# A Case Study of Automated Curation of Digital Archives

**Fabiola Hanna**
fhanna@ucsc.edu
UC Santa Cruz, United States of America

Access to digital archives has been well problematized in recent years. For example, should one have access to an archive by default or should one belong to a community in order to gain access to that knowledge,– such as with Mukurtu CMS which builds on knowledge heritage in indigenous groups (Christen 2007)? Another thread with regards to access and archives is the trend of dumping all the data, and claiming that because of this, the individual and/or the organization is somehow more transparent (as seen with various initiatives such as data.gov). But many have also shown that there is no raw data (Gitelman 2013). I propose to build on these two threads in order to argue that 1) paying attention to the medium and its parameters is important, and 2) that archives need to be sifted and curated in order for them to be properly accessible. I will illustrate both of these arguments with my project, *We Are History: A People's History of Lebanon*, which is in its final stage of development and will have been released publicly before DH2017.

Digital Humanities (DH) projects are, in varying degrees, led by the desire to engage with a wider public. Some often include actions such as inviting participants to share their stories, images, audio clips, drawings, and videos. Many DH projects place these contributions in an audio or video database displayed in full on a webpage, not unlike oral history transcripts ending up in a dusty closet. To build on one of the most successful digital storytelling projects, it is useful to examine the Storycorps team, who have been collecting ordinary oral histories in video form and archiving them in full as a record of American history using booths and a mobile application. The Storycorps team knows very well that if it did not curate and edit together shorter versions, then few users would listen to the longer interviews in full, let alone several at a time.

Francis X Blouin and William G. Rosenberg trace the intersection of history and archives and found that Ranke, during the Enlightenment, conceptualized history as a scientific endeavor in that truth could be extracted from archives through rigorous methodologies. This led to the idea that documents could "speak for themselves" (Blouin, 24), as if simply making documents available, without providing context of any sort, would reveal their inner truth. This is one of many cases where the reading of documents is taken for granted. It also ignores the effect that archivists have on the collecting, saving, and indexing of documents. Influential archivists such as Terry Cooke, Richard Brown and Brian Brothman have brought about new attitudes to repositories with an acknowledgment of the effect that archivists have on documents (as quoted by Cox, 33). This relatively recent push in archival theory, therefore, points to the flaws in the claim that documents on their own can represent themselves: that would be ignoring all the various power relationships at play, as well as the medium itself in which the data is codified.

This is more directly seen in the tagging of videos and their categorization without additional interpretational work, such as in the Oral History Metadata Synchronizer (OHMS) tool developed at the Louie B. Nunn Center for Oral History University of Kentucky Libraries. This *will to not "add"* to these stories seems to come from the premise that these testimonies should "speak for themselves"; that no added interpretation is needed, even that any added interpretation distracts from the directness of the stories. But this often also means the medium and its effects on these stories are not carefully examined.

In pursuit of generating communal dialogue in the context of inability to have conversations about our contested history in Lebanon, I set out to build an Artificial Agent that would sift through an oral history video archive of testimonies of daily life with the task of figuring out common threads, sometimes confirming and sometimes contesting each other, and automatically editing many different versions of possible histories. This automatic montage machine addresses two problems in the Lebanese context: first, it circumvents the tiring accusation of being biased since a machine is now the moderator (presenting a multiplicity of stories might be the closest one can get to strategic objectivity) and second, it opens up the possibility of conversation by weaving various and often opposing perspectives in order to start imagining what our histories could look like. The project, which would reside online as well as in booths in public spaces across Lebanon, invites people to listen to an automated montage of oral histories and to then share their own stories and memories. Each newly contributed story is added

to the archive, analyzed using new developments in computational corpus-based linguistics, automatic story generation, and social computing, and tagged with its transcript, which enables the interface to incorporate newly added video interviews into the pool concerning the event discussed