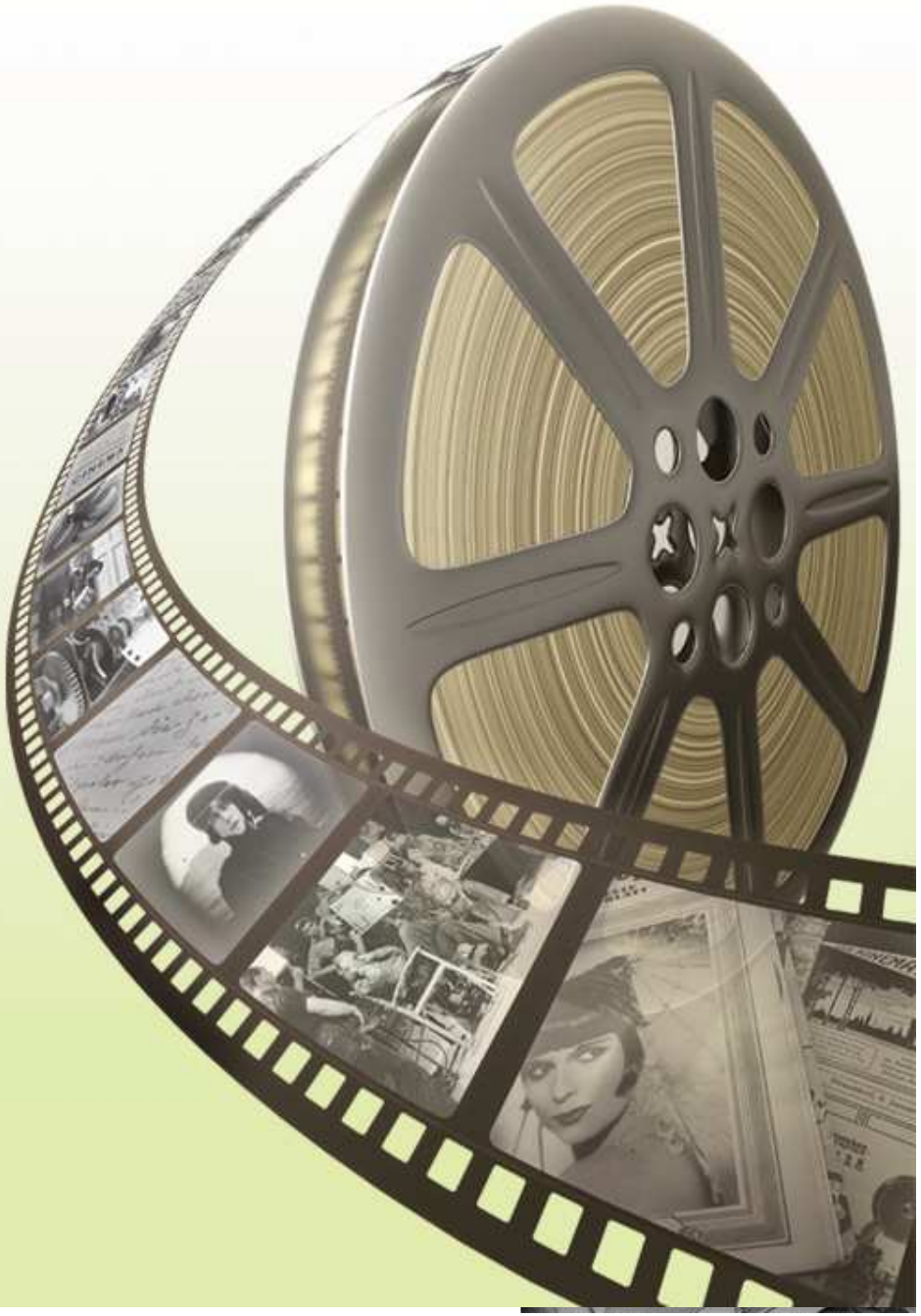
EFG
european film gateway

EFG´s data cleaning and enrichment work



CEN/TC 372 Workshop
Copenhagen, 15 April 2011

Francesca Schulze
Deutsches Filminstitut - DIF
f.schulze@deutsches-filminstitut.de

Pernille Schütz
Det Danske Filminstitut – DFI
pernilles@dfi.dk

Co-funded by the Community Programme eContent*plus*

EFG
european film gateway

Overview of presentation

- Challenges EFG Work Package 3
- Cataloguing Work in Film Archives
- EFG Vocabularies and Matchings
- EFG Backend Tools
- Achievements, Visions, Next Steps

EFG
european film gateway

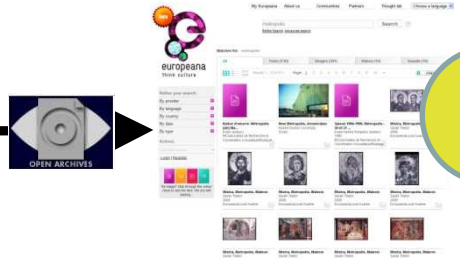
Aggregation aims

Reduced set of EFG metadata:
Access to information about digital items, related names and titles (11 out of 16 Providers, 375.000 Items)

EFG format
(> 900.000 XML records):

Aggregation of filmographic and biographic information, data on digital archival resources

2nd idea of EFG: common registry of film works and persons active in the film domain (120.000 film works, 230.000 persons, 25.000 corporations, + event records)



ESE format:
340.000 Items in Europeana

Metadata

Metadata

Metadata

Metadata

Ca. 13 more sites (16 Providers)

Danish Filmography (DFI)

Filmportal.de (DIF)

Luce site

EFG european film gateway

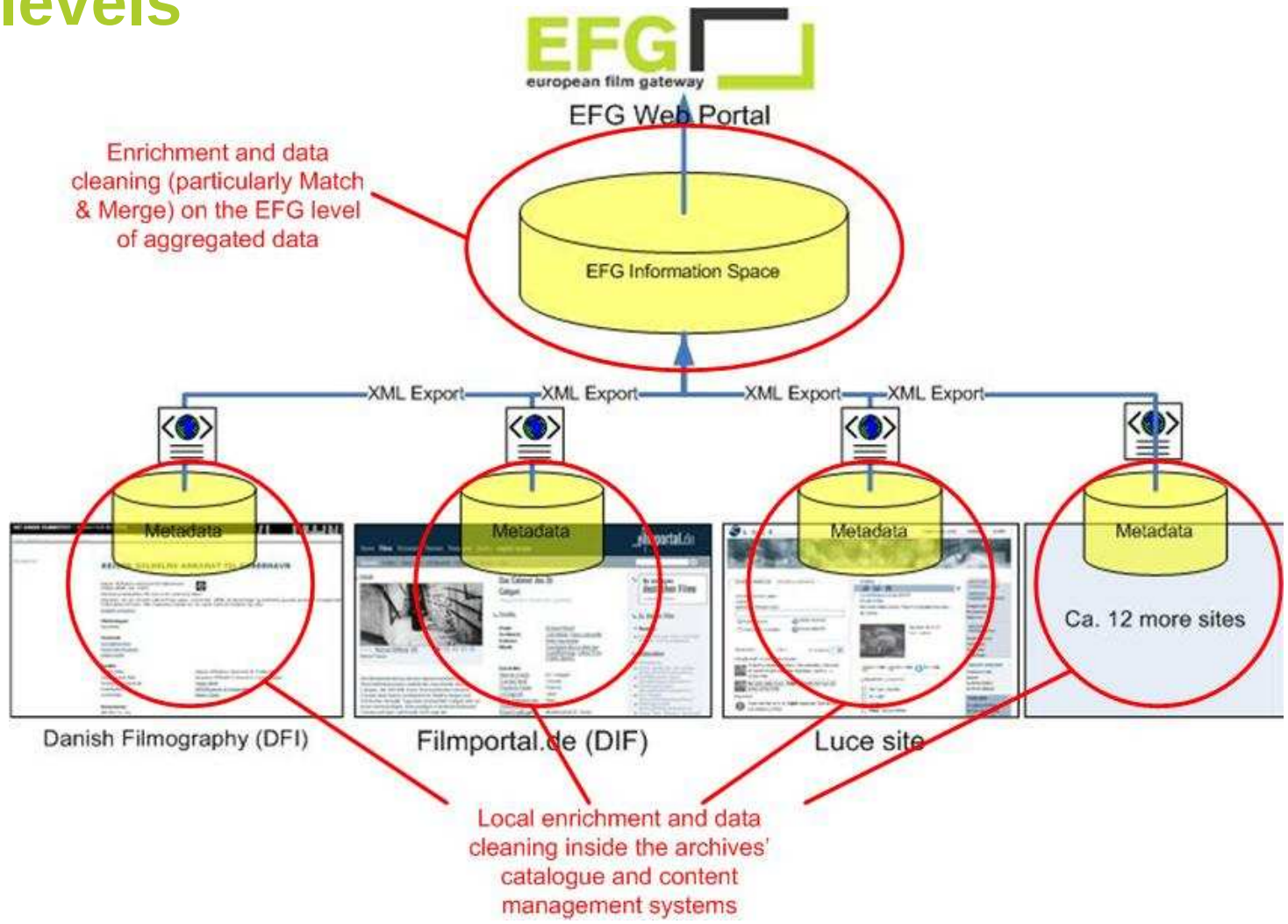
Challenges

- harmonizing heterogenous and multilingual metadata (values) from 16 film archives
- different quality of metadata
- no common standards for metadata schemas, cataloguing rules and vocabularies applied 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.europeanfilmgateway.eu/guidelines_and_standards.php

Data cleaning and enrichment on two levels



Data cleaning and enrichment in local databases

- Survey on cataloguing rules applied in film archives
  - ➢ No common cataloguing rules at EFG archives
- Provided archives with general guidelines on how to enrich and clean their data for EFG needs
- Collected evaluation sheets from archives on cataloguing work done
- Individual cataloguing plan for each archive

Data cleaning and enrichment in local databases

- ✓ 385.000 enriched and cleaned records
  - Addition of metadata to existing records
  - Creation of new records
  - Harmonization and standardization of records
- ✓ 275.000 established relationships
  - Local ID of the digital object record was embedded into the local authority record.
  - …and vice versa.

Data cleaning on EFG level:
Controlled vocabularies and matchings

45 mini-vocabularies:

- harmonize and contextualize metadata in EFG
- display values coherently in EFG and Europeana portal
  - ✓ vocabularies established for end users (public)
- express archival data consistently in a reference model
  - ✓ EFG schema: ER model with subject predicate object triples
- compiled value lists for attributes & semantic relationships
- assigned vocabularies with elements of EFG schema (textual documentation)
- translated into 13 European languages
- public version available on EFG project website:
  - ✓ www.europeanfilmgateway.eu/guidelines_and_standards.php

EFG european film gateway

Controlled vocabularies in EFG

What do they cover?

- Formal aspects to describe a film work
  - ✓ form/category, format, provider name, etc.
- Formal aspects to describe a non-film work
  - ✓ document type, format, provider, etc.
- Types for attributes of different entities
  - ✓ date & name types, title types, activitiy types, etc.
- Spatial coverage
  - ✓ countries, regions
- Semantic relationship types
  - ✓ cast & credits, etc.

➢ no harmonization of subject terms

Controlled vocabularies in EFG

How were they compiled?

- FIAF Glossary of Filmographic Terms, 2008
- ISO Country Codes 3166-1
- AFNOR
- Marc Geographic Area Codes
- ISO Language Codes 639-1, 639-2
- Marc Relator Codes
- EBU P/META 2.0 Concept Schemes
- EAC Beta
- IANA MIME Media Types
- ➢ Value lists of EFG partner archives

EFG european film gateway

EFG vocabularies:

38 value lists with 1.700 term concepts

Example: "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

ocabulary matching:

16 tables with 11.000 matched source values

Example: "Colour"

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

EFG european film gateway

Vocabulary matching

Example: Terms matched to "Assistant camera operator"

1. 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

Heterogenous source values

- ✓ 60 % terms from free-text fields
- ✓ 40 % terms from controlled vocabularies (diverse vocabulary lists at archives)

EFG european film gateway

EFG vocabularies:

7 lists with 210 semantic relationship concepts

Example: "Director"

EFG Term	Vocabulary Name	Domain	Display Term	Display Term Inverse	Range	Lan-guage	Gen-der
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	Rendezı	Rendezı	Person	hu	U
Director	AgentRelation	AVCreation	Regissør	Regissør	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

EFG european film gateway

Contextualisation through semantic relations

Film work

M *(Original title)*

Director

Release version

Has image / Is image of

Person

Fritz Lang *(Preferred name)*

Manifestation

M. Dein Mörder sieht Dich an *(Distribution title)*

Film still

Has image / Is image of

Has text / Is text of

Depicts / Is depicted in

Portrait

Censorship card

Prüf-Nr. 28843

Antragsteller: Nero-Film Aktiengesellschaft, Berlin W 8 Unter den Linden 21

Hersteller:

Haupttitel: „M."

Die Nero-Film A.-G. zeigt:

M ist einer Hand.

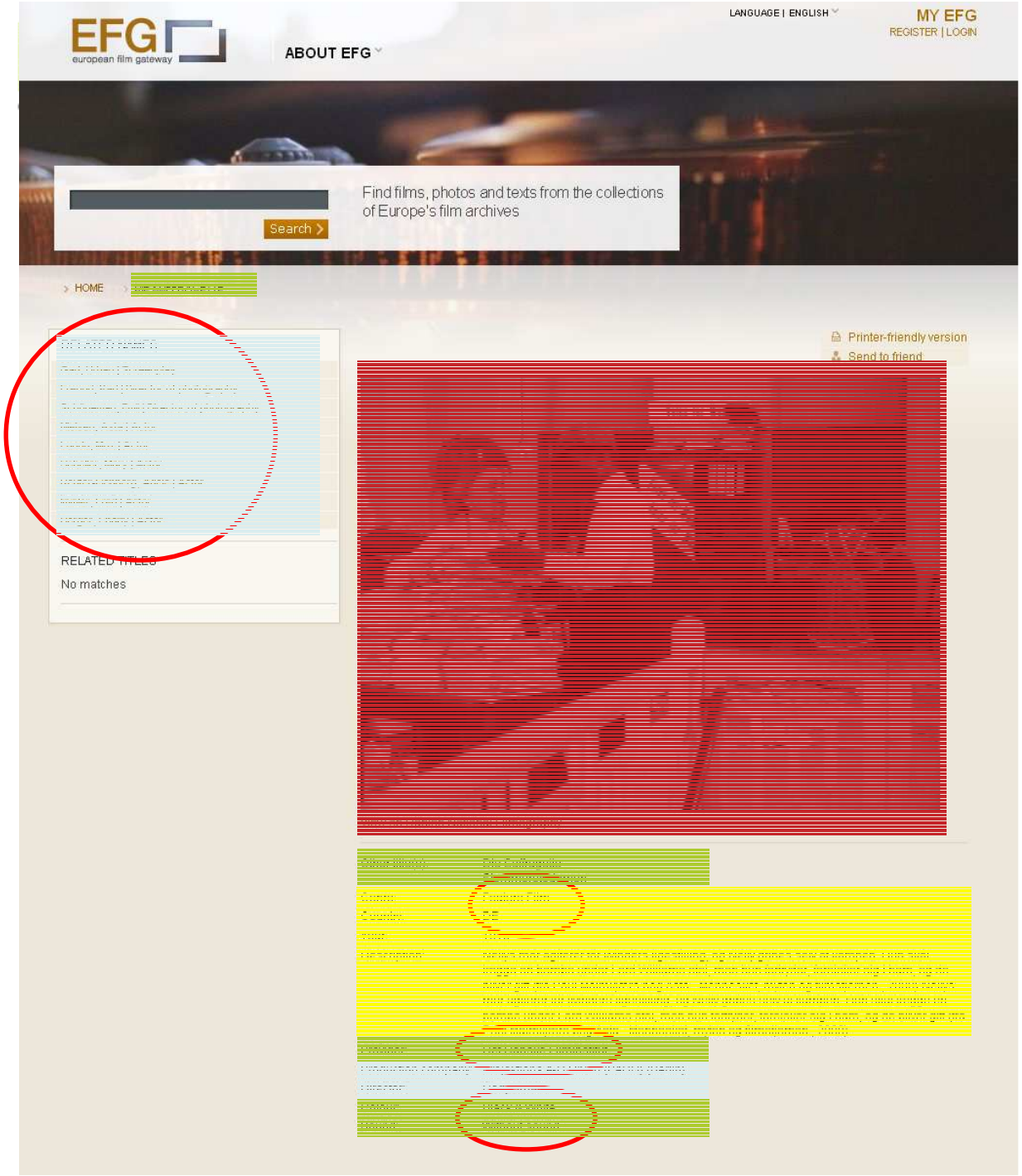
Ein Fritz Lang Film.

Zwischentitel. 1. Rolle. 1. Warte, warte nur ein Weilchen, Bald kommt der schwarze Mann zu Dir. Mit dem kleinen Hackebeilchen, Macht er Schabefleisch aus Dir. Du bist neun! Warte, warte nur ein Weilchen, Bald kommt der schwarze Mann zu Dir, Mit dem kleinen Hackebeilchen, Macht er Schabefleisch aus . . . 2. Da sollt der verfluchte Lied nicht singen, hab ich Euch gesagt. Könnt Ihr

Person

Peter Lorre *(Preferred name)*

Ladislaw Lowenstein *(Alternative name)*

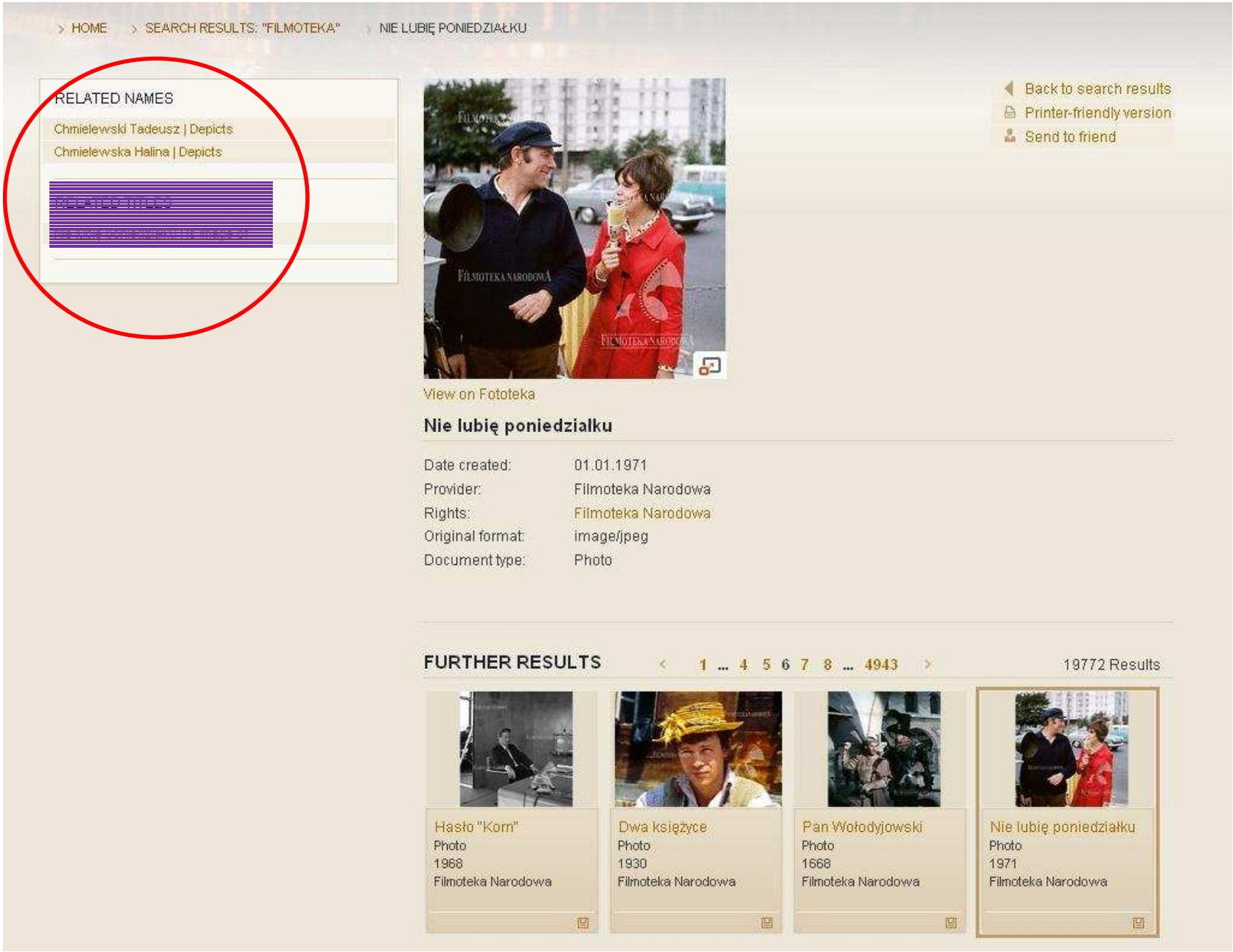


Work

Manifestation

Agent Relation

Item



[redacted] Work Relation

EFG Backend Tools

- Vocabulary Checker
- Authority File Manager
- Metadata Editor

- ✓ Support archives to enrich, clean and check the quality of their data in EFG

Vocabulary Checker

- Data cleaning on EFG level
- Tool used to verify vocabulary matchings

- matching intellectual work by partner archives
- no vocabulary management tool
- vocabularies and matching tables are managed outsite the EFG System

Vocabulary Checker

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_2BAB82FFFFF44B6DBD3FB5178DFA1D5D </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 eContent*plus*

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

EFG
european film gateway

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 Gülstorf	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 Kirchner	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		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

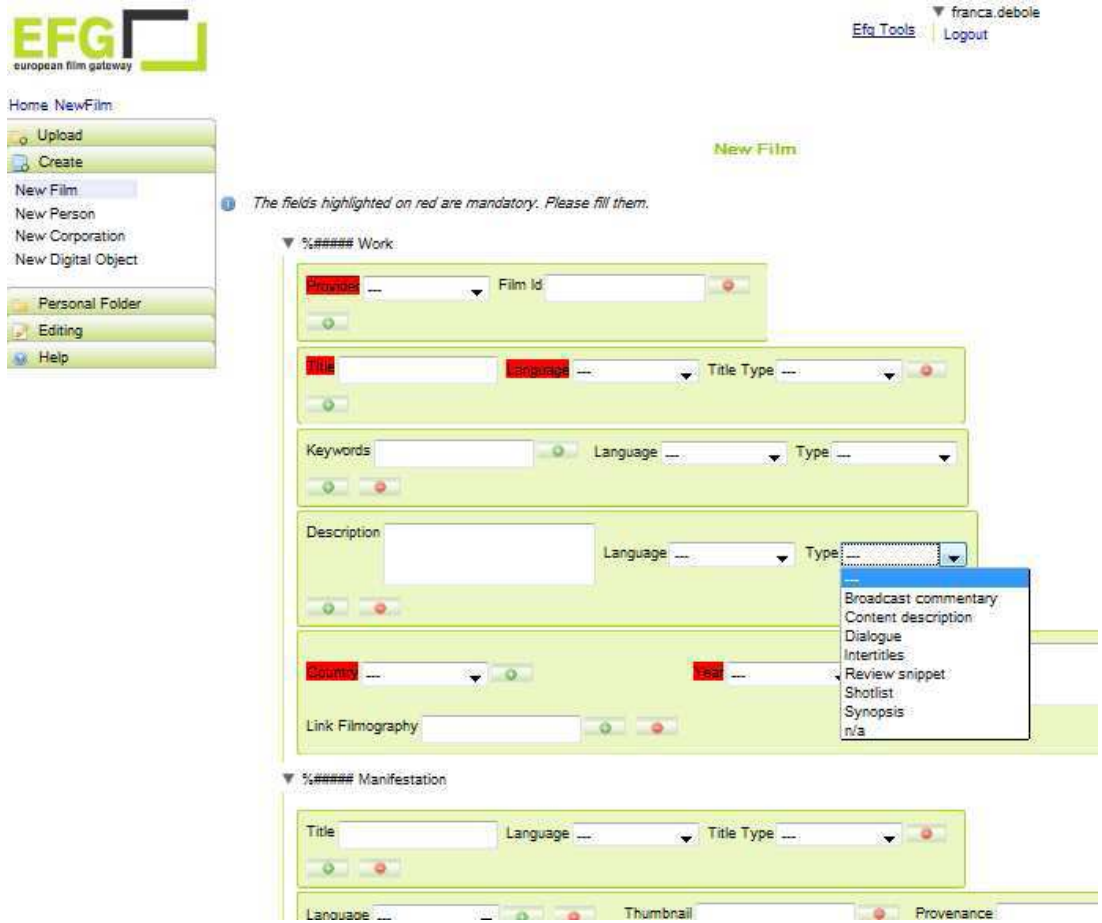
Metadata Editor

- Upload and ingest XML records into the EFG database
- Data enrichment on EFG level
- Tool allows partners to improve the quality of their metadata records

Most important enrichments:

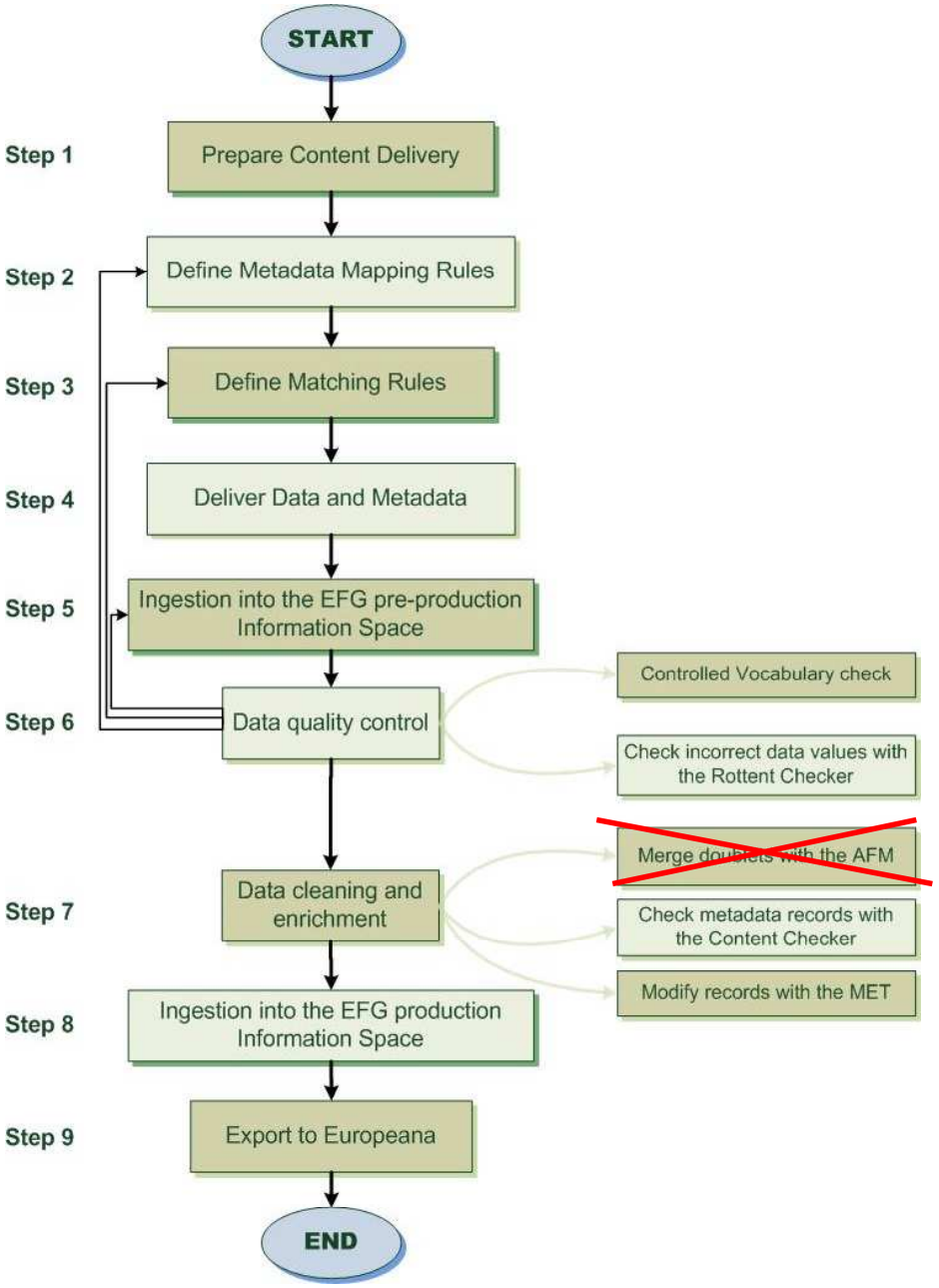
- add missing links to digital objects
- establish relations to persons depicted on an image, from digital object to a creator, et. al.

Metadata Editor



EFG
european film gateway

… how does all this fit into a workflow?



EFG european film gateway

Achievements: What has been done

- Compiled film-relevant mini-vocabularies
- Harmonized heterogenous and multilingual source values in EFG XML records
- Improved quality of film archival data in EFG through normalisation
- Contextualised film archival data through semantic relationships
- Display values consistently in EFG and Europeana portal
- Film archives improved the data quality in their local cataloguing systems
- Laid groundwork for common European filmography
- Developed semantic technology tools and tested workflows

Visions: What could further be done...

- Merge doublets of person and films to EFG authority records for a common European filmography
- Express EFG vocabularies in XML and link these files to the EFG metadata schema
- Provide EFG vocabularies in a language of the semantic web to the film archival community
- Implement multilingual vocabularies into EFG portal
- Apply a vocabulary management tool to update EFG vocabularies
- Find s
- Enrich film archival data with references to open ontologies or integrate them into external ressources (e.g. authority records)

Next steps: What will be done until Aug 11

- Evaluation of cataloguing work done by EFG partner archives
- Enrich EFG data with relevant relations and links (Metadata Editor)
- Implementation of EFG vocabularies into Linked Open Data
- Workshop on data quality and semantic interoperability issues in European film archives
  - ✓ 30 May in Frankfurt
  - ✓ Dedicated to film archival community
  - ✓ Bring forward standardisation of cataloguing and vocabulary work within the film archival sector

EFG
european film gateway

Thank you!

Questions?

www.europeanfilmgateway.eu