Roeland Ordelman[1,2], Franciska de Jong[1,3]

Distributed Acces to Oral History collections:

Fitting Access Technology to the Needs of Collection Owners and Researchers

(1) University of Twente, Enschede, The Netherlands

(2) Netherlands Institute for Sound and Vision, Hilversum, The Netherlands

(3) Virtual Knowledge Studio, Erasmus University, Rotterdam, The Netherlands

Introduction

In contrast with the large amounts of potential interesting research material in *digital multimedia repositories*, the opportunities to unveil the gems therein are still very limited. The Oral History project 'Verteld Verleden' (Dutch literal translation of Oral History) that is currently running in The Netherlands, focuses on improving access to *spoken testimonies* in collections, spread over many Dutch cultural heritage institutions, by deploying modern technology both concerning infrastructure and access. Key objective in the project is mapping the various specific requirements of collection owners and researchers regarding both publishing and access by means of current state-of-the-technology. In order to demonstrate the potential, Verteld Verleden develops an Oral History portal that provides access to distributed collections. At the same time, practical step-by-step plans are provided to get to work with modern access technologies. In this way, a solid starting point for sustained access to Oral History collections can be established.

Technology and daily practice

The Verteld Verleden project builds upon years of academic research on access technology for spoken word archives [1,2,3] deploying among others automatic speech recognition, text-to-speech synchronization and fragment-level search. This research resulted in a number of demonstration applications in close cooperation with cultural heritage institutes [4,5,6]. Although these demonstrations have been very well received both by the public and involved institutes, it was observed that there is still a gap between academic, 'technology-driven' pilots and the daily practice of Dutch cultural heritage institutions and researchers. One important observation was that Oral History

collections are managed in many different ways: from very adequate to not at all. There are a few 'forerunners' that do already make use of various professional infrastructures and technologies for disclosure and access but the larger part of the collection owners, in spite of acknowledging the virtues of modern access technologies, often do not have the knowledge or means to really start using these. In practice, access to Oral History collections in general is very limited and publically accessible overviews on available collections are missing. On the other hand, we see that collection owners and researchers have very specific requirements with regard to management, disclosure and access, and have very detailed knowledge of their collections and their contexts.

Knowledge transfer

In order to be able to catch up with the advantages of the digital networked society, cultural heritage institutes need practical handles that are specifically geared towards their specific use cases. The Verteld Verleden project follows this practical approach by mapping the available solutions and best practices in The Netherlands on a diversity of relevant topics ranging from digitization of audio-visual data, format conversion, online access to collections, and (semi)automatic metadata generation, and linking collections to other information sources, to dealing with privacy/copyright issues. On an academic level, special attention is addressed towards transfer of knowledge and awareness on methodology and theory of Oral History research, and the design of Oral History research in combination with modern technology.

Showcase

To showcase how these best-practices and solutions could work out in practice the project builds an embeddable Oral History portal that enables access to distributed oral history collections. The general approach is that collections owners are urged to comply with interoperability standards on the dissemination of metadata and content. By adopting these standards, content owners allow aggregators to channel content into local portals (e.g., Verteld Verleden) or even international portals such as the Europeana portal [9]. The Verteld Verleden portal serves here as a so called thematic portal.

Verteld Verleden promotes OAI-PMH, the Protocol for Metadata Harvesting and stimulates content owners to have their content available using a streaming media protocol to enable play-out of search results. The Verteld Verleden portal harvests metadata from associated institutes and provides

centralized search for searching and browsing the collections that are linked up. As the portal's user interface can also be embedded in the local websites, content owners can be provided with search functionality for their own content.

State-of-the-art

The portal is equipped with state-of-the-art search technology and a flexible user interface that allows the project to adapt it easily to the requirements of researchers and content owners that are expected to advance during the project as a result of discussions at workshops and local expert sessions. An important requirement of researchers is evidently to have sophisticated means to access and analyse available Oral History collections. To a large extent however, access to collections is rather limited due to the lack of appropriate *fragment-level* semantic descriptions. Metadata is often only sparsely available, forcing scholars to play an A/V item in full in order to decide if, and if so, which parts of the material are of interest for their research. Moreover, exploring possible *correlations and connections* both *within and across large data collections* requires an additional layer on top of the metadata for the interlinking of multimedia content sources and/or collection fragments. Ultimately, also dedicated technology for *browsing, accessing, analysing and comparing sources* effectively during the various phases of research (exploration, analysis, publication, verification) are a prerequisite for the innovation of the methodological framework of humanities researchers and for the formulation of new questions and the renewal of research agendas. In order to successfully exploit these technologies for the purpose of humanities research, their development must strongly be steered by the demands and requisites of the researchers and their research paradigms.

Speech Recognition

A special role in the project is assigned to the use of speech recognition technology. Speech recognition can play an important role in the process of making Oral History content better accessible, either directly via the conversion of speech to text or indirectly using available textual transcripts and a technology derived from speech recognition often referred to as forced-alignment. Verteld Verleden offers associated content owners the use of a speech recognition service supported by the Dutch CATCHPlus program [10], that aims to valorise scientific research results to usable tools and services for the Entire Dutch heritage sector.

Deploying speech recognition brings up an additional challenge with regard to metadata models and harvesting standards: it encompasses the need to incorporate time-labelled into the metadata model. Approaches are currently investigated in close collaboration with CLARIN-NL, a project on Common Language Resources and Technology Infrastructure [7].

References

[1] Goldman, J., Renals, S., Bird, S., de Jong, F. M. G., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D. W., Stewart, C., & Wright, R. (2005). Accessing the spoken word. Int. Journal on Digital Libraries, 5(4), 287–298

[2] F.M.G. de Jong, D.W. Oard, W.F.L. Heeren and R.J.F. Ordelman Access to recorded interviews: A research agenda, ACM Journal on Computing and Cultural Heritage (JOCCH), 1(1):3-29, ISSN 1556-4673, 2008

[3] R.J.F. Ordelman, W.F.L. Heeren, M.A.H. Huijbregts, F.M.G. de Jong and D. Hiemstra Towards Affordable Disclosure of Spoken Heritage Archives, Journal of Digital Information, M.A. Larson, K. Fernie and J. Oomen (eds), 10(6):17-33, ISSN 1368-7506, 2009

[4] Radio Oranje Demonstrator (alignment of historical speeches): http://hmi.ewi.utwente.nl/choral/radiooranje.html

[5] Searching interviews bombarding of Rotterdam: http://www.gemeentearchief.rotterdam.nl/brandgrens/navigator/interviews.php

[6] Access to interviews with survivors of World War II concentration camp Buchenwald: http://www.buchenwald.nl

[7] CLARIN-NL: http://www.clarin.nl

[8] Verteld Verleden homepage: http://www.verteldverleden.org

[9] Europeana portal: http://www.europeana.eu/portal/

[10] OAI-PMH: http://www.openarchives.org/pmh

[11] CATCHPlus program: http://www.catchplus.nl/en/