Data documentation, modeling and analysis of a digital oral history archive: “The Connected Histories of the BBC” project as a case-study

Dr. Anna-Maria Sichani, University of Sussex

This paper aims to discuss various data-related challenges around the development of a digital oral history archive, using the “Connected Histories of the BBC” project as a case study.

The “Connected Histories of the BBC” project is an interdisciplinary project towards the creation of a digital oral history archive of the BBC. The project is hosted at the University of Sussex (School of Media, Film and Music & Sussex Digital Humanities Lab), will run for nearly five years in the lead-up to the Corporation’s centenary in 2022, and is being funded by a grant from the Arts and Humanities Research Council (AHRC-UK). The project is also supported by key partners in the field: the BBC itself, the Science Museum Group, the Mass Observation (MO), and the British Entertainment History Project (BEHP).

The project aims to the tell the story of the BBC directly through the voices of the people who worked there. Their oral interviews offer unique accounts of how the BBC has developed the arts of broadcasting and seen the world of politics and culture. Yet, not only are they inaccessible to all but a select few; they are also unusable - scattered, un-catalogued, preserved in multiple formats, from videotape, audio files to crumbling paper.

This project aims to enrich Media History and Oral History with Digital Humanities methodologies and tools: from digitisation of the oral interviews, data management, curation and structuring to innovative data analysis techniques, in order to create a new digital catalogue of the entire oral history collection. Furthermore, the project seeks to employ 'linked open data' (LOD) technologies that would allow the diverse dataset to be interconnected with other resources. The resulting digital catalogue will allow historians, scholars and the general public - with their own memories of the BBC - to search for the first time ever this archive for a myriad of links between people, places and events, spanning decades of broadcast history.

By bringing together recent developments on the areas of data documentation, structuring, modelling and analysis, we are informing our technical development procedures with current research trends from the fields of Digital Humanities, (Audiovisual) Data Curation and Digital Cultural Heritage

This paper will focus on three pivotal research and development areas of the project related to the digital oral history archive of the “Connected Histories of the BBC” project, by discussing mainly the challenges and the related decisions on these fronts:

1. **Data documentation**

The BBC oral history archive, currently containing a variety of files and formats, needs to be digitised to highest standards, documented and tagged in such a way that will allow its further processing. Issues of metadata standards for various data formats and qualities as well as preservation standards need to be addressed too.

2. Data modeling

A robust data modeling strategy will provide the project team with the conceptual tools in order to describe more accurately data structures, relationships and data semantics. By adopting a curation and research-driven data modeling strategy, a linked open data framework will be developed as well as a domain-specific ontology for oral history.

3. Data analysis

The project aims to develop and apply a set of sophisticated data analysis techniques for the oral history interviews, such as entity extraction, data mining, topic modelling etc. One of the biggest challenges for the project team is to move beyond the text-based analytical approaches to oral history collections, and to experiment with audio feature analysis and Music Information Retrieval (MIR) methodologies in order to extract semantic information from the oral history dataset.