published as: Hercher, J. & Sack, H. Indexing German Audiovisual Heritage - Skosification made Easy 9th European NKOS Workshop at the 14th ECDL Conference, 9-10 September 2010, Glasgow, Scotland., 2010

Indexing German Audiovisual Heritage — SKOSification Made Easy

Johannes Hercher and Harald Sack

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH,
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
{johannes.hercher,harald.sack}@hpi.uni-potsdam.de
http://www.hpi.uni-potsdam.de

In order to foster creative industries the European Union supports the digitization and indexing of cultural heritage objects. This is also the case for the domain of broadcasting- and film-archives. Various projects have been funded by European or National governments, as e.g., European-Film-Gateway (EFG)[1], EUscreen[2] or Mediaglobe[3]. First results have already become visible as the EUscreen project launched the videoactive[4] platform, which makes use of controlled vocabularies for facetted browsing and filters, as e.g., on genre, language, time, and topics.

Mediaglobe[5] as being a small and medium enterprise (SME) spinoff project of the THESEUS/CONTENTUS[6] research program funded by the German federal Ministry of Economic and Technology, aims to improve workflows in media archives and broadcasting agencies. In particular more than 1.000 hours of historical film collections are to be digitized by Mediaglobe's project partners. These collections include both raw and edited material, and focus on major historical events of the 20th century — providing recordings of interviews and public speeches, weekly newsreel productions, and other documentaries. By making use of sophisticated multimedia analysis techniques, harnessing semantic metadata annotations, based on open linked data (LOD), and using crowdsourcing approaches, we infer high-level descriptions from media features like date, title, personal- and geographical-names as well as from other low level feature analysis. But, these descriptors have to be taken from controlled vocabularies in order to enable the exchange of data between collections that have different provenance and structure but deal with similar topics. Therefore, vocabularies such as classifications or thesauri play a key role to align heterogeneous collections, even in a semiautomatic manner [1, 2]. Furthermore, controlled vocabularies are useful to create facetted search and help to gain a better user experience [3, 4]. Publishing well established knowledge-structures, by using open standards such

[1] http://www.europeanfilmgateway.eu/ — last visited 21 Jun 2010
[2] http://www.euscreen.eu/— last visited 21 Jun 2010
[3] http://www.projekt-mediaglobe.de/ — last visited 21 Jun 2010
[4] http://www.videoactive.eu/ — last visited 21 Jun 2010
[5] http://www.projekt-mediaglobe.de/ — last visited 21 Jun 2010
[6] http://theseus-programm.de/anwendungsszenarien/contentus/ — last visited 21 Jun 2010

as the Resource Description Framework (RDF)[7], and the Simple Knowledge Organisation System (SKOS)[8], may be a nucleus for a variety of governmental, educational and commercial applications.

A strong commitment is required, as previous research has shown a lack of public available controlled vocabularies to support indexing of audiovisual material [5]. In contrast there is a vast amount of film related metadata standards as, e.g., P/Meta[9], MPEG-7[10], or BBC's Standard Exchange Framework (SMEF)[11]. But, these are primary designed for formal metadata exchange. For example MPEG-7 defines metadata structures and semantics — also for content-management and description — but its vocabulary terms, e.g. for classification, subjects and genre are strongly limited *c.f. ISO/IEC 15938-5)*. In contrast, controlled vocabularies are appropriate for collection-alignment on a content oriented layer by providing fine-grained collections of terms.

In fact, there already exist classifications and thesauri dedicated for indexing audiovisual material, but they are only available as print version, as e.g., [6], or so far 'hidden' in private or public media archives. Though, there is an increasing demand for audiovisual related vocabularies, we have encouraged major film archives in Germany, as e.g., *German National Archive* (BA), *National Broadcasting Archive* (DRA) and *German Broadcasting Compound* to make their controlled vocabularies available for public use. The following contributions to the scientific- and film-industry community, have motivated this step:

- Reuse of controlled vocabulary on films and film related material for indexing, facetted browsing, and retrieval issues.
- Interlinking and alignment of heterogeneous collections on the content-layer by using thesauri and classifications with semantic web technology such as RDF, SKOS and LOD.
- Promotion of open standards for knowledge exchange and adding value for small and middle-size archives by providing a conversion and hosting service.

We plan to support the public funded media archives with manpower to make their vocabularies available in SKOS-format. SKOS is the W3C's standard for building Knowledge Organisation Systems (KOS) for the web, and is well accepted because of its interoperability [7]. We will pay special attention on the correct use of the SKOS standard, hence many vocabularies and data within LOD cloud suffer from inaccuracy and of low quality [8]. In this context we discuss the feasibility of a *plain text to SKOS transformation* by exploiting indention-level and descriptor-labels with an automated procedure. At the end, we show the practical use of SKOSified thesauri for facetted search within video assets by the example of the Mediaglobe search interface and ontology.

[7] `http://www.w3.org/RDF/` — last visited 21 Jun 2010
[8] `http://www.w3.org/2004/02/skos/` — last visited 21 Jun 2010
[9] `http://tech.ebu.ch/docs/tech/tech3295v2_1.pdf` — last visited 21 Jun 2010
[10] `http://mpeg.chiariglione.org/` — last visited 21 Jun 2010
[11] `http://www.bbc.co.uk/guidelines/smef/` — last visited 21 Jun 2010

Meanwhile, other projects such as EFG and EUscreen have similar intentions. But, EFG is limited to translation of cataloging rules, as e.g., [9] into different languages in order to harmonize formal metadata exchange. Our approach is different, because we focus on the content level. Also, our approach differs from EUscreen as to our knowledge, they do neither use the vocabularies of German audiovisual archives nor do they plan to make vocabularies available for public use. Regarding SKOSification, there was already major research on converting existing thesauri. But van Assem et.al., focussed on XML-encoded thesauri and provided scripts for transformation of, as e.g., Medical Subject Headings (MESH) and WordNet [10]. In contrast, we focus on the audiovisual domain and try to convert existing thesauri from plain text to SKOS. In fact, various tools are available for creating and maintaining vocabularies, based on common web standards such as RDF, SKOS, and OWL.[12] But, these tools don't provide straightforward import of plain text formatted thesauri or export to RDF-SKOS file.

Up to now, we have distinguished only one thesaurus to index audiovisual material [6]. Moreover, a survey in German national audiovisual archives showed that limited resources seem to retard the publication of controlled vocabularies in this domain, although most archives are willing to contribute. Currently, getting access to their authority-files and thesauri is being negotiated. In parallel, we are developing tools for efficient *text to SKOS transformation* to support also small and middle-sized film archives in the publication of their vocabularies via a webservice.

References

1. Isaac, A., Schlobach, S., Matthezing, H., Zinn, C.: Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. Library Review **57**(3) (2008) 187 – 199
2. van der Meij, L., Isaac, A., Zinn, C.: A Web-Based repository service for vocabularies and alignments in the cultural heritage domain. In: The Semantic Web: Research and Applications. Springer (2010) 394–409
3. Morville, P., Rosenfeld, L.: Information architecture for the World Wide Web. 3rd. edn. O'Reilly, Sebastopol, Ca, USA (2007)
4. Troncy, R.: Bringing the IPTC news architecture into the semantic web. In Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., eds.: Proceedings of the 7th International Conference on The Semantic Web (ISWC'08). Volume 5318 of Lecture Notes in Computer Science. Springer, Berlin ; Heidelberg (2008) 483–498
5. Oomen, J., Smulders, H.: Multimatch : First analysis of metadata in the cultural heritage domain. Technical Report D 2.1, Netherlands Institute for Sound and Vision (May 2006)
6. Intner, S.S., Studwell, W.E.: Subject access to films and videos. Soldier Creek Press, Lake Crystal, Minn., USA (1992)

[12] c.f.: Protégé `http://protege.stanford.edu/` , PoolParty `http://poolparty.punkt.at/`, Neologism `http://neologism.deri.ie/` — last visited 21 Jun 2010

7. Pastor, J.A., Martinez, F.: Advantages of thesauri representation with the simple knowledge organization system (SKOS) compared with other proposed alternatives for the design of a web-based thesauri management system. Information Research **14**(4) (December 2009)
8. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: Proceedings of the Linked Data on the Web WWW2010 Workshop (LDOW 2010), Raleigh, North Carolina, USA (April 2010)
9. Harrison, H., Commission., F.C.: The FIAF cataloguing rules for film archives. K.G. Saur, München ; New York (1991)
10. van Assem, M., Malaisé, V., Miles, A., Schreiber, G.: A method to convert thesauri to SKOS. In: The Semantic Web: Research and Applications. Springer, Berlin ; Heidelberg ; New York (2006) 95–109