How to Dodge the Born Digital News Memory Hole || Linking Born-Digital News and Social Media Collections via Automated Entity Detection and Authority Matching || Online-Only Media: 21st Century Collection Crisis?
Wednesday, October 28 • 9:00am - 10:30am

Session Leaders (include Twitter handles)
Peter Broadwell, @peterbroadwell
Martin Klein, @mart1nkle1n
Edward McCain, @e_mccain
John Vallier, @vallier
Frederick Zarndt
Sam Meister, @samalanmeister

Slides
https://docs.google.com/presentation/d/1YFhAg16x-_vVoAyxZGW2RrU6tmt3gVl3txNDaajFkeQ/edit?usp=sharing

https://drive.google.com/file/d/0B1aoGJ7nJiYSMEx4N3V2LU9MTUk/view?usp=sharing

Notes
Presentation 1
How to Dodge the Born Digital News Memory Hole
The year is 2104. For her Europa University sociology studies your great granddaughter must research the distribution of news through early 21st century social media like Facebook and Twitter, predecessors to MyUniverse, the largest communications and social network on Earth, Mars, and most of Jupiter's moons, including Europa. Unfortunately, even though her University is a founding member of MyUniverse with full privileges, she cannot find much news about social media from the early 2000s. She asks you to help. You must tell her why news from this time is scarce.

Sound farfetched? Unfortunately it isn't. The present state of copyright laws, the lack of born digital legal deposit legislation in this country and many others, and rapid technology changes in the production and distribution of news combined with neglect of

born digital news by cultural heritage organizations make your great granddaughter's research misfortune altogether too likely.

In a keynote at the Dodging the Memory Hole conference Clifford Lynch reminded us that news is important not only now but also in future as "the first draft of history". An accurate and easily accessible journalistic record is important to many people and organizations: Historians, politicians, genealogists, scholars of all sorts, academics, sociologists, economists, governments, and, of course, to your great granddaughter.

The preservation challenges are formidable. It's no longer as simple as subscribing to the Washington Post and binding issues into volumes. Or piling them in a safe, neglected corner. Now to preserve news we must consider copyright, the dizzying variety of formats for digital news, the composite nature of many news feeds, its algorithmic personalization by Facebook, Google, and others, changes due to changing governments, citizen journalists and bloggers, as well comments about the news by readers. Once news was scarce. Now we are drowning in it! What can you do?

Presenters: Frederick Zarndt (Digital Divide Data), Sam Meister (Educopia Institute), Edward McCain (Donald W. Reynolds Institute, Missouri School of Journalism)

Katherine Skinner – Why is born digital news content a preservation problem?
Va Tech began collecting born digital in 1999. Biggest problem is that no one is collecting. There's a few pockets of information, but that's it. We've transitioned to digital-first news in the past few years. From 404 errors and link rot to renderable bits not altered/corrupted – to people wanting to manipulate what history looks like. The news does not just chronicle history; it influences history. If we don't have the news record, we lose the rough draft of history, and the understanding of how history unfolded.
The technical side is very confusing. We need to be keeping this info, and we're not. A couple other countries are doing this better, but no one has this mastered yet.

Frederick : How much digital-only news is not preserved and permanently lost?
Looked at home pages of huffington post and pro publica and watched the number of changes on the home pages on a single day. Quite a few. The Internet Archive does not even capture daily on most sites.
Ben Wells, LA Times: what does a typical publishing workflow look like?
A reporter, but also a database person. P2P custom built system built by Tribune publishing system—most news orgs use custom home grown databases. Knowing what to archive and how to integrate it into some other system is a huge problem. Many aspects were added over time and some no longer are used. There's interlinking between other articles as well. Revision history (many keep this in their own way) can be extensive. Constantly changing entries on a page.

Giant IT challenge.

What stakeholders need to be invested and involved to solve this problem? (Edward McCain)

Journalists have the content, and need to be able to access news from the past. However, their focus is making revenue, not preserving content.
Journalism Digital News Archive agenda and outreach initiative is the Dodging the Memory Hole program from MU Libraries and the Reynolds Journalism Institute – Journalists, librarians, technologists – we need more IT people. Thank goodness for Marc Wilson at Town News.com, a major content management system provider. Calls himself an "accidental archivist" – trying to convince his newspaper clients to save their digital content. TownNews has petabytes of data. Many publishers not interested in paying to store old content. TownNews.com has about 1700 newspapers using its CMS.
McCain is particularly interested in smaller and minority community newspapers. NYT keeps theirs! But most do not. Upcoming Dodging the Memory Hole conference at UCLA - sign up for info at: http://www.rjionline.org/events/dtmh2016

Frederick: How do policies and actions in digital news preservation compare across national boundaries?
Survey of libraries – 19 responded, from multiple countries. Asking about legal deposit of content of news. Northern European countries are doing pretty good – they've at least started. Poland, no policy at all. US, not very good. The first step is a legal framework that supports or requires deposit of the news.
What technologies are available now to harvest/preserve born-digital news? What tech needs to be invented?

Herbert: Similar to web-based scholarship, so speaking to that.
3 approaches;:
Backdoor approach: database dump/export (predominant approach – U of NTx, U of KY, Miinn Historical Society) -- exports best to introduce uniformity. Will have combo of manual and automated processes. Scalability issues. Most important – how representative is this of what the public interacted with? It represents the process of generating the news, but does not show social interaction and linking.
Front door approach: passive web archiving Advantage of a uniform technology stack – using proven capture approaches Example (Goelt? GDelt?) Problems with quality of what you archive. Missed content, changing content not captured, embedded and linked resources not captured.
Better: Active web archiving (still front door): engage with a web archive, existing or dedicated Proactive to allow the web archiving to do a better job. A growing field. Using Site maps, ResourceSync, W3C packaging on the Web format.
New, just coming out – allows all the content, even linked content in a page to be included in response to a single request.
None of these are mutually exclusive.

Presentation 2
Linking Born-Digital News and Social Media Collections via Automated Entity Detection and Authority Matching
Digital libraries increasingly have the means to assemble large collections of born-digital news materials and social media records. These collections can be opaque and difficult

to use on their own, yet become much more valuable when linked to each other, thereby increasing their collective exposure and discoverability and ultimately enabling the assembly of multi-perspective histories of significant events. We have been developing techniques to discover and use such inter-archive connections and now wish to share with the DLF community the results of our latest project, which links social media materials to a large database of computationally indexed news broadcasts.

The tools and resources we have employed in this project include named-entity extraction software specially tuned for each type of media, which then reconciles discovered entities with entries in the Virtual International Authority File (VIAF) service. To demonstrate the effectiveness of this technique and present a practical example for comment and evaluation, we describe our experiences linking particular sections of the UCLA NewsScape television news collection to material captured from Twitter that pertains to specific events of regional or global interest.

Steps we have taken to evaluate the effectiveness of these efforts include enumerating the types of entities we were able to discover and link between the two collections, as a first-order quantification of how our linking techniques can increase the visibility and discoverability of the resources in each collection. We also have assembled case studies of the types of multi-perspective narratives that users can build using online search aids that incorporate discovered inter-collection links into their results. We welcome input from other forum participants regarding next steps to take, new tools and resources to use in the future, and the potential for subsequent collaborations and re-use of our work in other archival projects.

Presenters: Martin Klein (University of California, Los Angeles), Peter Broadwell (University of California, Los Angeles)

Proof of concept work. Collections related to news events, collected by researchers, collected by archivists, diverse in format. Dynamic and ephemeral. International digitizing ephemera -- ex tweets about major events. Social media content disappears in no time. Using an open source tool called social feed manager. Social-feed-manager/readthedocs.org

NewsScape collecting TV news, 2005-present, from 13 countries, 9 languages, 38 networks. Searchable by captions, on-screen text, official transcripts.

Philosophy: Social Local Global (SoLoGlo)
Linking social media, TV news, and web news.

Linking via automated entity detection, in order to discover and highlight commonalities and relationships between disjoint collections on related news events. Want to establish automated workflow for linking, and integration with search and discovery interfaces.

Applying DBPedia Spotlight Named Entity Recognition software

Presentation 3

Online-Only Media: 21st Century Collection Crisis?

Today's music and movie industry is increasingly favoring streaming and download-only, direct-to-consumer distribution. No longer can librarians or archivists expect to collect sound recordings and videos on tangible media (e.g., CDs and DVDs) where first sale doctrine applies. At an ever-increasing rate, librarians are discovering that many titles are only available via such online distribution sites as iTunes or Amazon.com. These distributors require individual purchasers to agree to restrictive end-user license agreements (EULAs) that explicitly forbid institutional access and such core library functions as lending: "Upon payment for Music Content, we grant you a non-exclusive, non-transferable right to use the Music Content only for your personal, non-commercial, entertainment use..." (amazon.com).

With this presentation we describe failed attempts to negotiate a library-friendly EULA with tech/music industry representatives, and give an overview of an IMLS funded project tasked with investigating the issue. Called the "National Forum on Online-Only Music," the project has enabled us to hire legal consultants, work with the Library of Congress' National Recording Preservation Board, as well as envision a range of possible solutions: from licensing online-only works directly from artists to creating a closed collection of files that would be released when the content is no longer commercially available.

Presenter: John Vallier (University of Washington Libraries)

Collecting online music and video. Online-only -- only available as licensed download or stream; unavailable for purchase on tangible media; not limited to born-digital media.
Streaming is growing -- on demand streaming sales are rising, while CD sales are plummeting. By 2016, streaming should exceed media sales (video) and in 2017 will exceed box office sales as well.
The Download Dilemma -- DJ Hoek quoted. This compromises preservation of our culture.
If we leave preservation in the hands of industry, some bad things can happen -- like Universal Music's fire -- tremendous amount of content lost.

Challenge: Libraries can't purchase and collect online-only audio / video due to restrictive licenses. What do we do? We try to come up with solutions.
NEH and IMLS grants -- wrapping up the latter one now.

Moving toward direct licensing, but still problematic