts of Kolvenbach are not registered in a way the records can be presented. For instance news items and copies with different content are clustered in one catalogue record. Languages: Five fields concerning languages (intertitles, spoken, commentary, dubbed, subtitling) cleaned. Valuelist was and is present but in system it is possible to place more than one choice in the same field. Actual change during conversions. ISO-codes/terms 369 are used. Additional notes: Total records: 3579, 29-3-2010. Short descriptions: 1483 according rules. Genres: 3202. Subject keywords: 3267. Fiction film (1351) with cast: 951. Special attention to corpora: Dutch East-Indies, Desmet, Dutch directors, etc.</p> <p>These corpora are richer than the metadata standard for EFG. Extra research for data, missing or unsure or wrong), historic background, links with persons, other films etc.</p>	<p>All fields from the EFG schema that can be stored in our current catalogue. Short description: 614 records in place with a short description but approximately 314 have to be checked. Approximately 300 short descriptions according recently set up rules for a short description.</p>	Hand-edited	XXX	
-----	-----------	------------	--	---	-------------	-----	--

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
EYE		Kolvenbach - Geography	Added planets/codes according to MARC, Place of birth and death are now free text fields. We attached the places to the geography table. Actual change in new cataloguing system during conversions.	Automatically: changing internal tables to match geographical names to value list, correcting tables between entities. Technical work.	Both (Hand-edited and automatically)	XXX	
EYE Total							XXX
NNB	Person/Film Work/Corporation	Småfilm/Commercials	Merges doublets name authority records. (i.e. titles).	OrganisationName biographyPublic summary synonyms><name firstName lastName dates nationalities gender synonyms>name> <synonyms><firstName> biographyPublic	Hand-edited	XXX	

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
NNB	Person/Film Work/Corporation	Ekofisk/industrial films	Merges doublets name authority records. (i.e. titles).	OrganisationName biographyPublic summary synonyms><name firstName lastName dates nationalities gender synonyms>name> <synonyms><firstName> biographyPublic	Hand-edited	XXX	
NNB	Person/Film Work/Corporation	Miscellaneous + Filmography/long fiction	Already catalogued in the past / The database does not have an event entity.		N/A	XXX	
NNB	Film Work	Feature films	Cataloguing and adding new authority records to NFIs Mavis database, feature films for the Norwegian filmography.	All according fields to describe Norwegian feature film works in the NFI Mavis database. (total number of titles including what has been catalogued in the past, before the project started, = 771)	N/A	XXX	
NNB Total							XXX
TTE	Person	All collections	Added data to translations	Name, Type, Gender	Hand-edited	XXX	
TTE	Corporation	All collections	Added data to translations	Name, Type	Hand-edited	XXX	
TTE Total							XXX

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
CP	Film Work	Silent non-fiction films	Addition of data according to the fields necessary for EFG	Genre / category	Hand-edited	XXX	
CP	Person	Silent non-fiction films	Addition of new authority records in the local database	Date of birth and death, place of birth and death, function	Both (Hand-edited and automatically)	XXX	
CP	Corporation	Silent non-fiction films	Addition of new authority records in the local database	Date of birth and death, place of birth and death, function	Both (Hand-edited and automatically)	XXX	
CP Total							XXX
NFA	Film Work	Collection of Czech Documentary films	Adding new authority records to NFA's filmographic database.	Full revision and further indexing of documentary films which were screened within working of Cataloguing Commission.	N/A	XXX	
NFA	Film Work	Czech Documentary Films selected from period 1898-1928	checking of complexity of cataloguing records, adding new records to NFA's filmographic database	genre, country, year, runtime, description, keywords, production company, director, director of photography, colour, sound, language	Hand-edited	XXX	
NFA	Film Work	Czech Feature Films selected from period 1911-1920	checking of complexity of cataloguing records, adding new records to NFA's filmographic database	genre, country, year, runtime, description, keywords, production company, director, director of photography, actors, colour, sound, language	Hand-edited	XXX	

Providing institution	Kind of data	Data set	Description of enrichment work	Enriched fields	How	Enriched records	Totals
NFA	Film Work	Czech Documentary Films selected from period 1898-1928	selection of images from film material and description	description	Hand-edited	XXX	
NFA	Film Work	Czech Feature Films	checking of complexity of cataloguing records for photos	type, size, state of the picture, colour, author of the picture, actors and roles	Hand-edited	XXX	
NFA Total							XXX
CF							
CF Total							XXX
LUCE							
LUCE Total							XXX
Total all							258.228

4 DATA CLEANING AUTHORITY DATA. Total overview of replies

Question 4: Please enter in the table hereunder which kind of enrichment work you carried out for EFG on your authority data until 1 September 2008 until 31 August 2011. These are the numbers from the beginning of the project until the date you are answering to this evaluation sheet. Please list your indications as precisely as possible (e.g. to break down the numbers for Film Works into genres or categories). Please use one row for each record type and enter only one type (either “Person”, “Film work”, “Corporation” or “Event”) in the column “Kind of authority data”.

Note: Activities which were supported by the EFG Authority File Manager Tool are **highlighted in blue**.

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
CCB (Archive)	Corporation	Cleaned and corrected names of the same corporations listed in slightly different ways (example: "Corona Cin.ca" and "Corona Cinematografica"; or in other cases some Person names of Producers were wrongly listed under Production Company – i.e. Corporations))	Corporations Name	N/A	XXX	
CCB (Library)	Film Work	We cleaned mainly the original titles	Angelo Novi Photos Collection	N/A	XXX	
CCB (Library)	Corporation	We cleaned mainly original titles and production company	Poster collection	N/A	XXX	
CCB (Library)	Person	We also cleaned information about the Persons (wrong names, etc)	Censorship visas db	N/A	XXX	
CCB (Library)	Corporation	Cleaned and corrected names of the same corporations listed in slightly different ways (examples: "Ambrosio" and "S.A. Ambrosio"; "Pasquali" and "Pasquali e C."; in other cases some Person names of Producers were wrongly listed under Production Company – i.e. Corporations)	Corporations Name / 50 AV 1173 Non Av	Both (Hand-edited and automatically)	XXX	
CCB (Library)	Person	Checking spelling mistakes	Persons / 20 AV 1173 Non AV	Both (Hand-edited and automatically)	XXX	
CCB Total						XXX

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
DFI	Person	For our cleaning work of person names for EFG we have verified possible doublets by looking at our website under Person where our person names are sorted alphabetically: http://www.dfi.dk/FaktaOmFilm/ Nationalfilmografien/ NFAIphPerson.aspx?type=person and we have used the Authority File Manager. We have located different ways of spellings of person names and we have been checking around 80.000 person names and merged around 4000 person names. We did start to use the extractions of doublet lists from the EFG authority file manager which we have used to detect possible doublets.	Person name	Hand-edited	XXX	
DFI	Person	Cleaning and enrichment of Famous Danish directors and actors' "died" and "born".	Født dato år (Born - date - year) + Død - dato - år (Born date year)	Hand-edited	XXX	
DFI	Film work	Cleaning of doublets, categories		Hand-edited	XXX	

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
DFI	Corporation	For our cleaning work of corporations for EFG we have verified possible doublets by looking at our website under Selskaber (Coporations) where the corporations are sorted alphabetically: http://www.dfi.dk/FaktaOmFilm/Nationalfilmografien/NFAlphSelskab.aspx?type=company We have here located different ways of spellings of the corporations. So we have here been checking around 500 corporations and merged around 100 corporations.	Credit - Produktions-selskaber (Production companies)	Hand-edited	XXX	
DFI Total						XXX
DIF	Person	Entries in the "type of activity" field were not controlled by a vocabulary in the past. So the field entries were heterogeneous and not spelled correctly in all cases. This field has been cleaned in a semi-automatic process. The field was parsed and all activity types were extracted into a separate list. In this list, multiple entries were separated into single ones by splitting them after the comma. For each activity type the number of its occurrence in the field	Type of activity	Both	XXX	

		<p>was indicated. All entries with a high number of occurrences became automatically terms of the controlled vocabulary "Activity types for persons". All other entries came into the pool of uncontrolled activity types. The uncontrolled terms were displayed in the user interface of the filmportal database where they were linked to the person record in which the according entry existed. A cataloguer has checked the list of uncontrolled terms and corrected the respective person records connected to these. Within this process further activity types were defined for the controlled vocabulary. Special activities such as "Chancellor of Germany 1998 - 2005" were matched to broader terms (here: "Politician"). The specialization was copied into the field "Bibliographical note". The type of activity is still a free-text field but cataloguers have to apply the newly established controlled vocabulary for this field. A special tool was developed to clean the activity field.</p>				
DIF	Film work	Cleaning date and time indications which have not been entered according to the syntactical rules for this field.	Several fields for dates and times	Both (Hand-edited and automatically)	XXX	

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
DIF	Film work	Identifying doublet records of film works and merging these to single records. The EFG Authority File Manager was used to detect possible doublets within DIF's film work contribution to EFG. The possible duplicates were exported in an excel list. A filmographer checked whether the record pairs in the list were real doublets or not (doublet, no doublet, unknown – treated as no doublet). The result was: 99 identical, 270 not identical, 42 unknown record pairs. Hence, in total 411 record pairs were checked (in total 822 records).	Film title	Both (Hand-edited and automatically)	XXX	
DIF	Person	Checking and merging of doublet person records. The EFG Authority File Manager was used for this task. The possible duplicates were exported in an excel list. A filmographer checked whether the record pairs in the list were real doublets or not (merged – doublet, ignored – no doublet, unknown – treated as no doublet, remarks). During the process the correct spelling of names was checked, too. On a relation level of appr. 91,1% the plausibility of finding real	Persons	Both (Hand-edited and automatically)	XXX	

		doublets was very low. Some of the possible duplicates were just alternative names. The result was: 66 identical, 1167 not identical, 67 unknown record pairs. Hence, in total 1300 record pairs were checked.				
DIF Total						XXX
FAA						
FAA Total						XXX
LCA	Person	Revised, corrected and unified orthography of person names in order to prepare a proper display of names in the EFG Portal. Checking whether there are doublets, checking and correcting all spellings mistakes in person names.	Person Name	N/A	XXX	
LCA Total						XXX
MNFA	Film Work	Cleaned and corrected the wrong file names.	Film Title	Hand-edited	XXX	
MNFA Total						XXX
EYE	Person	Sex: was imported from former database as Male. Already cleaned. Now still 41 Unknown (part really Unknown or Animals but final check takes place) Still sexes are changed because we only cleaned globally and not individual records.	Sex	Both (Hand-edited and automatically)	XXX	

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
EYE	Film Work	Comparing physical characteristics of the digital object, the source of the digital object and the characteristics of the original film work. Physical characteristics in the filmography will be sent to EFG. Runtime is partly available in another database, to be transferred. Changing multiple filmographies for one film to one filmography. Still in progress.	Year, colour, sound, runtime	Hand-edited	XXX	
EYE	Corporation	Checked metadata of several corporations. Number can't be extracted from database.	Particular name (and subfields of the name) - Several cleaning rounds of corporations of Dutch films, esp. early films. For a considerable part belonging to Kolvenbach.	Hand-edited	XXX	
EYE Total						XXX
NNB	Person/Corporation	Misspelling and double recordings in the NFI-Mavis database: Miscellaneous + Filmography (Long fiction). The EFG Authority File Manager was used to support this task.	Navn / Names See comments regarding filmography	Hand-edited	XXX	

Providing institution	Kind of data	Description of cleaning work	Names of cleaned fields	How	Cleaned records	Totals
NNB	Person/Corporation	Cleaning in connection with the Mavis database merger: misspelling and doublets	See comments regarding Mavis merger	Hand-edited	XXX	
NNB Total						XXX
TTE	Person	Cleaned double entries and spelling mistakes	Name	Hand-edited	XXX	
TTE	Corporation	Cleaned double entries and spelling mistakes	Name	Hand-edited	XXX	
TTE Total						XXX
CP	Film work	Checking data for spelling mistakes + Merging doublets	Title, genre/category	Both (Hand-edited and automatically)	XXX	
CP	Person	Merging doublets, correction of spelling mistakes	Person name	Hand-edited	XXX	
CP	Corporation	Correction of spelling mistakes + Checking data for spelling mistakes	Name	Hand-edited	XXX	
CP Total						XXX
NFA						
NFA Total						XXX
CF						
CF Total						XXX
LUCE						
LUCE Total						XXX
TOTAL						48.804

5 DIGITAL OBJECTS related with AUTHORITY RECORDS. Total overview of replies.

Question 5.1: Please make indications on how many objects you connected with film titles and/or names from 1 September 2008 until 31 March 2011. If your archive delivers separate XML exports for digital objects, for film works, for persons and for corporate bodies to EFG please do the following: Count the number of object records that were connected to authority files *and* the number of authority files that were connected to the digital object records.

Providing institution	Digital objects connected to persons	Digital objects connected to film works	Digital objects connected to corporate bodies	Persons connected to digital objects	Film works connected to digital objects	Corporate bodies connected to digital objects
CCB	XXX	XXX	XXX	XXX	XXX	XXX
DFI	XXX	XXX	XXX	XXX	XXX	XXX
DIF	XXX	XXX	XXX	XXX	XXX	XXX
MNFA	XXX	XXX	XXX	XXX	XXX	XXX
EYE	XXX	XXX	XXX	XXX	XXX	XXX
NNB	XXX	XXX	XXX	XXX	XXX	XXX
TTE	XXX	XXX	XXX	XXX	XXX	XXX
CF	XXX	XXX	XXX	XXX	XXX	XXX
TOTAL	68213	136107	28321	47083	75133	8481

6 DIGITAL OBJECT RECORDS enriched with PERSON NAMES and FILM TITLES: Total overview of replies.

Question 5.2: Please make indications on how many objects you connected with film titles and/or names from 1 September 2008 until 31 August 2011. If your archive delivers the digital object record together with the person name(s), film title(s) or corporation name(s) in one XML export please do the following: Count the digital object records in which you inserted a film title *and* the number of digital object records in which you inserted names.

Content provider	Records in which you inserted person names	Records in which you inserted film titles	Records in which you inserted corporate body names
LCA	XXX	XXX	XXX
MNFA	XXX	XXX	XXX
NNB	XXX	XXX	XXX
CP	XXX	XXX	XXX

TOTAL	2127	2232	1937
--------------	-------------	-------------	-------------