Acquisition of audiovisual Scientific Technical Information from OSGeo by TIB Hannover: A work in progress report

Peter Löwe[1], Margret Plank[1], Paloma Marín-Arraiza[1]

[1] German National Library of Science and Technology (TIB)

Abstract

This paper gives a work in progress report on the application of the TIB|AV Portal for audiovisual OSGeo content. The portal is a web-based platform for audiovisual media combining state-of-the art multimedia analysis with semantic based analysis, and retrieval. It meets the requirements by special libraries for reliable long term preservation, scientific citation via persistent identifiers, and applies metadata enhancement to enable innovative services for search and retrieval.

Keywords

Audiovisual content, Digital Object Identifiers, Citation, OSGeo, Preservation, Semantic Web

1 The Challenge of Scientific Technical Information

Communication within OSGeo involves organisation committees, project maintainers, developers, application of related topics, education and reach-out. From the library perspective, much of this consists of scientific technical information (STI) and is conducted on alternative channels beyond the traditional journal-based scientific discussion. The advent of ubiquitous digital recording equipment and screen-cast-software has led to a steady rise of audiovisual content, such as lecture recordings or data animations. This trend continues and is likely to accelerate further due to the widespread use of handheld end-user devices to access such content, a diversification of application scenarios (“software mash-ups”), and the growing number of OSGeo-approved software projects.
Currently the majority of this content is distributed via proprietary commercial Web 2.0 platforms (e.g. Youtube). Access to and reuse of such content is hampered due to the lack of appropriate metadata, citation, and long term preservation. These topics exceed the scope of the Web 2.0 audiovisual platforms that are currently used., However, theyare crucial for the discoverability, the long term availability and re-use of audiovisual STI in general. This is part of a larger challenge currently affecting academia regarding access to heterogeneous STI content as part of the general trend towards data-driven science.

2 The German National Library of Science and Technology

The German National Library of Science and Technology (TIB) ranks as one of the largest special libraries worldwide. Its task is to comprehensively acquire

and archive literature from around the world pertaining to all areas of engineering, architecture, chemistry, information technology, mathematics, and physics. TIB actively tracks relevant output from OSGeo projects, whichis significant for the field of applied computer science.
Moreover, the information portal of the TIB, GetInfo, provides access to more than 160 million data sets from specialised databases, publishers, and library catalogues.
The access to non-textual materials such as audiovisual media, 3D objects, and reseach data, as well as the use of these materials, also concerns the TIB . This is the main task of the Competence Centre for non-textual Materials. Tools and infrastructure that enable the easy publication, discoverability, and long-term availability of non-textual materials are developed to support users in scientific work processes.

3 The TIB|AV Portal Video Platform

Launched for practical use in April 2014, the TIB|AV-Portal is a bilingual (English/German) web-based portal for audiovisual media that optimises access to scientific videos in the field of science and technology. It was developed together with the Hasso-Plattner Institute and designed to overcome the limitations encountered in current commercial Web 2.0 video portals according to the requirements of a data-driven special library. By combining state-of-the-art multimedia retrieval techniques with semantic analysis, it provides content-based access to videos at the segment level, and link content to new knowledge. The processing workflow of the video analysis includes structural analysis based on video shot detection, optical character recognition, automated speech-to-text transcription, and visual concept detection.

3.1 Metadata of the TIB|AV Portal and the Linked Open Data Context

Automatically generated metadata are processed via linguistic and semantic analysis. That is, named entities are identified, disambiguated and mapped to an authoritative knowledge base. This knowledge base consists of subject specific parts of the Integrated German Authority Files (GND), available as Linked Data Services of the German National Library. The English labels were gained by mapping GND entities onto other authority files.
The portal uses it own schema for the authoritative metadata (formal, technical and content-describing) to describe and manage its content in a standardised manner (Lichtenstein et al., 2014).
The automatical and authoritative metadata of each video are stored in RDF form in a local RDF Store (also known as Triple-Store). The connections between elements stored in the RDFStore are annotated by using an internal ontology. It is planned to map and merge this ontology onto external vocabularies available as Linked Open Vocabularies in other to improve the interoperability of the data.
Furthermore, additional information concerning authors and institutions will be included to facilitate the searchability and discoverability of the content.

3.2 Features and Services

Each video is registered by a digital object identifier (DOI). A DOI name clearly identifies the video, akin to the use of ISBN in books. In addition, the TIB|AV

Portal offers a time-based citation link, enabling a citable DOI to be displayed for each video segment using the open standard media fragment identifier (MFID).
A visual table of contents provides a quick overview of the video facilitating access to individual video sequences. Content-based filter facets for search results enable the exploration of the increasing number of video assets. The term search is not only performed within authoritative metadata but also within metadata from video analysis, giving different term weight when searching. These techniques allow the users to search more efficiently and find content that otherwise would remain hidden. Producers of scientific films, such as the OSGeo communities, can upload their video to the TIB|AV-Portal free of charge. Once the quality of the video has been checked, it is published in a legally watertight manner, indexed according to international standards, transcribed, digitally preserved, and given a DOI name. This ensures an optimal discoverability of scientific films.

3.3 Collections

Films from the TIB´s fields are collected in both German and English. This covers recordings of conference presentations, panel discussions, and recordings of experiments (microscopic images, modelling, simulations and presentations of specific software), among others. Audiovisual content from the OSGeo communities has been steadily acquired since its launch in 2014. The content available by March 2015 ranges from conference recordings (FOSS4G-EU 2013, FOSSGIS 2014, FOSS4G 2014) to thematic-renderings of code evolution (e.g. http://dx.doi.org/10.5446/14652) and historic GIS footage dating back to 1987 (e.g. http://dx.doi.org/10.5446/12963).

4 The European Perspective on audiovisual Scientific Technical Information

The large amount of audiovisual STI generated by OSGeo exceeds both thematically (i.e. topics beyond TIB's acquisition foci) and qualitatively (i.e. multilingual content) the range of acquisition and customer services offered by TIB.
A comprehensive acquisition strategy addressing the multilingual and multitopical diversity of OSGeo's scientifically relevant output remains a challenge to be taken on by several special libraries. This will lead to added value for the OSGeo communities, libraries and their users, and also the general public. The experiences gained from the TIB|AV-Portal can serve as a starting point towards a modular best-practice based approach for this objective.

5 Conclusions

Digital audiovisual content has become an important communication channel for STI within and beyond OSGeo. This content is currently hosted on publicly available commercial web portals whose functionalities do not meet the needs for reliable long term scientific preservation, access, and citation. Since the production of audiovisual content diversifies and accelerates, best practices are needed to address and solve this challenge on a global scale. The TIB|AV-Portal for audiovisual STI meets the requirements to preserve such content and to provide innovative services for search and retrieval. Simultaneously, the

research and development for improved services continues. Quality checked audiovisual content from the OSGeo communities is constantly being acquired for the portal as a part of TIB’s mission to preserve relevant content in applied computer sciences for science, industry, and the general public.

6 References

[1] Berners-Lee, T. (2006, July 27). Linked Data-Design Issues. Available in : http://www.w3.org/DesignIssues/LinkedData.html

[2] Lichtenstein, A., Plank, M., & Neumann, J. (2014). TIB's Portal for Audiovisual Media: Combining Manual and Automatic Indexing. *Cataloging & Classification Quarterly*. 52(5), 562-577. doi: 10.1080/01639374.2014.917135.

[3] Strobel, S. (2014). Englischsprachige Erweiterung des TIB|AV-Portals: Ein GND-DBpedia-Mapping zur Gewinnung eines englischen Begriffssystems. *o-bib*. 1(1). 197-204. doi: http://dx.doi.org/10.5282/o-bib/2014H1S197-204.