

## EFG's data cleaning and enrichment tools

**Europeana v1.0 WP3 meeting**  
*Vienna, 29 March 2011*

Francesca Schulze  
Deutsches Filminstitut -DIF  
[f.schulze@deutsches-filminstitut.de](mailto:f.schulze@deutsches-filminstitut.de)



## Project aims

- digital „showcase“ for collections of Europe ´s film archives and cinemathèques
  - Films and film clips, trailer (mostly: non-fiction)
  - Images: stills, posters, photos, set designs, drawings
  - Text material: correspondences, reviews, censorship documents
- make collections accessible through a common access point
- link Europe`s film institutions with *Europeana*
- 3 years project (ends in August 2011)
- portal launch expected in May, URL:  
[www.europeanfilmgateway.eu](http://www.europeanfilmgateway.eu)
- In August 2011: ca. 600.000 Items in *Europeana*



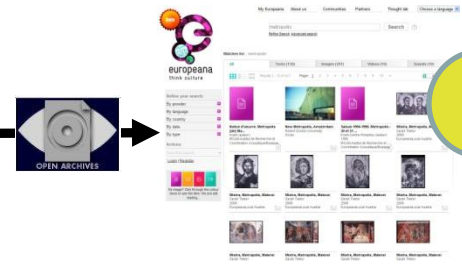
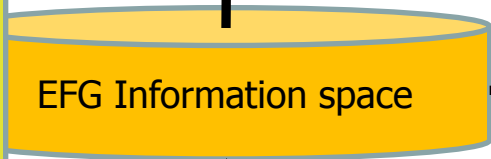
# Infrastructure setup

Aggregation of filmographic, bibliographic and data on archival resources

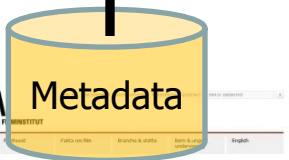
2<sup>nd</sup> idea of EFG: common registry of film titles and persons active in the film domain:  
120.000 film works  
230.000 persons  
25.000 corporations  
+ event records



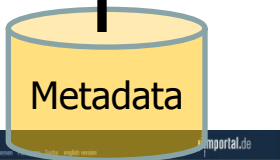
Access to information about digital objects, related names and titles (11 out of 16 Providers, 375.000 Items)



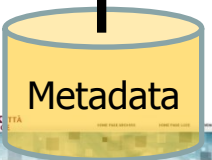
340.000 Items in Europeana



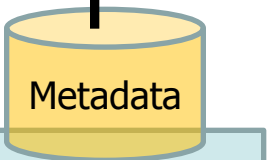
Danish Filmography (DFI)



Filmportal.de (DIF)



Luce site



Ca. 13 more sites (16 Providers)





# Challenges

- harmonizing heterogenous and multilingual data from 16 film archives
- different quality of metadata
- no common standards for metadata schemas and vocabularies in film archival domain
- information about the same person or film work can come from different data sources due to numerous European co-productions
- complex metadata schema based on FRBR oriented Cinematographic Works Standard
  - [www.filmstandards.org](http://www.filmstandards.org)



## Authority File Manager

- Data cleaning
- Tool to identify doublet records of the same film work or person (active in the film domain) in EFG Information Space
- allows partners also to detect doublets within their own databases
- merge doublet records to authority records
- uses basic concepts of „Match & Merge“ by librarians



# Authority File Manager

EFG - Authority File Management Welcome franca.debole | E

Persons

[Back to Dashboard](#) [Help](#)  
[merged \(2\)](#) [ignored \(1\)](#)



Duplicates: 14715

Pages: << < [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >>

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 Meinecke	portal DIF	Eva-Maria Meinecke	portal DIF
99.1%	J. J. Johnson	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 Kirchner	portal DIF	Hans-Peter Kirchner	portal DIF
99.0%	William Stranz	portal DIF	William von Strantz	portal DIF
99.0%	Theresa Scholz	portal DIF	Theresa-Sophie Scholz	portal DIF
99.0%	Steven Posters	portal DFI	Steven B. Poster	portal DFI
99.0%	Heinz Konrad	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
99.0%	George Connors	portal EYE	Bad George Connor	portal DFI
99.0%	K.C. Colwell	portal DFI	K. C. Colwel	portal DFI



# Authority File Manager

## EFG - Authority File Management

### Person Record Merge

[Back to detail page](#)

[Help](#)

Choose which record to keep and select which version of the field value you want to preserve:

First name Eva Maria

First name Eva-Maria

Last name Meineke

Last name Meinecke

Date 1923-10-08

Date

Type of activity Actor

Type of activity Actor

Sex f

Sex f

Place of birth Berlin

Place of birth

Provider Deutsches Filminstitut - DIF

Provider Deutsches Filminstitut - DIF

Identifier: person.filmportal.de/DIF\_person\_7E2201C16:

Identifier: person.filmportal.de/DIF\_person\_F909E564D

[Switch to this record](#)

[Merge preview](#)

Version 0.9.4

This software is heavily under development, stay tuned.

Tested only with Firefox >= 3.5 and Google Chrome >=5





## Metadata Editor

- Data enrichment
- Tool allows partners to improve the quality of their metadata records
- encouraged archives to enrich their data locally with EFG relevant information
- most important enrichments: descriptions for non-audiovisual digital objects (also in English), establish relations to persons depicted on an image, creator, et. al.
- individual cataloguing plans





# Metadata Editor

Home NewFilm

Upload

Create

New Film

New Person

New Corporation

New Digital Object

Personal Folder

Editing

Help

franca.debole  
Efg Tools | Logout

## New Film

The fields highlighted on red are mandatory. Please fill them.

### Work

- Broadcast commentary
- Content description
- Dialogue
- Intertitles
- Review snippet
- Shotlist
- Synopsis
- n/a

### Manifestation



## Vocabulary Checker

- Data cleaning
- Tool used to verify vocabulary matchings
- matching intellectual work by partner archives
- 46 EFG mini-vocabularies (value lists for attributes & semantic relationship types)
- FIAF glossary of filmographic terms, ISO country and language codes, Marc relator codes, et al.
- translated into 13 European languages
- available on EFG project website:  
[www.europeanfilmgateway.eu/guidelines\\_and\\_standards.php](http://www.europeanfilmgateway.eu/guidelines_and_standards.php)



# Vocabulary Checker

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_2B4B82FFFFF44B66DBD3FB5178DFA1D5D </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.



# Open workshop on data quality and semantic interoperability issues in European film archives

30 May in Frankfurt

Thank you!

[f.schulze@deutsches-filminstitut.de](mailto:f.schulze@deutsches-filminstitut.de)

[www.europeanfilmgateway.eu](http://www.europeanfilmgateway.eu)

