Data Integration for the Media Value Chain

Henning Agt-Rickauer[1], Jörg Waitelonis[2], Tabea Tietz[1], and Harald Sack[1]

[1] Hasso Plattner Institute, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
{firstname.lastname}@hpi.de,
[2] yovisto GmbH, August-Bebel-Str. 26-53, 14482 Potsdam, Germany
joerg@yovisto.com

1 Introduction

With the switch from analog to digital technology the entire process of production, distribution, and archival of a film and tv program large amounts of data are created. Besides recorded and processed audiovisual information, in each single step of the production process and furthermore throughout the entire media value chain new metadata is created, administrated, and put into relation with already existing metadata mandatory for the management of these processes. Due to competing standards as well as to proprietary and incompatible interfaces of the applied software tools, a significant amount of this metadata cannot be reused and is not available for subsequent steps in the process chain. As a consequence most of this valuable information has to be costly recreated in each single step of media production, distribution, and archival. Currently, there is no generally accepted nor commonly used metadata exchange format that is applied throughout the media value chain. But, also the market for media production companies has changed dramatically towards the internet as being the preferred distribution channel for all media content. Today's available limited budget for media production companies puts additional pressure to work in a cost and time efficient way and not to waste resources due to the necessity of costly reengineering of lost metadata. The *dwerft* project aims to apply Linked Data principles for all metadata exchange through all steps of the media value chain [4]. Starting with the very first idea for a script, all metadata is converted according to either existing or newly developed ontologies to be reused in subsequent steps of the media value chain. Thus, metadata collected during the media production becomes a valuable asset not only for each step from pre- to postproduction, but also in distribution and archival.

This paper presents results of the *dwerft* project about the successful integration of a set of film production tools based on the Linked Production Data Cloud, a technology platform for the film and tv industry to enable software interoperability used in production, distribution, and archival of audiovisual content.

2 Linked Production Data Cloud

The core of the *dwerft* project is the **Linked Production Data Cloud** (LPDC), a technology platform for the film and television industry that allows lossless

interoperability between software and hardware tools used in production, distribution, and archiving of audiovisual content. Based on Linked Open Data principles [1] the LPDC stores and publishes semantic metadata originating from different subtasks of the film production process under a unified ontology schema. Fig. 1 provides an overview of the LPDC and connected production tools of an example show case. The key components of the LPDC are: an extensible vocabulary for metadata storage, a set of pre-defined converters for RDF data generation, a framework to develop customized converters, a tool to manage inserts and updates of RDF data including versioning, and a triplestore for RDF data management and querying.

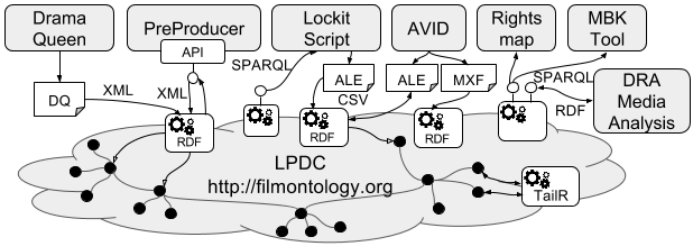

Fig. 1. Data integration use case for tools and applications in the media value chain

The **Film Ontology**[3] vocabulary was designed in collaboration with domain experts to create a suitable terminology describing the different tasks of media production and all associated metadata. The ontology schema is capable of representing film scripts (e.g., scenes, scene content, characters, sets, etc.), production planning metadata (e.g., film crew, departments, cast, filming locations, shooting schedule, used equipment, etc.), on-set information (e.g., shots, takes, and associated clips), post production metadata (e.g., timecodes, codecs, resolutions, and formats of recorded and further processed clips), as well as metadata for quality assessment of archived audiovisual material (e.g., surface damages, splices, bulges, glued areas, etc.). Where ever possible, already existing vocabularies have been reused, mapped, and interlinked, such as e.g., Broadcast Metadata Exchange Format (BMF)[4], EBUcore[5], or DBpedia Ontology[6]. The collaborative design of the Film Ontology was carried out with WebProtégé [2]. Currently, the vocabulary is further extended with rights management information, film editing metadata (e.g., cut information), and technical metadata of rendered movie containers for delivery and distribution (e.g., Material Exchange Format (MXF)). None of the participating software applications was originally capable of importing, exporting, or processing RDF data. First, a set of cus-

[3] http://filmontology.org
[4] https://www.irt.de/en/activities/production/bmf.html
[5] https://tech.ebu.ch/MetadataEbuCore
[6] http://mappings.dbpedia.org/server/ontology/classes/

tomized converters was developed to transform proprietary metadata produced by the tools into RDF representations conforming to the Film Ontology. The analysis of the production workflows has shown that most of the created production metadata is encoded in XML and CSV formats. Therefore, the *dwerft tools* converter framework has been developed to efficiently create customized CSV/XML-to-RDF converters[7]. The framework includes predefined converters for a set of film production applications as well as a generic CSV/XML-to-RDF converter that allows to create the required transformations on custom metadata based on lightweight mapping definitions.

RDF Metadata generated by different converters is stored in a RDF triplestore and can be queried via SPARQL. As a proof of concept, semantic metadata originating from a test film production at the Tempelhofer Feld in Berlin is available for further use[8] and can be searched[9].

In a setting where data from heterogenous sources is transformed, aggregated, and stored in a triplestore, it is essential to manage updates of the data. In our approach, we have integrated the linked data versioning system TailR [3]. RDF data generated by converters is first uploaded to TailR. In case the original data is changed and converted again – as it usually often happens, as e.g., during filming, when changes are made in dialogs to adapt them according to the intention of the director or the preferences of an actor – , TailR stores each version and generates RDF diffs. These are used to derive respective SPARQL insert and delete statements in order to update the RDF data in the RDF store accordingly.

3 Integrated Film Production Applications

An exemplary set of tools, representative for the different stages pre-production, planning, shooting, post-production, distribution and archiving, was chosen, analyzed with respect to interoperability and connected to the Linked Production Data Cloud. *DramaQueen*[10] is a script writing software to develop, visualize, and analyze stories. It allows working from the first idea to the final script using predefined formatting, storylines, characters, outline, synopsis, and story charts. DramaQueen is a Java based standalone application and uses a proprietary data format based on XML to store script projects. *PreProducer*[11] is a film production management software to support the complete preproduction planning process. It features general project management, script analysis, management of crew, cast, inventory, and filming locations, development of shooting schedules, budgeting and financial calculations. PreProducer is a web-based application and offers partial export and import based on XML documents via a REST API. *LockitScript*[12] is a mobile web application used during film shooting. It supports

[7] The *dwerft tools* framework is available at https://github.com/yovisto/dwerft
[8] http://filmontology.org/resource/DWERFT
[9] http://filmontology.org/search/
[10] http://dramaqueen.info/about-en/?lang=en
[11] http://www.preproducer.com/index.html
[12] http://lockitnetwork.com/home/

the script supervisor to oversee the continuity of the movie and keeps track of the daily progress. It also manages the linking of scenes and takes to filmed clips and uses a special hardware device to directly synchronize camera data with its backend. LockitScript offers limited export facilities for daily reports and camera metadata in the web interface. *AVID Log Exchange* (ALE)[13] is a file format used by various cameras and post-production tools (e.g., Arri Alexa, AVID Media Composer, DaVinci Resolve, Silverstack) to exchange metadata about filmed movie clips. The integration of ALE is challenging, because each tool defines custom columns in the CSV format. While the previously described tools primarily produce metadata, the distribution phase of a film production usually requires metadata of all steps of the production process. Two tools already benefit from the early availability of semantic metadata using SPARQL queries: *rightsmap*[14], a licence management solution for film and tv productions, and the "*Medienbegleitkarte*" (MBK), a metadata set based on the Broadcast Metadata Exchange Format (BMF) mandatory for delivery at German public-service tv broadcasters. Finally, media condition analysis tools by the German Broadcasting Archive directly insert analysis reports as RDF data into the LPDC.

4 Conclusion and Outlook

With the *dwerft* project and the LPDC framework a first subset of applications and tools has been integrated for lossless metadata exchange in the media production cycle. Metadata from media production and archival thus become a valuable asset used to enable better search and retrieval as e.g. for video on demand platforms, where it can also be used to support content-based recommendation and customized advertising.

Acknowledgement: This work has been funded by the German Government, Federal Ministry of Education and Research under project number 03WKCJ4D.

References

1. T. Heath and C. Bizer. *Linked Data: Evolving the Web Into a Global Data Space*. Synthesis Lectures on Web Engineering Series. Morgan & Claypool, 2011.
2. M. Horridge, T. Tudorache, C. Nuylas, J. Vendetti, N. F. Noy, and M. A. Musen. Webprotege: a collaborative web based platform for editing biomedical ontologies. *Bioinformatics*, page btu256, 2014.
3. P. Meinhardt, M. Knuth, and H. Sack. Tailr: a platform for preserving history on the web of data. In *Proc.s of the 11th Int. Conf. on Semantic Systems*, pages 57–64. ACM, 2015.
4. H. Sack. From Script Idea to TV Rerun: The Idea of Linked Production Data in the Media Value Chain. In *Proc. of the 24th Int. Conf. on World Wide Web Companion*, WWW '15 Companion, pages 719–720, 2015.

[13] http://www.avid.com/en/media-composer/features (Log and track metadata)
[14] http://www.recoupmentpro.de/