Visualizing Television Archives

by Courtney Michael, Mayo Todorovic and Chris Beer

This article will outline efforts by the WGBH Media Library and Archives (MLA) and the WGBH Interactive team to create online access points into our multimedia archive using innovative visualization techniques that create visual alternatives to traditional search, browse and faceted navigation and that highlight WGBH's rich media collections. The main objective in this effort is to allow scholarly researchers to access the collections, search and pinpoint useful records and discover related items they may want to explore. It has been an interesting challenge to strike a careful balance between allowing for targeted search and passive browse while also providing for serendipitous discovery opportunities. The depth of metadata essential for our expert audience led to our use of linked data in our cataloging processes, which enabled further experimentation with visualization.

Background

WGBH, Boston's PBS station, produces over one-third of nationally broadcast PBS television programming. The WGBH archives, therefore, hold hundreds of thousands of hours of moving image content, as well as thousands of linear feet of related documentation and still images. Not only do researchers find finished documentary films from our flagship productions (Frontline, NOVA and American Experience), but also all of the production elements that went into the making of these films.

This scale of acquisition makes for an extremely robust and yet complex set of collections. Like other media archives, the barriers to access for researchers in the WGBH collections include intellectual property constraints, unreadable or fragile media carrier formats and unfamiliar access tools. While the WGBH MLA has experimented with offering traditional finding aids, media collections are not well served by this hierarchical, text-based and largely linear format. Internally, the MLA operates through item and shot-level databases, but these tools are not useful to external researchers without a steep learning curve and significant investment of time. Therefore the MLA has experimented with repository-based sites that bring together the searchability of an electronic finding aid and the multi-dimensional orientation necessary to understand our collections.

All three authors are affiliated with WGBH in Boston. Courtney Michael is project manager and can be reached at courtney_michael<at>wgbh.org. Mayo Todorovic is an interactive designer and can be reached at mayo_todorovic<at>wgbh.org. Chris Beer is web developer and can be reached at chris_beer<at>wgbh.org.

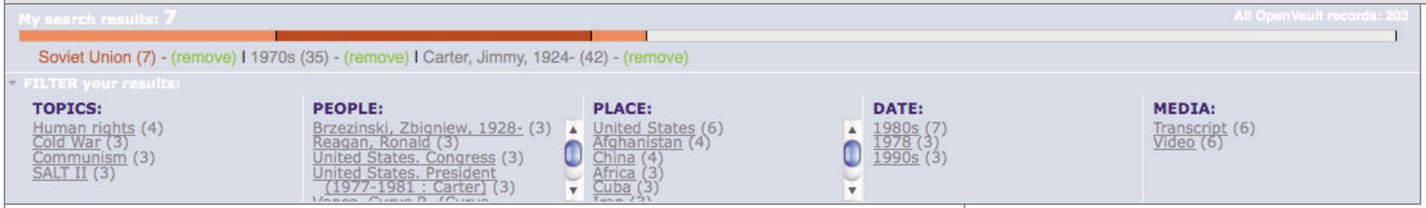
In 2006 the WGBH MLA launched *Open Vault* [1], a searchable and browsable site that highlights four television series in our archives. This site includes video clips and transcripts from the elements that went into the making of the series and is aimed primarily at educators and the general public. As the central presence of the archives online, it has also greatly increased the WGBH archives' visibility to scholarly researchers.

The MLA has found, however, that the scholarly research audience has very different needs from those of the general public and educators. At the most basic level they seem to want it all, rather than to be hand-held or guided through a collection. They want to see an entire collection in order to ensure that their research is thorough and want to view full-length archival videos, not just pre-selected or curated clips. In short, they want to do the data mining and curation work for themselves.

In 2008, with funding from the Andrew W. Mellon Foundation, the MLA and WGBH Interactive developed a small-scale prototype [2] to discover how academic researchers would like to find and use our materials. A password-protected site was shared with selected scholars in order to

NEXT PAGE >

FIGURE 1. Results bar and facets



My search results: 7 All OpenVault records: 203

Soviet Union (7) - (remove) I 1970s (35) - (remove) I Carter, Jimmy, 1924- (42) - (remove)

FILTER your results:

TOPICS:
Human rights (4)
Cold War (3)
Communism (3)
SALT II (3)

PEOPLE:
Brzezinski, Zbigniew, 1928- (3)
Reagan, Ronald (3)
United States. Congress (3)
United States. President (1977-1981 : Carter) (3)

PLACE:
United States (6)
Afghanistan (4)
China (4)
Africa (3)
Cuba (3)

DATE:
1980s (7)
1978 (3)
1990s (3)

MEDIA:
Transcript (6)
Video (6)

solicit their feedback, and further user tests continue. The lessons the WGBH team is learning through this experiment in communicating large amounts of data to humanities scholars could well inform others facing similar challenges.

Visualizing our Data

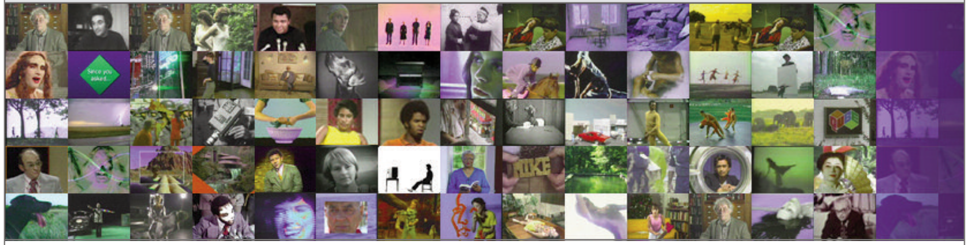
The first visual technique we employ to communicate the depth of the data is a subtle one. In an attempt to orient researchers to the collections and to the diversity of our materials, we expose the faceted navigation system provided by the Solr search index. This view allows users to see not only the results of their search but also the underlying information about their aggregate result. They can then select different facet attributes to narrow their results set. While facets are not generally perceived as data visualization, they orient the user through visual cues to enhance understanding of the content.

Results bar. To the faceted navigation we have also added a visual results bar (Figure 1) that acts as a breadcrumb trail through the evolving found set. As the user applies additional facets, our results bar compares the relative size of the resulting set to the user's original search results and allows the user to remove previously applied attributes in non-linear order. In this way the user is acclimated to the scale of our collections and encouraged to interact with and learn about them even as they mine them.

Faceted classification works well for the WGBH collection because the materials are multi-leveled and multi-faceted. It allows users to narrow and expand their found sets within increasingly familiar frameworks. What the faceted navigation lacks, however, is a multi-dimensional view that illustrates the relationships between materials and provides a visual representation of the media archive.

The mosaic. As can be seen on the homepage of our current *OpenVault* site, we sought to communicate visually the diversity of materials within our unique collection. As shown in Figure 2, our mosaic design uses the thumbnail images representing individual videos within the collection. On rollover the user can discover additional metadata about each record and can click on the image to go directly to the record. Initial reactions to this design were very positive as it is colorful and playful and does the job of communicating the idea that the WGBH archives has a wealth of interesting materials.

FIGURE 2. Mosaic



Academic researchers, however, seem to want something more structured and useful to their work. While they appreciate the aesthetics of the mosaic, they want additional information up front. When shown a different mosaic categorized by people, place and television series they reacted positively because it communicated more information. We took it one step further, however, when they shared with us their goal of quickly making time management decisions regarding archival research.

The relationship map. Scholars want to quickly find out what we have, how they can access it and if a particular record is worth their time to follow through. For example, a researcher searching for information on President Kennedy's actions during the Cuban Missile Crisis may be excited to come across an hour-long interview with Robert McNamara, but they might soon

TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >

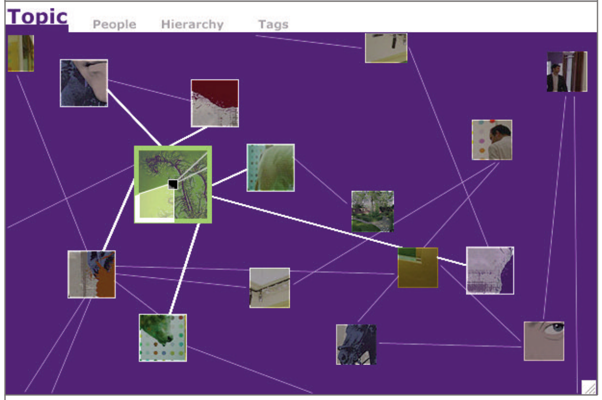
be frustrated or even angered when they spend an hour listening to the recording only to discover he never mentioned the event.

Responding to this need, we developed a visual graph we call a "relationship map" illustrating the depth of content in each asset and the explicit and implicit relationships among assets. While we are still working to implement aspects of the map and to collect the necessary metadata, the prototype's functionality of relating items and displaying these relationships has proven exciting to our users and is many steps closer to communicating the multi-dimensional aspect of our materials than the mosaic.

Figure 3 depicts the design concept for the map. The highlighted thumbnail image indicates the central record or the record the user is interested in. The user can see the connections from segments of the content within the pertinent record to other records, represented by the other thumbnails.

Our objective here was to visually dissect the video record, for example, into the different topics (or people or places) it covers. The segments of a pie chart (as represented in the highlighted thumbnail) represent the percentage of the video that address a specific topic. In this way, a researcher can understand visually the breakdown of topics within a record and make informed decisions about how much time to invest in an individual record. This functionality, not yet fully implemented, will be enabled by an additional cataloging process where we segment or chapterize video transcripts and assign topics and timecode to the chapters. We can then calculate the extent of the total video length from the associated timecode to construct the pie chart.

FIGURE 3. Relationship map

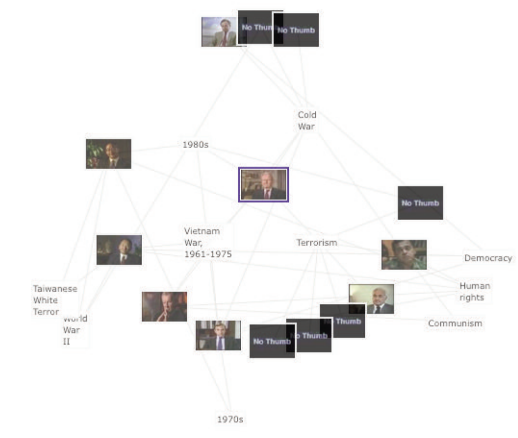


From the design concept, we have built a dynamic radial graph using the JavaScript Information Visualization Toolkit. The radial graph illustrates the relationships among and between records through the lenses of people, topic and hierarchy or collection context (Figure 4).

In this example, the central record is an interview with former National Security Advisor Robert McFarlane. Users can see that he discusses terrorism, Vietnam, the Cold War and the 1980s in general. They can go one step further and see what other records in the collection touch on these same topics. They can click on each thumbnail or topic to re-center the map and explore the relationships, and they can also zoom in and out to discover records located farther afield.

The intensive cataloging that enables a map like this one presents challenges as well. Records are cataloged to varying degrees, which may skew their relevancy and therefore their proximity values on the map. For example, a record for which we have little data may not appear on the map at all, giving the researcher a false picture of the collection. On the other hand, a record that has been fully cataloged may connect to dozens of other records, making for a crowded and hard to use map display. We are currently experimenting with automatically pruning certain ill-defined relationships in order to avoid overcrowding and overlapping data.

FIGURE 4. Radial graph



While we originally were excited about the relationship map serving as both a browse tool and a homepage feature, we have found that it tends to intimidate users when we reveal it up front on the homepage. The last thing we want to do is turn users away from the collection because they don't understand the homepage. As Jared Spool recently explained at the IA Summit 2009 [3], we need to minimize the "tool time" – or the time it takes

TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >

a user to figure out functionality and get their bearings – and maximize "goal time" – in our case the time allotted for finding value in the content. We may still incorporate the map into the homepage in a less intimidating way, but we recognize it should not be the first feature users encounter.

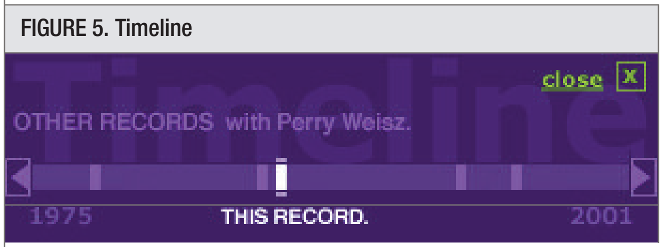
Another challenge we have pinpointed regarding the relationship map is a lack of contextual information. The user often gets lost within the map. We need to better communicate the relationships (the significance of the connecting lines) and the relevancy (or proximity) of one record to another. In some ways these challenges stem from too much data, in other ways they might be solvable with continued experimentation and refinement.

The lesson here is that, for our scholarly audience, the attractive yet random mosaic was "pretty" but not substantive enough. Yet the working relationship map, while it has the possibility of being very interesting, is not intuitive as currently implemented. We need to spend more time striking a balance between utility and usability for this feature.

All of these visualizations are made possible by the depth of cataloging we have for our prototype record set. We have cataloged to the item level and sometimes have gone farther to tag people, places, dates and topics within the content of a record. In the Robert McFarlane example above, a PBCore record describes the interview media asset, while a TEI-encoded transcript provides the links into the video content itself [4].

Other visualization possibilities we can explore with this depth of data include tag clouds, maps and timelines. We are looking into placing records in context along a timeline of life or collection dates that would look something like what is illustrated by Figure 5.

FIGURE 5. Timeline



This tool would serve two functions: it would allow users to see related records and would also show them the date range of similar records. In the example above, the user has found an interview with Perry Weisz. From this timeline the user can see that the archives contain four other records related to Perry Weisz, spanning from 1975 to 2001.

Conclusion

One of the first discoveries we made working with humanities scholars was the intense level of contextual information they require about a particular archival resource. We translated this need into a high level of cataloging at the item and even sub-item or shot-log level.

While they want this density of information made available to them, the scholars also want to be able to quickly pinpoint the exact content that will enhance their work. This paradox of "give me all your information" versus "give me only what I need" led us to experiment with alternative methods of communicating our metadata.

Sophisticated data-visualization techniques ideally allow us to save our users hours of time reading through pages of text. We have found that senior scholars, accustomed to spending those dedicated hours mining an archive, are intrigued and yet skeptical of some visualization techniques. As we continue to improve our website and refine our visualization tools, we anticipate that the next generation of digital natives, said to be more visually oriented and more extractive in their research behaviors, will embrace these tools as part of their research, analysis and scholarly production workflows.

Acknowledgements

This work was funded in part by the Andrew W. Mellon Foundation and the Institute for Museum and Library Services. The authors also wish to thank Karen Cariani and Peter Pinch, co-principal investigators, and our advisors, Drs. John Dower, Peter Winn, James Blight and Janet Lang. ■

RESOURCES on next page

TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >

Notes and Resources for Further Exploration

[1] The WGBH Media Library and Archives' OpenVault website: http://openvault.wgbh.org OpenVault for Researchers (closed prototype): http://openvaultresearch.wgbh.org

[2] The prototype is built on a Fedora repository architecture (www.fedora-commons.org/) and uses Solr, an open source search server (http://lucene.apache.org/solr/). The relationship map uses techniques from the JavaScript Information Visualization Toolkit (http://blog.thejit.org/javascript-information-visualization-toolkit-jit/).

[3] Jared Spool presented these ideas during his talk on "Revealing Design Treasures from the Amazon" on Saturday, March 21, 2009. Find an overview of his talk at http://iasummit.org/2009/program/presentations/revealing-design-treasures-from-the-amazon/ or read about Goal and Tool time at www.uie.com/brainsparks/2006/04/20/dividing-user-time-between-goal-and-tool/

[4] PBCore is an XML schema originally designed for the exchange of public broadcasting materials. Learn more about PBCore at www.pbcore.org/ and www.pbcoreresources.org/. TEI or the Text Encoding Initiative is also an XML standard used for encoding texts. We are primarily using elements from the "Transcriptions of Speech" guidelines. Learn more about TEI at www.tei-c.org/

The *Visual Thesaurus* uses relationship mapping to illustrate language connections: www.visualthesaurus.com/

The Many Eyes project allows users to play and experiment with data sets and multiple ways of visualizing them: http://manyeyes.alphaworks.ibm.com/manyeyes/. Also explore Martin Wattenberg's personal site highlighting his other data visualization projects: http://bewitched.com/song.html

Jonathan Harris' *Ten by Ten* illustrates news stories organized by date from three sources through image mosaics: www.tenbyten.org/

TOP OF ARTICLE

< PREVIOUS PAGE