MINUTES

EFG Workshop on Data Quality and Semantic Interoperability Issues in European Film Archives

Frankfurt am Main, 30 May 2011

Attendees:

Alessandra Bani (Cinteca del Comune di Bologna), Detlev Balzer (System developer, Lübeck), Anke Bel (EYE Filminstituut Nederland, Amsterdam), Françoise Bourdon (Bibliothèque nationale de France, Paris), Tatiana Berseth-Abplanalp (Cinémathèque Suisse, Lausanne), Cornelius Bradter (Film and Television University Potsdam-Babelsberg), Atanas Chuposki (Kinoteka na Makedonija, Skopje), Robina Clayphan (Det Kongelige Bibliotek, Europeana, Den Haag), Beate Dannhorn (Deutsches Filminstitut – DIF), Edith van Dasler (EYE Filminstituut Nederland, Amsterdam), Maria José de Esteban (British Film Institute, London), Melanie Eichler (Deutsches Filminstitut – DIF), Eleonore Emsbach (Deutsches Filminstitut – DIF), Christiane Eulig (Deutsches Filminstitut – DIF), Annette Groschke (Deutsche Kinemathek – Museum für Film und Fernsehen, Berlin), Christiane Grün (Deutsche Kinemathek – Museum für Film und Fernsehen), Monika Haas (Deutsches Filminstitut – DIF), Adelheid Heftenberger (Österreichisches Filmmuseum, Vienna), Barbara Heinrich-Polte (Bundesarchiv-Filmarchiv, Berlin), Jim Heller (Frankfurt), Mervi Herranen (Kansallinen audiovisuaalinen arkisto, Helsinki), Antje Hirsch (Deutsches Filminstitut-DIF), Regina Hoffmann (Deutsche Kinemathek – Museum für Film und Fernsehen, Berlin), Jiří Horníček (Národní filmový archiv, Prague), Aline Houriet (Cinémathèque Suisse, Penthaz), Pavla Janásková (Národní filmový archiv, Prague), David Kocher (Lichtspiel – Kinemathek Bern), Markus Krottenhammer (Checkpoint Media / Österreichisches Filmmuseum, Vienna), Sylvie Lapeyre (La Cinémathèque française, Paris), Börje Lewin (Swedish National Heritage Board, Stockholm), Jutta Lindenthal (Information consultant, Lübeck), Ronny Loewy (Deutsches Filminstitut – DIF), Kurt Majcen (JOANNEUM RESEARCH, Graz), Anke Mebold (Deutsches Filminstitut – DIF), Randi Østvold (The National Library of Norway, Oslo), Barbara Pfeifer (Deutsche Nationalbibliothek, Frankfurt), Maria Assunta Pimpinelli (Cineteca Nazionale – Rome, FIAF Cataloguing Commission), Lisbeth Richter-Larsen (Det Danske Filminstitut), Uschi Rühle (Deutsches Filminstitut – DIF), Kirsten Rydland (The National Library of Norway, Oslo), Pasquale Savino (ISTI-CNR, Pisa), Margret Schild (Filmmuseum Düsseldorf), Francesca Schulze (Deutsches Filminstitut – DIF, EFG data co-ordinator, Frankfurt), Pernille Schütz (Det Danske Filminstitut, EFG WP 3 leader, Copenhagen), Uffe Smed (Det Danske Filminstitut), Barbara Vockenhuber (Österreichisches Filmmuseum), Julia Welter (Deutsches Filminstitut – DIF, EFG project manager, Frankfurt)

- Agenda

Content:

1) Welcome

1.1 Welcome and practical information

Pernille Schütz welcomes the participants and presenters on behalf of the Work Package 3 leader team of the project "EFG The European Film Gateway". EFG Work Package 3 "Content enrichment and semantic interoperability" is co-ordinated by the Danish and the German Film Institute. Aim of this workshop is to bring forward standardization efforts of cataloguing and vocabulary work in the film archival sector. The workshop wraps up what EFG Work Package 3 has achieved in the fields of data cleaning and enrichment so far – and what could be the benefits from this work for the film archival community. The workshop consists of four sessions (Cataloguing, Authority Files, Linked Open Data, Vocabularies). Furthermore, in each session external presenters report on the current status of work and discussion within their own project or initiative.

2.2 Short introduction to EFG

Julia Welter gives a short introduction to the EFG project. Beyond the main aim to develop a portal that gives access to digital material presently held in 16 European film archives, the project deals with issues like metadata interoperability, Intellectual Property Rights and usability. EFG has been a 3-years project under the eContentplus programme of the European Commission which will run out in August 2011. The portal is currently available as an internal beta version. It will be publicly launched by end of June under the URL www.europeanfilmgateway.eu.

2) Session 1: Cataloguing

2.1 Lessons learned from the EFG project

- Presentation 1: Pernille Schütz (Det Danske Film Institute)

Pernille Schütz, EFG Work Package 3 leader and head of DFI´s library and non-film collection department, explains that the overall objective of the EFG Work Package 3 leader team was to structure and harmonize the heterogeneous metadata contributed by 16 European film archives to the common EFG database. Main challenges were that the film archives do not use common standards for metadata structures, cataloguing rules and vocabularies. Furthermore, EFG has to deal with the issue that the same person or film work can be contributed from different archival databases to EFG. Guidelines and web tools were provided to the cataloguers of the partner archives so they could clean and enrich their data for EFG purposes. This also had a positive effect on the data quality of the local databases. The EFG Metadata Editor Tool is used by the archives to catalogue their metadata on the EFG level. The EFG web tools were developed by the technology partner ISTI-CNR who is leader of EFG Work Package 2 "Technical Interoperability and Access".

In August 2011, a final report summarizing the results from the archives´ cataloguing work for EFG will be available in the Outcomes section of the EFG project web site. This document (Deliverable 3.2 "Final report on type and quantity of archival resources tagged") will include the results from the cataloguing

activities the archives performed locally as well as directly in the EFG database. The report will also contain a lessons learned section.

Q & A outcomes:

- The numbers on the EFG archives´ data cleaning and enrichment work presented on slide 8 of the power point presentation include both, person and film work records.
- The EFG WP 3 leader team prefers that data cleaning and enrichment work is carried out in the local cataloguing systems. Only if this is not possible, a partner archive can perform this work directly in the EFG database with the help of the Metadata Editor Tool.
- Through the upload and ingest function of the Meta Data Editor archives can only integrate XML data in the predefined EFG XML format into the EFG system. This function should primarily be used by archives that have a small amount of records to contribute.

2.2 Cataloguing for EFG: An EFG partner´s experience

- Presentation 2: Edith van Dasler (EYE Filminstituut Nederland)

Edith van Dasler, co-ordinator of EYE´s cataloguing department, reports on the experiences she and her team has made when preparing data for EFG purposes. EYE is an archival partner of the EFG project and therefore member of EFG WP 3. Writing instructions for the mapping of data elements from the local database to the elements of the EFG database was also part of the EYE´s data work for EFG. Main challenges were to implement a new cataloguing system and to manage dependencies that exist with other projects EYE is involved in to make its digital material available online (for instance: Images for the Future). A lot of cataloguing work was carried out and the quality of EYE´s metadata improved during the EFG project. An achievement is that the improved metadata and technical implementations can be further used in the context of future projects (synergy effect). Two short films, so called “Living postcards”, are shown during the presentation. The film “Naar’t Tolhuis” shows the place where EYE’s new building is currently erected.

Q & A outcomes:

- The WP 3 leader team provided the cataloguers of the EFG partner archives with general guidelines on how they can enrich and clean their data for EFG needs. Edith states that EYE considered the priorities named in these guidelines. It would have been impossible for EYE to follow all recommendations as these have not been fully in line with the local cataloguing practice.
- EFG´s data cleaning and enrichment guidelines are publicly available in the EFG Data Provider Handbook (which contains all relevant information for institutions that wish to contribute data to EFG), on the project website: http://www.efgproject.eu/guidelines_and_standards.php.

2.3 News about the FIAF cataloguing rules

- Presentation 3: Maria Assunta Pimpinelli (Centro Sperimentale di Cinematografia - Cineteca Nazionale – Rome)

Maria Assunta Pimpinelli is a film preservation and restoration specialist. She deals with cataloguing issues of film collections at the CSC-CN and is a member of the FIAF Cataloguing and Documentation Commission. Her presentation focuses on the revision of the FIAF cataloguing rules and on the missing points of the former version: Originally, the FIAF rules were established for analogue film works and not for digital film objects. Thus, changes have become necessary and a revision of the rules is in progress since 2008, after a survey on FIAF cataloguing rules started in 2004.

Main aim of the FIAF cataloguing rules revision project is to make the rules compatible with MARC, CEN Metadata Specifications for Cinematographic Works and other metadata standards. Latest drafts of the new rules and further information about the project can be found on the wiki "filmstandards.org" which is primarily used by the FIAF working group as common communication platform: http://www.filmstandards.org/fiaf/wiki/doku.php.

Furthermore, Maria Assunta presents the activities of the FIAF Commission to revise of the Glossary of Filmographic Terms, which is now available as an electronic publication. The English version of the glossary is available on the FIAF web page since 2008: http://www.fiafnet.org/uk/publications/fep_Glossaryoffilmographicterms.cfm. Currently, terms and definitions are being translated into other languages: French, Italian and Portuguese translations have been completed; work on Spanish and German translations is in progress. The Commission is looking for volunteers helping to translate the glossary in the other languages. If you are interest in support this work please get in touch with Maria Assunta: mariaassunta.pimpinelli@fondazionecsc.it.

Q&A outcomes:

- Maria Assunta answering to a question by Françoise Bourdon about the relationship with the new cataloguing standard RDA (Record Description and Access, based on AACR – Anglo-American Cataloguing Rules), states that RDA provides hardly specific rules for audiovisual works at the moment. It could be useful if FIAF and RDA developers get in contact with each other on this issue. However, as RDA has already been released it would be necessary to wait for its next revision.

3) Session 2: Authority Files

3.1 EFG´s approach to authority file building

- Presentation 4: Francesca Schulze (Deutsches Filminstitut – DIF)

Francesca Schulze, data co-ordinator of the EFG project, gives an overview of the challenges and achievements of creating authority files in EFG. She introduces the EFG metadata schema as a tool to aggregate the heterogeneous filmographic data contributed by the partner archives to the common EFG

database. The complex EFG metadata schema is based on the FRBR oriented Cinematographic Works Standard (EN:15907). Furthermore, she presents the web tool developed by ISTI-CNR to establish authority files in the EFG database: The Authority File Manager. A recommendation for the future is that film archives could cooperate with national libraries in order to establish common authority files for persons and corporations with the help of the Linked (Open) Data technology.

Q&A outcomes:

- The EFG Authority File Manager Tool is available as open source software. The web tool works best in the Firefox browser. A public demo version and further information can be found here on the webpage of the Europeana ThoughtLab: http://www.europeana.eu/portal/thoughtlab_improvingmetadata.html
- More information on the EFG metadata schema can be found in the Guidelines & Standards section of the EFG project website: http://www.efgproject.eu/guidelines_and_standards.php.

3.2 Using authority files across domains. Introduction to a planned collaboration between Deutsche Nationalbibliothek (DNB) and Deutsches Filminstitut – DIF on using Linked Data technology for access to person records

- Presentation 5: Francesca Schulze (Deutsches Filminstitut – DIF)

Francesca Schulze presents the planned cooperation project between the German Film Institute and the German National Library. The institutions applied for funding at the Deutsche Forschungsgesellschaft (DFG). Aim of the 2-years project, which is expected to start in October 2011, is to match & merge person records of both institutions in order to offer access to the person data and their related resources from the two web portals: filmportal.de and dnb.de. The persons and their related works will be displayed in a common result page. Within this pilot project a new cooperation model will be set up and tested by using the Linked (Open) Data technology.

Q&A outcomes:

- A main aim of the project is to make person authority records sustainable by introducing the GND (Gemeinsame Normdatei) as a common reference for person records to the film archival domain and by improving the quality of the exchanged data in both data sources.
- Each record from filmportal.de will get a GND ID, respectively URI.
- An advantage of the cooperation for the German National Library is that search access to information about film will be enhanced which is new for the library sector.
- The data shall be exchanged monthly, so that records can be updated regularly.
- This pilot project bears potential for possible future cooperations of the German National Library with other institutions outside the library sector.

3.3 Contributing data to international authority files: The Virtual International Authority File (VIAF)

- Presentation 6: Françoise Bourdon (Bibliothèque nationale de France, Paris)

Françoise Bourdon, Deputy Head of the Bibliographic and Digital Information Department at BNF, gives an introduction to the cataloguing reference tool VIAF – The Virtual International Authority File developed by OCLC (Online Computer Library Center). She explains how institutions can contribute their authority data to the VIAF and what advantages this has. The VIAF is primarily used by national libraries as a common reference tool for cataloguing authority files (currently: persons and corporate bodies, later: geographic names and works). VIAF creates common IDs for the collected records which are internationally unique. One VIAF authority file consolidates all authority records contributed by various data source for the according person or corporate body. Thus, under one VIAF ID several authority records coming from different authority files from one country can be found. For instance, the National Library of Israel contributed three different authority files (one in Latin characters, one in Arabic characters and one in Hebrew characters), so for the person "Miguel de Cervantes Saavedra" (VIAF ID:17220427) this library contributed 3 different authority records.

Future plans foresee that also other institutions than libraries (e.g. archives and museums) join the VIAF (Quotation from Françoise Bourdon: "Future is on the semantic web! Beyond libraries!"). If you are interested in joining VIAF please get in touch with Françoise Bourdon: francoise.bourdon@bnf.fr.

Referring to Francesca Schulze´s earlier quote from EFG´s description of work (see: presentation 4, slide 20) Françoise Bourdon concludes her presentation that common international authority files facilitate the highest possible precision in information and content retrieval in various contexts.

Q&A outcomes:

- VIAF currently examines geographical names.
- Subjects are not supported by the VIAF tool.
- The VIAF tool should support cataloguers to with RDF and uniform work titles.
- Since Pernille Schütz wonders why the Danish National Library is not a partner of VIAF, Françoise Bourdon encourages all institutions to join the initiative. As long as institutions agree with the philosophy behind VIAF they are welcomed to participate.
- The process of clustering data in VIAF is completely automatic work. OCLC has constructed a technical solution for that.
- VIAF authority records are very well documented as they are proved for correctness any time a new partner joins in. Each contributed data element needs to be identified.

4) Session 3: Linked Open Data

4.1 Update on Europeana activities: Data enrichment, Linked Open Data and the Europeana Data Model

- Presentation 7: Robina Clayphan (Det Kongelige Bibliotek, Europeana, Den Haag)

Robina Clayphan, Interoperability Manager at Europeana, introduces the new Europeana Data Model (EDM). She states that the elements of Europeana´s current metadata schema "Europeana Semantic elements (ESE)" became a part of the new model. Furthermore, she reports on Eurpoeana´s latest data enrichment activities. To enrich the collected data with time periods, Europeana has developed a time ontology (http://annocultor.eu/time/). With the help of the AnnoCultor tool the data are automatically tagged with links to the according concept of this ontology. Robina Clayphan presents Europeana´s efforts and experiences in applying linked data techniques to Europeana´s metadata in the Linked Open Data Pilot.

Q&A outcomes:

- Robina Clayphan will inform the German Film Institute in how far Europeana´s time ontology and the AnnoCultor tool (http://annocultor.eu/) can be used by other projects as well after the workshop.
- So far, there is not much user feedback on the new Auto-tag function in the Europeana portal. This new function has not been advertised and feedback was not requested.
- Europeana is currently setting up a web page which provides guidance on how to use SPARQL in the frame of EDM. More information on Europeana´s Linked Open Data activities can be found here: http://version1.europeana.eu/web/lod/
- Europeana has not collaborated with VIAF so far but working with VIAF data is planned for the near future.

4.2 Introduction to Semantic Web technology: Pilot implementation of controlled film archival vocabularies into Linked Open Data / How to use a shared platform for standardisation purposes: The new wiki on www.filmstandards.org

Detlev Balzer, system developer and consultant, continues to bring the Semantic Web into focus and invites the audience to use a new wiki on "filmstandards.org" for submitting and discussing issues and proposals.

- http://filmstandards.org/fsc/index.php/Main_Page

He explains some of the motivation behind Linked Open Data (LOD) technology and gives a live demonstration of using different LOD datasets within a distributed query using SPARQL, one of the backbone technologies used for working with LOD. He stresses the importance of namespaces and universal resource identifiers (URIs) as prerequisites for the Semantic Web. His proposals for long-term maintenance of namespaces can be found in a FAQ section on filmstandards.org.

Practical uses of LOD are illustrated in a brief example: A poster from a poster collection contains details about screenings of various films. Augmenting the metadata for this poster would enable LOD-aware applications to offer new avenues for exploring the historical context of cinematographic artifacts. This and further examples of creating link-friendly filmographic metadata can be found in the CEN TC 372 Workshop Compendium:

- http://filmstandards.org/fsc/index.php/TC_372_Workshop_Compendium

Developing a strategy for LOD for a film archive requires several planning decisions in addition to representing and expressing filmographic metadata. Long-term stability of identifiers, suitable namespace policies, as well as adoption and maintenance of appropriate predicate vocabularies are among the crucial ingredients for a promising LOD strategy. A first-hand collection of guidelines can be found in

- www.linkeddatabook.com/editions/1.0.

In the last part of his talk, Detlev Balzer introduces the standard for cinemathographic works, EN 15907: http://filmstandards.org/fsc/index.php/EN_15907 that was developed by the Technical Committee TC 372 of the European Committee of Standardization – CEN. EN 15907 was designed as an interoperability specification addressing the multitude of databases that exist in audiovisual heritage institutions throughout Europe. Its data model builds on some concepts from FRBR, an influential reference model from the library community. In addition to supporting interoperability, EN 15907 can also serve as a reference specification for revising filmographic information systems (or for designing new ones) with an emphasis on co-operative cataloguing, data exchange, and the Linked Open Data paradigm.

Detlev Balzer encourages participants to come up with suggestions and comments on the filmstandards.org wiki (registration will enable write access).

Q&A outcomes:

- Börje Lewin (Swedish National Heritage Board) offers his help regarding RDF issues. Anyone interested can contact him via borje.lewin@raa.se.

5) Session 4: Controlled Vocabularies

5.1 EFG´s vocabulary and matching work

- Presentation 10: Francesca Schulze (Deutsches Filminstitut – DIF)

Francesca Schulze reports on the vocabulary and matching work which has been carried out in EFG so far. Main objective of this work has been to harmonize the heterogeneous and multilingual source values in the common EFG database for a coherent display in both the EFG and Europeana web portal. The vocabularies were compiled according to the needs of the portals´ end users. Whenever possible, existing standards were applied. A major challenge is that many source terms have been contributed via free-text fields which have increased the necessary matching work especially when updates of contributions are delivered to the common database. Furthermore, she demonstrates the EFG Vocabulary Checker Tool which is used to verify the vocabulary matchings established by the partner archives. This tool was

developed by EFG´s technology partner ISTI-CNR. Eventually, an overview of the complete EFG data ingestion and metadata editing workflow is given including all cataloguing activities that were introduced earlier during the workshop. The EFG vocabularies, the EFG metadata schema as well as EFG´s data cleaning and enrichment guidelines (part of "EFG Data Provider Handbook") are available on the Guidelines & Standards section of the EFG project web site: http://www.efgproject.eu/guidelines_and_standards.php.

Q&A outcomes:

- A positive result from the archives´ local cataloguing work for EFG is that controlled vocabularies were introduced for fields that were formerly managed as free-text. For instance, the Danish and the German Film Institute are now using an according vocabulary for the person's functions / activities.

5.2 Controlled vocabularies: What they are and how they perform in various contexts / How to create and maintain vocabularies: The web-based tool xTree

- Presentation 11: Jutta Lindenthal (Information consultant, Lübeck)

Jutta Lindenthal introduces four main types of controlled vocabularies together with examples from the film archival domain: keyword list, taxonomy, thesaurus and ontology. Depending on the context of use each vocabulary type has its advantages and disadvantages. She underlines the need for uniquely addressable concepts for vocabulary control. Guidance for vocabulary management can be found in the new ISO 25964 standard "Thesauri and interoperability with other vocabularies".

In the second half of her presentation, Jutta Lindenthal demonstrates the vocabulary management tool xTree on behalf of Axel Vitzthum (digiCULT-Verbund eG) who could not attend the workshop. Development of the xTree tool was started within the framework of the 3-years digiCULT project initiated by the Christian-Albrecht University Kiel and the Schleswig-Holstein Museum Association in order to digitally record and publish inventories of museums. The project expired in 2010 and the work has been followed up in the digiCULT network since. xTree is a web-based tool that allows to manage thesauri and other vocabularies collaboratively. Jutta Lindenthal demonstrates the tool by means of selected film archival vocabularies.

Q&A outcome:

- The xTree tool is based on open source software. For more information please contact: Lütger Landwehr (digiCULT-Verbund eG: http://www.digicult-verbund.de/index.php?p=Kontakt)