The Media Ecology Project: Using Linked Data to Support Distributed Analysis of Visual Culture

John Bell

New Media Department
University of Maine
Orono, Maine
john.bell@umit.maine.edu

Mark Williams

Department of Film and Media Studies
Dartmouth College
Hanover, New Hampshire
mark.j.williams@dartmouth.edu

Abstract. The Media Ecology Project (MEP) is an architecture, methodology, and partnership network that facilitates new forms of scholarly production based on distributed analysis of moving image materials. Based on standards like the W3C Open Annotation Data Model, MEP seeks to connect a variety of specialized tools together using a third-party metadata repository that sits outside of any single tool or video archive. Doing so serves several purposes: it allows existing software platforms to be used in a more open, collaborative manner; it maintains the ability of individual tools to specialize in focused areas of functionality without the pressure to serve all needs for all users; and it frees metadata created for a specific analysis to be shared with other scholars. A critical outcome of this model is that it allows metadata created by subject area experts outside of source archives to contribute metadata they generate about an asset back to the archive itself, expanding the archive's contextual knowledge about its own materials. The first stand-alone tool developed for MEP, Onomy.org, is available now at http://onomy.org/ .

Keywords. Linked Data, Open Annotation, time-based annotation, RDF, scholarly collaboration, video archives, Scalar, Mediathread, Onomy.org, Media Ecology Project

1 INTRODUCTION

The Media Ecology Project is a digital resource at Dartmouth College which is creating a dynamic environment in which researchers digitally access archival moving image collections and contribute back to the archival and research communities through the fluid contribution of metadata and other knowledge.

Our moving image heritage is at enormous risk. Moving image archivists and digital repository advocates are developing solutions to these problems, but we cannot sustain interest in "preservation" without a better sense of the historical value of these materials. "Access" is not enough; new knowledge production is required in order to connect archival materials with audiences and prompt preservation and access efforts. MEP is working to produce cooperation and motivated engagement surrounding archival materials within academic communities by building new software tools and connections that encourage novel modes of scholarly production.

MEP is designed to create a metadata loop by which scholars and archives both benefit from the critical work being done by each other. The MEP metadata loop begins with an archive making media files available online. When scholars access those files they are linked to by a variety of tools that allow new knowledge about them to be created through annotation and tagging. Those annotations, created in a variety of distributed software clients, are then aggregated back to a centralized metadata server. Once on the metadata server the annotations are available for use in other clients, but more critically, they are also available to be harvested back into the original archive, thus closing the loop and increasing the overall available knowledge about that media file.

2 MEP TOOLSET

2.1 MEP Metaserver

The technological heart of the Media Ecology Project is the MEP Metaserver. The MEP Metaserver is an indexing system that accepts Open Annotation metadata about 3rd party media files. The MEP Metaserver accepts annotations about any media file that has a canonical URI so long as the annotation is submitted through an API call from an authenticated client system.

The MEP Metaserver is designed for flexibility and data exchange. In order to increase its ability to interoperate with a variety of host archives and client software packages, it only stores minimal metadata about each media file it indexes. Instead, it acts as a clearinghouse for annotations and implements the concept of annotations as an open-ended data structure.

Annotations may contain text comments, named tags, and provenance information about their own source. However, because the text comments are themselves unstructured, they may contain whatever metadata any client software package wants to define. For example, one application that has been explored in a pilot project is creating annotations containing machine-vision derived gestural classifications; not the traditional use of an "annotation" field, but one which allows for very granular descriptive metadata to be shared without the need for a new metadata specification. Clients that do not understand these annotations can simply ignore them while clients that do understand them can take advantage of a new pathway for data exchange.

The MEP Metaserver publishes its records as RDF XML and uses a record structure based on the W3C Open Annotation Data Model [1]. It supports FOAF identification of users for annotation provenance and controlled vocabulary tagging based on

externally defined RDF taxonomy files. Annotations are versioned on the server, with previous versions maintained at permanent URIs.

2.2 Onomy.org

Onomy.org is the first stand-alone tool designed by MEP to facilitate working groups sharing annotations across software platforms and will be demonstrated as part of the Linked Media Workshop at the European Semantic Web Conference 2014. It is a web service that allows collaborative development of taxonomies and folksonomies for use as controlled vocabularies. Onomy.org supports hierarchal categories and definitions, API-based term suggestions, and export in HTML, RDF, and JSON formats [2].

Shared vocabularies are a simple but integral aspect of the distributed toolset the Media Ecology Project is enabling. While full-text annotations can be associated with media files using Open Annotations, full-text presents difficulties for certain computational analysis techniques like entity identification and topic analysis. Tags based on a shared controlled vocabulary enhance metadata and prepare annotations for a variety of network analyses in different software packages, increasing their value as a flexible knowledge resource. By creating a shared vocabulary that exists independent of any one analysis tool, Onomy.org creates a resource that allows many different tools to speak the same language and behave as a true read/write ecology across platforms.

2.3 Connected Platforms

MEP is modifying 3rd party software platforms to connect to the MEP Metaserver as clients that can read and write metadata about media files available over open networks. Future work includes integration with the Shibboleth federated identity system to allow greater access to rights-restricted archive materials. The initially-supported platforms, Mediathread and Scalar, were chosen because they both focus on adding commentary to and displaying externally-linked media files rather than uploading files to a closed content management system.

Mediathread. Mediathread is a Django-based classroom platform developed by the Columbia Center for New Media and Teaching [3]. It is designed to allow students to collaboratively annotate web-based media files and integrate those files with essay text. When integrated with the MEP Metaserver, Mediathread provides an environment that supports small group analysis of video content that can be dispatched to other platforms for publication and reuse.

Scalar. Scalar is a semantic web-publishing platform developed by the Alliance for Networking Visual Culture [4]. It presents a simple platform for authors to create online books built from linked pages containing integrated text, media, and annotations. Each page is a semantic object than can be assembled in a variety of non-linear paths within the book or accessed through an API for external reuse in a variety of rich-media interfaces.

2.4 Metadata Loop

The MEP metadata loop connects different components of the MEP toolset to each other and to host archives. For example, one instance of the metadata loop workflow might see a user find a video on the Library of Congress web site. They hit a bookmarklet that grabs the URL of that video and imports it into Mediathread, where they create annotations describing the video. When those annotations are made public, Mediathread updates the MEP Metaserver with them. Other users in Mediathread or Scalar that query that URL will be able to load their annotations and the Library of Congress can harvest them to update their own records on the file.

3 PILOT PROJECTS

3.1 Library of Congress Paper Print Collection

In conjunction with the United States Library of Congress, MEP is developing a project regarding scholarly analysis of early silent film era materials with an emphasis on the historically significant Paper Print collection. The Paper Print collection consists of films dating from 1894-1915 and encompasses all period genres [5]. Scholars including Prof. Tami Williams (University of Wisconsin at Milwaukee), Philippe Gauthier (Universite de Montreal and Harvard University), and several additional members of the DOMITOR research society will engage the Paper Print materials using software platforms and ontologies connected by MEP for this pilot study, with the goal of publishing commentary on the films in Scalar. The Library of Congress has provided a first batch of 32 Paper Print media files with related metadata for use in this pilot study and will continue to supply additional titles as the project proceeds.

3.2 In the Life

A second pilot study will focus on an important public television program called In the Life, which assays the history of gay and lesbian lived experience in the United States [6]. The entire run of that program, plus all of the associated materials involved in the production of that program, will be provided and placed online by the UCLA Film and Television Archive. MEP has begun to assemble a group of prominent scholars to work on these materials, including Prof. Matthew Tinkcom (Georgetown University), Prof. Michael Bronski (Harvard University), and Prof. Stephen Tropiano (Ithaca College). The UCLA archive anticipates that the programming materials will start to be made available online in 2014.

3.3 Newsfilm Archives

The third pilot study involves the participation of multiple archives and is dedicated to providing more and better access to historical news materials: newsreels, news telecasts, news film, and other associated footage. Archives who will participate include WGBH in Boston, The UCLA Film and Television Archive, The University of South

Carolina, The University of Georgia and the Peabody Award Archives, Northeast Historic Film in Maine, and the Library of Congress. The core group of scholars dedicated to this pilot includes Karen Cariani (WGBH Archive), Prof. Mark Garret Cooper (University of South Carolina), and Prof. Ross Melnick (University of California at Santa Barbara).

4 Conclusion and Future Work

The notion of ecology is central to the Media Ecology Project in several ways. Those who work on media history recognize all too well that the materiality of historical media is fated. These historic materials simply will not endure without active preservation. In a fundamental sense this is a sustainability project: we are working to protect and ensure cultural memory in the form of historical media collections.

Inherent to any idea of ecology, though, is continual adaptation. MEP has been designed as a flexible medium capable of supporting a variety of emerging research interests and methodologies. One use of the MEP Metaserver that has drawn repeated interest is the capacity to facilitate networked "big data" applications of moving image metadata. Discussions about MEP with stakeholders including the Northeast Historic Film Archive, The Internet Archive, and the American Archive have revealed an interest in using the MEP Metaserver to store, update, and distribute information on collections ranging from hundreds of thousands of items to tens of millions of items. Conversely, machine vision applications have the capacity to produce large volumes of deep, granular metadata about small subsets of collections. The MEP toolset could support both methodologies and both would lead to unique types of interpretation and scholarship. The Media Ecology Project will pursue 21st-century pedagogies and research procedures that would contribute to the development of interdisciplinary approaches to visual literacy in relation to media history. In addition to extending the research profile of MEP as a networked resource, this will facilitate the widespread production of qualitative metadata that can support the essential work of the archives.

5 References

1. Sanderson, Robert, Ciccarese, Paolo, Van de Sompel, Herbert (eds.) (2013) "W3C Open Annotation Data Model", consulted March 6, 2014. Available: http://www.openannotation.org/spec/core/20130208/index.html
2. Bell, John (2014) "Onomy.org", consulted March 6, 2014. Available: http://onomy.org/
3. Columbia University Libraries/Information Services (2013) "Mediathread", consulted March 6, 2014. Available: http://ccnmtl.columbia.edu/digitalbridges/projects/mediathread.html
4. Alliance for Networking Visual Culture (2013) "About Scalar", consulted March 6, 2014. Available: http://scalar.usc.edu/scalar/
5. United States Library of Congress (2010) "Motion Picture & Television Reading Room", consulted March 6, 2014. Available: http://www.loc.gov/rr/mopic/earlymps.html
6. UCLA Film & Television Archive (2011) "'In the Life' Collection", consulted March 6, 2014. Available: http://www.cinema.ucla.edu/collections/life-collection