Cultural Archives: Enhancing Materials, Collections, Formats, Data for Culture Analytics

Peter Broadwell (UCLA Digital Library)
Goki Miyakita (Graduate School of Media Design, Keio University)
Lindsay King (Yale Arts Library)
Mehrdad Yazdani (Qualcomm Institute at UCSD and the Open Medicine Institute)
Ekaterina Lapina-Kratasyuk (National Research University Higher School of Economics)
Timothy R. Tangherlini (UCLA Scandinavian Section)

"We turn our own lives into an information archive by storing all our emails, SMS, digital photos, and other "digital traces" of our existence." - Lev Manovich, 2005

Introduction

Culture emerges from human interactions with each other, with the environment and with ideas, and is recorded in the entire continuum of human expression, from literature, art, and drama, to the documentation of everyday life in tweets, advertisements, and patterns of web activity. At times, these expressions are consciously preserved, while at other times they are not. The "stuff" of culture analytics constitutes these vast, disparate archives and, as such, represents a different conceptualization of what has traditionally been conceived of as "the archives." Yet conceptualizing archives as data types in a broad, integrated fashion provides a sound basis for analytic work that considers the cultural processes and the outcomes of these processes that shape our world.

In order for Culture Analytics to succeed, scholars need to be able to access, discover, and traverse these vast archives of cultural data in a coherent manner. This challenge recognizes the heterogeneous nature of such data, and that data appears in many different forms, in many different languages, and often in many different formats. There are several second-order challenges that are fundamental to the challenge of archives in Culture Analytics. For example, data preparation and data resource discovery remain significant barriers to current work. Ideally, future scholars will be free to use the right datasets for their research questions, not just the most conveniently available or correctly formatted datasets.

Challenges

The archives group identified six core challenges related to data preservation, access and storage for Culture Analytics. Each of these challenges are interrelated and recognize the dynamic nature of culture, which is always in a state of change. The explosive growth in both new expressive forms and the digitization of historical resources for the study of culture means that these questions are central to the success of Culture Analytics.

1. How do we alleviate the bottleneck of data preparation and data discovery to facilitate data-driven culture analytics?

Creating curated cultural datasets will facilitate data-driven culture analytics by addressing some of the common pitfalls in data selection. In machine learning, having well-curated datasets has been an important factor in recent advances made in algorithms (cf. "Datasets over Algorithms"). Machine learning has benefitted from having many "baseline" datasets for students to use in learning to apply existing methods and for researchers to test and develop new methods. Similarly, since data and computational methods are such an integral part of culture analytics, establishing relevant datasets that pertain to cultural questions will allow researchers to quickly iterate over different computational methods to advance the field. For example, prior to embarking on an analysis of a set of visual art images, the images would be vetted for color fidelity and authoritative metadata. (cf. Artistic AI: Analyzing Visual Culture Through the Eyes of a Machine).

2. How can we combine text, 2D and 3D images, video and audio in the archival realm? Typical computational methods deal with a small range of specific media and data structures (networks, text documents, images, etc.). Cultural processes and the products of those processes are often far more diverse, and require that archives be able to store the heterogeneous resources. For example, on social media websites, social interactions take place over a graph and the interactions themselves are expressed through content such as texts, images, and videos. We need to establish computational methods that combine these heterogeneous data sources in a robust, consistent and predictable manner. Recent advances in deep learning, particularly dealing with images, languages, and time-series data, suggest a promising framework. For example, new image caption algorithms provide an example of a model of combining text and images (cf http://cs.stanford.edu/people/karpathy/deepimagesent/).

3. How can we make non-English data as accessible as English data? Globalization and advanced internet infrastructure have made it possible to access datasets around the world. Nevertheless, there are still many walls between nations, including language and, interestingly, aesthetics. Making multilingual datasets available to multilingual audiences will allow for considerable gains in the understanding of cultural processes that span linguistic areas, as well as those that are related to one or another linguistic group, be that a current group or a historical

group. Given the promising advances in recent years, deep learning methods for language translation may emerge as a central technology for facilitating this type of work.

4. How do we integrate historical, non-digital archival resources with born-digital and hybrid archival resources? Scholars need to be able to move seamlessly between formats and types of data that address a given research question. Tracing influence on and reception of a living artist's work, for example, might involve looking at personal networks in two formats: reading handwritten letters as well as considering Facebook postings, in order to preserve original context within the archive. Without that context, the mode of communication could become a confounding factor.

5. How do we integrate proprietary data sources (restricted by copyright, vendor license, or privacy concerns) with public data sources? Ever larger amounts of cultural data are being digitized by libraries, and also by corporate vendors. These data sources may be restricted by privacy concerns (as with Holocaust testimonies), by copyright (as with contemporary authors or artists), or by vendors (as with historic magazine or newspaper collections), but they represent part of the cultural landscape that needs to be traversed as well. Considerable effort will need to be expended in addressing these issues, to ensure that research is not unnecessarily hampered by overly restrictive access regimes, while recognizing the significant issues related to privacy (Sandvig 2016).

6. How do we facilitate discovery of unofficial or underground archives that may counteract prevailing narratives? In closed or semi-closed societies or during wartime, official archives are often inaccessible and/or neglected. To balance this situation, participatory archives are flourishing, but their audience is limited and the data collected about past or contemporary events is not a focus of decision-making processes or academic research. Can we find efficient ways to combine crowdsourced efforts and professional ways of collecting data? Moreover, is there a possibility to provide the community with tools which allow them to make the data collected in the crowdsourcing process more visible to people researching in archives?

Future directions

Culture Analytics as an emerging field is enriched by a diversity of scholars having access to rich historical data. The development of accessible digital archives lessens the digital divide, while collective sourcing empowers individuals to work on public collections and the building of interoperable datasets and tools, particularly those that are based on open-source software and Linked Open Data. An extraordinary example of this can be found in the work of the New York Public Library (cf Human-Computer Collaboration at NYPL Labs). Democratization of access opens up the field to scholars at varying levels of technological skill and institutional funding. A scholarly community of practice for Culture Analytics would aim, in turn, to move from

constellations of discrete digital projects to networks of shared applications and data within and across disciplines.

Already we are seeing exciting developments in these areas. Preservation of cultural materials enabled by rapidly mobilized technological tools for archiving allows real-time analysis of cultural occurrences along with the archiving of current events as they happen. For example, we can conceive of social-media discussions about ISIS as a cultural archive to be analyzed as events are unfolding and later, when events have passed into history, we can see the value of synchronic and diachronic analyses of this kind of data. Similarly, easily deployed mobile technology taking advantage of mobile phone cameras and GPS can be used to help document imperiled archaeological sites or artwork, be the threat from natural disasters or social unrest.

Some of these archival challenges have computational solutions, while others require social or institutional changes. The awareness of disparate types of data in multiple archives will be key to future scholarship.

Relevant presentations from the Culture Analytics workshops

All of the presentations and discussions throughout the Culture Analytics long program involved “archival” materials in one form or another. Yet because presenters focused largely on analytical techniques (in keeping with the theme of the program), the role of the archive was not strongly foregrounded in most talks. Even so, the opportunities inherent in these projects as well as the archive-related barriers they encountered pointed to the key challenges for developing next-generation cultural archives formulated above.

The following list highlights presentations from the Culture Analytics workshops that explicitly presented new archival techniques.

Crowdsourcing and advanced public access interfaces for archives

Amanda Visconti (Purdue University)
Reading with the crowd: InfiniteUlysses.com & designing social interfaces for public literature participation
Sebastian Chan (Australian Centre for the Moving Image)
The intersection of museums, collections, interfaces and visitor data
Mauricio Giraldo (New York Public Library)
Human-Computer Collaboration at NYPL Labs

Advanced techniques for data discovery and description, focusing on images

Taylor Arnold (Yale University)
Recommender Systems for Digital Archives
Lindsay King (Yale University)
Data Mining beyond Text in the Vogue Archive

Improving processes of data discovery and preparation, attempted multilingual resource development

Peter Broadwell (University of California, Los Angeles (UCLA))
Library Systems and Paradigms of Cultural Data

Interrelationships of the archival challenge with other areas of Culture Analytics:

- Multilingual description, translation and correlation across archives should be addressed by advancements in **language** processing.
- Access to archive materials needs to support analyses at all scales: macro (accessing and analyzing multiple sources across institutions and archives), meso (integrating local communities of archives and collections), micro (close reading of specific items, access to digitized versions of the original item). The **multiscale/multiresolution** focus area of Culture Analytics will continue to develop approaches and paradigms to address this aim.
- An intensive focus on creating practices for **translational data science and collaboration** across data types should help address the challenges of working with multimodal data in archives and correlating text, images, video, sound.
- Similarly, work on advanced **interfaces for data analysis** will provide improved models for archival data access, although **cross-archival search and discovery** is primarily in the purview of the archives focus area.