Citizen Journalists and Community News: Archiving for Today and Tomorrow
Library of Congress, November 3-4, 2010

Wednesday, November 03, 2010

Opening/Welcome (Martha Anderson, Library of Congress)
Martha welcomed everyone and thanked them for being willing to join the "conversation".

Brief Attendee Introductions (Martha Anderson, Library of Congress)
Meeting attendees briefly introduced themselves.

Overview and Introduction (Martha Anderson, Library of Congress)
Anderson explained the format for the meeting and outlined the reasons for hosting the meeting.

Anderson said the important thing is to focus on the fact that we're dealing with at-risk collections and that we're learning by doing. The Library recognized a gap in collecting non-corporate produced news through previous meetings with news organizations and an investigation by the Copyright Office that examined how newspapers are deposited electronically. The Library recognizes that news has been collected for ages. Now that we consume news online, the question becomes how do we collect it?

Through the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) partnerships, it has collected some digital news. Some of the items being preserved through NDIIPP partnerships and the Library's own collections include:

- Foreign news sources
- Hossein Derakshan, the father of Iranian blogs
- Twitter archives
- Local news sources

We're seeing more community blogs and "hyperlocal news" that is too small for most papers (TBD is one such community of local blogs). When we ask who reports the news now, the answer includes experts and individuals who just happen to be in the right place at the right time. Clay Shirky pointed out that the shift in technology has made the news shift from one-to-many communication (traditional news broadcast or newspaper distribution) to many-to-many (Twitter feeds, blogs, social networks, etc.) communication. How does this shift change our perception of what is valuable in the information space?

A November 2, 2010, blog post by the Archivist of the United States asks whether Facebook status updates, Tweets, YouTube videos, and blog posts by Federal agencies are federal records. He writes that if you can answer yes to any of the following questions, then the item is probably worth preserving as federal records:

- Is the information unique and not available anywhere else?
- Does it contain evidence of the agency's policy, business, mission, etc?
- Is this tool being used in relation to the agency's work?
- Is use of the tool authorized by the agency?
- Is there a business need for the information?

The Library hopes to achieve the following through this meeting:

- Develop a shared understanding of the long-term value of new forms of journalism and reporting
- Propose approaches to develop criteria for assessing long-term value
- Suggest new organizational models to best support this stewardship
- Understand that newspapers represent users of the future as well as those of today

Summary of Pre-Meeting Questionnaire Responses (Abby Rumsey, NDIIPP)
The questionnaire was non-scientific and random, with a very high response rate. NDIIPP asked what you think about citizen journalism.

When we discuss digital content, we talk a lot about what is at-risk content. But another question we should ask is how do we actually contextualize all the content we collect?

Discussion (Abby Rumsey, NDIIPP)
Local media has always been hard to obtain, but now it's easy to get access to media from more countries. To what degree does that access affect how society functions? As an example, most of the press about President Obama's trip to India is coming from India. It seems that the President's trip is much more important to India than to Americans.

More often than not, information from countries like Africa comes from a guy with a transmitter. We need to look at a huge portion of the world that is broadcasting their news. Can we look at ways to collect those broadcasts? We need to look at sources other than electronic and social media sources.

What will replace Facebook and Twitter? From the Internet Archive's point of view, they want to make sure that social media sites are preserved before they become obsolete. Social media sites are constantly changing the rules. What are the rules? The Internet Archive tries to captures publicly accessible Facebook and Twitter. The sites don't try to stop us, but they do make it difficult to capture the information. Now the Internet Archive works with their engineers on a daily basis to be proactive. Flickr is pretty good. They don't change stuff a lot. Facebook has a new Terms of Service in order to archive them. The Internet Archive submitted it

To what degree do we have to think strategically vs. tactically on what to capture? We can't take a snapshot of the entire Web.

What other issues should we get on the table for discussion?

- Non-geographic community news
    - Special interest items that previously had maybe a monthly newsletter or something, but now they are run on blogs that update regularly.
    - Capturing those can be important in looking at how knowledge flows through particular communities.
- Keeping current with Facebook/Twitter
- Unknowns – Broadcasts (international)
- Thinking strategically vs. tactically

- The medium itself: The nature of the material is in a database or in a stream of interactions, rather than something that is printed on a sheet makes it interesting to think about how we'll save it.
- Metadata: What are the relational links from one story to another that helps you contextualize the collected item?
- Resources over time
  - Working in some sort of timeline, rather than just looking at a one-day snapshot.
  - "Chunks"
- Citizen journalism exists alongside big media news sources. We can't have the Twitter archive and not the Tribune archive. We have to look at all the sources together to get the complete picture.
- Organization and access
  - Some sort of clearinghouse?
- Preserve the sources
  - Sources have been unbundled.
  - Direct to audience aggregator.
- Relationship between the sources and the users
  - Collect the tweets and who follows whom.
- What is the Library doing? What is the technical capacity? How much are you taking in? What's the scope of digital collections today at the Library?
  - 5 terabytes a month. 190 terabytes total. We have specific collections such as elections, legal blogs, international elections, Supreme Court transitions, etc. The intent is not that it all comes to our basement. NDIIPP works with partners around the country and the world to build collections and provide access to them.
- What are the existing access tools?
  - Using tools that can scale entire collections to create histograms and show relationships among items.
- What partnerships are needed to preserve and provide access? Who stewards the material?
  - Media
  - Information companies
  - Universities have a strong desire to preserve content.
  - Probably not tech companies as they come and go.
- Where does the money come from to preserve these items?
  - People who care. It's the people who will pay to have their content preserved.
  - Research institutions
- Who is the network? Universities are key.

Review how we strategize this meeting: It's important to come up with a common understanding of what we're talking about. We want to focus on a case study of a group of people who have been collecting in a traditional way. Tomorrow we want to talk about the users and about what's happening in the web archiving landscape. What organizational challenge does collecting this content present? Advise the library on what it should be doing and with whom it should be working.

Case Study: Legal Blawgs (Donna Scheeder, Law Library of Congress)

Scheeder began her talk by saying that she and her team never discussed whether or not to collect legal blogs; they just did it. The Law Library of Congress archives legal blogs because the blogs are temporary, relevant, often cited elsewhere, an enhancement to the collection, and a service to the Library's mission. People use blogs because they can publish quickly, and it's important for them to get their ideas out there and into the public discourse. Legal blogs are often cited in legal opinions and *Law Review*.

The Law Library began capturing about 90 legal blogs in 2007, added 38 in 2008 and about 200 more in 2009. Increasing our international and comparative blogs is a current goal for the collection.

Curators from the Law Library, staff from Web Archiving in OSI, and staff from Web Services collaborated on this project. First, the Law Library came up with selection criteria (Variety, Authoritative, Nominee Status). They organized the blogs by categorizing them into bucket terms, and continually monitor them to see if they are still publishing and for standards (scholarship) maintenance.

The Law Library wanted to build a library-wide process for capturing Web content, and we thought this project could be used as a case study for the rest of the Library. The Library of Congress has been collecting Web content since 2000, but most of that collecting has been focused on events rather than formats. The Web Capture team tracks and crawls the sites, and we monitor the status.

The Law Library doesn't collect anything that isn't available for display to the public. They contact the bloggers for permission, and the bloggers re generally excited and pleasantly surprised to learn that the Library of Congress thinks the material is important and worthy of preservation. Each collection has a permissions plan that specifies to the crawler what content to capture, what links to follow, and how far to drill down into the site.

A MODS catalog record is created for each captured Web site. Metadata comes from several sources: a tracking tool, the archive, catalogers, and the Web Capture team. XML records are created for catalogers to review and enhance.

There are some Web integration challenges, such as how to integrate the blog collection into the standard loc.gov design template and the existing Law Library content. Working with the Way Back application data was an interface challenge in terms of how to make the results more user-friendly.

Some planned enhancements are to convert the Web archives site to a standard design, allow searching and browsing of catalog/bibliographic records, and better integration with other collections.

Some final thoughts: Collect and preserve if you have the resources. The biggest challenge ahead will be to keep pace with changes in scholarly publishing. Will librarians need to become aggregators in order to collect and preserve and still maintain the context of things? We collect the past and the present so there is a more informed future for everyone.

Q & A:

- Do you crawl the entire blog all the time? It is crawled once a month and then de-dup. There is also a blog about legal blogs that the Law Library monitors to keep up with new legal blogs.
- Have you found that everything you're archiving is consistent? There is one guy who changes his URL every month.
- How much is English versus non-English? 95% English right now. Long-term goal is to get more languages.
- Do you have data on what's relevant to users? The Law Library is working on that right now. The vision is that people will be able to put in a search term and will get results sorted by types of things, i.e. books, blogs, THOMAS, images, current legislation, etc.
- What about citations? They would cite from the original even if they come in from our site. We include information in the record about the original URL. The URL that shows in the browser shows our site with the original URL at the end.
- What is Minerva? It's the original name of the the Library'S web archive. Has nothing to do with DoD Minerva.

Citizen Journalism and Community News: What is the "Nature of the Beast"? (Rebecca Harrington, Huffington Post; Dave Winer, Editor, Scripting News; Jason Fry, Reinventing the Newsroom; Christopher Grotke, iBrattleboro.com; Dan Gillmor, The Knight Center for Digital Media Entrepreneurship)

Dan Gillmor: Anything you want preserved, you have to do it yourself. Anything that embarrasses you will automatically be preserved. Consumers=creators, creators=collaborators. Who is a journalist is the wrong question. We're talking about news. The question has to be what is journalism.

- We can agree that the *New York Times* is journalism. Blah Blah Blahg is not. In between is the guy who captured the iconic image of the London underground bombing. Where do we put WikiLeaks?
- This is about one AND the other, not one OR the other.
- What about advocates like Human Rights Watch who is doing better reporting on their niche than regular journalists?
- Technology reporting has moved to the Web.
- How do we handle the very local blog that's doing local news better because no one else is doing local news?
- The data part of this is so important. On the site EveryBlock.com you put in your zip code, and the info changes every moment because they are scraping millions of sites. Ushahidi put up great stuff on Haiti earthquake information. FixMyStreet.com in London combines government with citizens in a collaboration. BlogHer, Salon and The Economist are examples of other collaborative, informational blogs.
- Saving a blog post is great, but so is saving the comments, the inbound and outbound links. What about Yahoo pipes and Wikipedia edits? Can this information be archived? Is it proprietary?

Rebecca Harrington: Huffington Post is a combination of news aggregation and citizen journalism in its most diverse, pure form and consists of a small reporting team and networks of 6,000 bloggers and 23,000 citizen journalists.

Citizen Journalists

- The network of citizen journalists was created in 2007 through a program called Off the Bus. It was a response to traditional journalism that is press corps on the bus during political campaigns. Mayhill Fowler did Obama's famous bittergate quote was the biggest scoop.
- The program evolved into Eyes and Ears. We've combined forces to make a social news team. There's a Facebook person, a Twitter person, a comment moderation person, and someone who handles all these citizen journalists.
- Our latest project was the mid-term elections, in which we followed 80 elections using on-the-ground news reporting by citizen journalists. We recruited them and offered editorial support. We had everyone from veteran reporters to college kids.
- All of our social efforts are best epitomized by liveblog, which we did yesterday during the elections. It was an aggregation of lots of interesting anecdotal evidence by scraping twitter hashtags and individual reports.

Dave Winer:
1994

- Built a website and began using it to publish his own ideas
- Wrote a piece "Bill Gates vs. the Internet" in which I said that he was wiped out. Within 10 minutes he got an answer from Bill Gates and published his response.
- Was a blogger, even though that wasn't a term in use then.
- Got an offer from *Wired* magazine.
- Developed blogging tools.
- Dave's Blog Archive

1997

- Wrote a script that would publish an XML version of his blog every time he published it.
- Dave's XML script combined with a script another group was using for tech publications and became RSS.
- Then Netscape exploded, and it became a battle ground.

2000

- Worked with Adam Curry of MTV to develop podcasting. Because the Internet was a lot slower then, Adam came up with the idea of having feeds that download automatically and then notified the user they're available. Dave integrated with his RSS feed. It took 3 years for the idea to take off.
- When NPR started podcasting, it went over the top.

2002

- *The New York Times* began using Dave's XML script for blogs

Even though it looks like these technologies came from the wilds, it took the *New York Times* and NPR to make blogs and podcasts go.

Jason Fry:
Discussed how he became interested in citizen journalism and citizenry.

- Lives in Broolyn Heights and visits the neighborhood blog every day.
- The blog server practically melted when the neighborhood put in a bike lane that was painted an odd shade of green.
- The character of his neighborhood has remained consistent over time: they tend to look askance at new things.
- Brownstoner blog is a blog for house geeks and for people who dream of having a house they can renovate. When people ask whether a neighborhood is safe, what they want to know is the racial composition.
- Park Slope blog put up a post about a lost boys' cap which prompted someone to say they were offended by the assumption that it was a boy's hat. The discussion then went awry into whether or not it could even be called a hat.
- Interested in how these blogs reflect the historical context of our time.

Repackaging and Remixing of Media

- Which parts are important? Fry thinks it's all important.
- People are talking amongst themselves and no longer through mass media. A lot of what information about the past is from the Census, land grants, etc. They didn't have those voices, but now we do. So it's important to preserve it.
- Storage is not going to be a problem. The cost of 1TB of storage has decreased dramatically over time.
- The problem is access. Even if we have the storage of items in public forums like Flickr, the access might be a problem because the accounts might just go dark. In a hundred years, if you want to look at my Facebook photos, go ahead. But how do we make that happen? How can we have a "digital commons" for all of us, whatever we might do with it 100 years from now?

Chris Grotke: Brattleboro is a town of about 12,000 people in southern Vermont. iBrattleboro started in 2003. Grotke moved there and then decided the town needed a website because there was only a chain paper that didn't cover much. He realized there were a lot of smart people in town, so he decided to merge their ideas and utilize technology to allow the townspeople to write their own news on the website. http://www.ibrattleboro.com/

Since February 2003, iBrattleboro has 28 million views, 2,200 registered citizens, 16,000 stories, 82,000 comments, and 300 original hand drawn mazes. The site probably isn't interesting to people outside Brattleboro because it's hyper local, but it's everything to the people of Brattleboro.

A love of history led iBrattleboro to add a Wiki where users can add history, old photos, and timelines. Sometimes the stories aren't stories, but questions. For example, one user asked if anyone know anything about the body found on the street. iBrattleboro covers town meetings. They also put the news in context for one another. The writers often have seniority over the journalists from the local paper. The local paper had to respond because iBrattleboro was doing more stories with better coverage. The original papers were citizen journalists. It's a reputation thing. If someone comments about the city manager, that person might run into him or be questioned by others.

Q & A:

- To Chris Grotke: Who archives your site? They do and keep a backup of it. The real fear is what would happen to it if something happens to the people who run the site? They'd love to give it to the local historical society and the local library. They'd love a tool to help give the information to them. The data is getting too big to manage at this point.

- To Jason: If the local editor assumed nothing was preservation worthy about the blog site, what can we do to make people understand the value? A lot of these folks are such one man bands, that they're caught up in the minutiae of doing this that it's hard to draw back and look at the bigger question. Word will spread quickly that the Library wants to preserve social media, if there was a proactive plan to get the word out there.

- Is there a plan? We do seek permission from all blogs we collect. They get it and post about it on their blogs and say hi to their future grandchildren.

- What stories on this blog are important? The real importance lies in the collective stories.

- To All: We have a good start to the conversation. As individuals or as representatives of an organization, if approached by an entity such as the Library saying that they think it's an important blog worth preserving, how many of you would say yes without question? For some blogs, having a 50-year embargo before releasing the information would allow more people to agree. As the LIBRARY, we have a long view, but for people who are using it they want it now. We tell them to go to the site itself. This is an archive for future reference, not a real time environment.

- Would anyone be interested in exploring the digital commons idea?

- For sites that have contributors, do you have agreements with them? Chris Grotke: We have a statement at the bottom of the page that says you own what you post. But he sees the site as an archive anyway and feels that he could give permission for the Library to archive the site. Getting permission from commenters for the purpose of an archive is a huge undertaking. If you keep info on commenters, you open yourself up for subpoenas and you have to disclose that upfront.

- How do we make this a safe enough activity that local libraries and/or state libraries can preserve these blogs and local information, rather than putting it all in the Library of Congress's basement? A library sanctioned plug-in for WordPress that would allow the posts to automatically convert to a format that is useable by all sorts of archivists and send it to the local library.

Close-out

Thursday, November 04, 2010

Recap of Day One (Martha Anderson, Library of Congress)
Anderson mentioned three takeaways from yesterday:

- Digital commons concept
- Match-ups with local libraries and historical societies
- WordPress or other blog plug-in to allow for instant archiving

Anderson asked that everyone write ideas on sticky notes and post them to the flip charts at the front of the room throughout the day. At the end of the day, the following sticky notes were posted:

- Digital Commons
- Collection Policy Statement on Journalism (from the Library)
- Best Practices for Site owners/ “Archive” friendly guidelines for community news sites
- Network as editor/selector
- Legal deposit mandate needed
- Matchmaking Web service:
    - flag your content as archivable
    - let Libraries sign up to archive
- Create an archive tag like robots.txt to schedule crawls, grant permission…? /Use robots.txt as mechanism to allow publishers to give permission to archive
- Register your personal blog with the Library (genealogy project)
- Understanding “transfer points” between medial outlets and media types
- Is there a catalog system/tag taxonomy etc. for archiving? How to make one?
- Is anyone archiving congressional tweets and Facebook info?
- Metadata on authority can help solve the filter problem.
- A lot of this is very labor intensive, how do we automate? Should we?
- Specialist communities (non-geographic)
- State universities land grant - partner with local news
- Line between citizen journalism and pro-journalism is blurry, so consider bring them together into one initiative.
- Create a model to capture/access digital newspaper content that can be implemented at the local/regional level – a network?
- Geotags – GeoRSS
- Plug-In!!!
- Better use of RSS feeds in selection/archiving
- Match-ups with Local Libraries and historical societies
- Blogger Con
- I-Schools partner with local sites
- Partner with IMLS regional digitization programs on archiving
- 50 state network = USNP/NDNP participants

Research Use: Now and Future (Eric Ulken, University of British Columbia Graduate School of Journalism; Jeff Ubois, The Bassetti Foundation; Erin Schumaker, Graduate student, Ohio State University; Kristi Conkle and Ann Toohey, Humanities and Social Sciences Division, Library of Congress; Kalev Leetaru, University of Illinois' Cline Center for Democracy)

Eric Ulken:

- Teaches journalism and was a consultant on aggregating social media and citizen journalism in Los Angeles.
- Participated in the effort to map news in LA neighborhoods: mashing data with maps allows users to find out what happened in a particular area.
- Time Space by the *Washington Post* uses *Post* content (photos and stories) and maps it.
- SIMILE Exhibit Project is another example of a data/map interface.
- Current landscape: geo-tags on social media (tweets, photos, etc.), less so in blogs (WordPress now supports geo-tagging).
- Aggregators work with what they've got, such as outside.in and Every Block.
- GeoRSS attaches geographic info to content. It isn't perfect, but it's getting better as technology improves. Example: Reviews for the restaurant Mexico City were placed on the map in the city Mexico City.
- Is there a role for "Big Media"? TBD, Huffington Post, and NPR's Argo Network are working on this now.
- The network as editor? How do you figure out what's worth archiving? Look at what big media curates or go with the network. Look at Twitter and other social networks to see what influential people are linking to. TwitterTim.es looks at the people you follow on Twitter and looks at the links they post. You get a personalized news feed.
- Think about the metadata you would grab. Timestamps, geotags, but what about beyond that? How would you filter results? Maybe by inbound links, Google page rank, document readability score, and more.

Jeff Ubois: Participated in joint project with Yahoo Research on local news production and consumption, in which they looked at what people did and how they engaged with the news.

- Looked at people in San Francisco who were involved in production of local news.
- They found that consumption habits are in transition. People weave together multiple sources of news and they use multiple types of media (radio, digital, etc.).
- People feel conflicted about the loss of paper. Newspapers mix different kinds of information, which is split apart online.
- Local news has special characteristics. It touches personal identity and makes a strong emotional connection.
- Local is not just about geographic location. There are immigrant communities that want to remain connected to their culture, and they can do that by reading news of their homeland online.
- People were frustrated about quality. How do you measure quality and decide what to keep? Credibility, findability, and aesthetics are important.
- Advertising is a big part of the paper experience. Online news ads are from 3rd party sites.
- Quality has lots of dimensions we could score for preservation.

- It's great that the Library of Congress is archiving Twitter; it legitimizes the preservation of social media. With respect to partnerships, we need to be careful.
- Network is helpful and has a payoff. Local news is not confident in their own archive. Great to see partnerships with journalism schools, the Newseum, and major news organizations like CBS.
- Partnerships with commercial organizations can work out poorly. Revisit past work on partnerships. In Europe, there's a framework for public/private partnerships. NARA is working on these types of partnerships, and there is a need to strengthen our hand in these partnerships.
- Need to rethink what it means to be a custodial institution. The cloud providers are saying there's no need to worry about storage, but custodial institutions need to deal with this and have control over their own storage. Libraries are becoming more like IT organizations, and in the tech world, it's winner take all.
- Glad the Web is being archived and that off-air recording is happening. Other archives are growing quickly. The Library of Congress' is like a pilot compared to this. Pain is in collection policy and selection. This could be reduced if there was more slack in the system. Hope the Library of Congress builds out its own digital infrastructure because that would be a way to remain relevant.

Erin Schumaker: Schumaker's research involves a study about the diffusion of news using Twitter.

- Social networks are a news source.
- Discussed how news spread after Kennedy was assassinated. How quickly do things spread now that we have more technology?
- Looking at what gets told and where. Who breaks the story?
- No cost to share news anymore. This creates a massive surplus of information and massive noise. What rises above the noise?
- Focusing on the communicative act – sharing of news.
- Call for research using technology would get high consideration for publication. People think it's difficult to get a handle on the mass amount of untapped data. Computational Social Science looks at content and structure of online interactions.
- Diversity of news consumption might predict power or performance.
- Look at how people are linked through what they read. Archiving of citizen journalism would show this.
- The online networks are reality. It's happening and worth preserving. It's important to look at how people construct and live in their social realities.
- Access and manageability of archived data is a concern.
- Examples of academic uses: credibility judgments, heuristics, rumors, intergroup social identity, social network analysis, selective exposure (do people avoid info), interpersonal communication, and how news is being edited (who is editing, how they change over time).

Kristi Conkle:

- This meeting provides validation that researchers do want this information.

- The Library knows that diaries, letters, manuscripts are important to researchers now. The local aspect gives richness. Helps add to the New York Times perspective since we know there's a wider range of opinions.
- The comments on blogs will be helpful in the future. Discussion in the Chronicle of Higher Ed comments about social work student who refused to counsel gay people. This will show more to the story than just what the Chronicle printed.

Ann Toohey:

- Genealogists come to libraries for 3 main things. The news and newspapers are important. Local newspapers are helpful to users to find obituaries.
- Sites like iBrattleboro are now writing a daily local history.
- More usual are compiled genealogies. They give a historical background. Example of a Miller in New York who moved to Illinois because of drought. Compiled genealogies give biographies too. Local histories used to be published as books and focused on local events.
- Newer formats are coming about, such as the Ewing Family Association online newsletter. The Ewing Family Association has a reading room and information exchange. It's a collaborative news site that's also a genealogy site. The Library would want to collect it as news.

Kalev Leetaru:

- Works on global views of hyperlocal events. We present citizen media as a new problem, but broadcast is where a lot of the news comes from. It's hard to archive broadcast news, like Africa's because it's low power. You need boots on the ground to get that information.
- Even with unlimited resources, you cannot get mainstream media very well. The world is too large so we hire people to summarize what's fit to print. News is how we as humans try to process the world. Citizen journalism is not new.
- The issue with the web is that it's now high visibility and gaining authority. What do we do with the process of news?
- News is a filter. We know news in China is being falsified. With millions of blogs, how do we know that? There's a lot of data. We can never get all of it. We can't trust in updates.
- Do we keep content or context? Who decides what to keep?
- Strategic vs. tactical archiving. We need a little of both. No media is bias-free.

Q & A

- Mainstream media is influenced by public relations officials. Compare sources to see where news comes from. Look at AP Newswire versus Twitter, etc. Interesting to see how they all flow together. Is it similar or different?

- How far will the the Library go in serving as a point of reference for the originals? For example, changing press releases on the White House website. Leetaru said they are working on something to fix this.

- Sites can self declare how often they update. Look at RSS feeds.

- In the past, obituaries were important. In the future, maybe it will be blogs. The government can ask people to tell them where they keep personal blogs. Maybe you could register with the Library to say, here's my SSN and my blog. Glad the Library is archiving Twitter, but wants them to avoid archiving based on platform/brand. It will affect history by only archiving this one source. The Library does solicit genealogies and allows people to block personally identifiable information. There would be a suspicion of the government is they asked for blogs.

- Abby Rumsey said that people thought they had a lot of technology when Kennedy was assassinated. The cost of communication seemed minimal because there was public fear. The cost of publishing and communicating now is low and so we spend more time communicating which is of high cost. We suffer from filter failure. It takes more time to figure out what to pay attention to. Sometimes filters are not trustworthy. For archiving purposes, there's strategic versus tactical collecting. The legal blawgs collection is tactical and is based on metrics. Going forward, we should think about both of these types of collecting. Maybe tactical will be institution driven, public libraries, etc. and strategic will be large organizations driven.

- Martha Anderson said she is talking to people in the UK who want to develop personal or organizational Twitter archives. Efficiency versus redundancy. There's room to explore both, even if there's overlap.

- Is there a cataloging system for web archives? There's bottom up and then top down. Someone else disagreed. Maybe there should be a half-blogger, half-librarians unconference. 50-100 people and have informal discussions about these issues.

The Web Archiving Landscape: What is Currently Being Done? (Abigail Grotke, Library of Congress)

It's not just the Internet Archive anymore. Some of the current partners are the International Internet Preservation Consortium, university libraries, public libraries, museums, state archives and libraries, other U.S. federal agencies, non-profits, and schools. Most of us use the same open source systems like the Wayback Machine.

Some individuals are using Archive-It to create personal archives. Various collection strategies: bulk, domain, selective, event, and thematic.

Smaller countries can collect all items from their domain, but it's difficult to determine the .us or even the .gov domains.

We do selective crawls because we don't have all the laws in place or the money required to crawl everything.

Various legal strategies: legal mandate (deposit laws, collection mandates), permissions based, no permission/crawl without asking, and follow/don't follow robots.txt. We have the Copyright

Office in our building, so we have to be a little more careful. Some places have no restrictions on collecting, but we do.

The IIPC is also doing a collaborative archive of the 2012 Olympics.

Library of Congress collections: probably want to add a genealogy section soon.

We can't follow links due to legal restrictions. We can only show that there was a link.

Blogs appear in all our collections. When we did the Sotomayor crawl of Twitter, we had to add a tweet to the hashtag saying that we were crawling in order to satisfy the legal department.

On Minerva, we have a built-in time delay. We just put up the archive of the 2006 election.

Challenges in Selection and Preservation of Citizen Journalism and Community News (Emily Howie, Library of Congress; Mark Sweeney, Serial and Government Publications Division, Library of Congress; Cathy Hartman, University of North Texas; Devon Akmon, Arab American National Museum)

Emily Howie: Howie just finished a four-month detail as the Acting Acquisitions Director of the Library.

Five minute summary of how the library performs collection development.

- The Library has over 20 reading rooms, and collects in all formats and in 460 different languages.
- The Library acquires materials in four ways: the first is through mandatory deposit (American publishers have to submit two copies when applying for copyright), purchase, donations, and sharing materials with other institutions.
- The library does not have a copy of every book nor of every book in America. They review every item coming in according to collection development policies and do not collect clinical medical materials or agricultural materials. With regard to self published works, they only collect family histories and genealogical materials.
- The Library has over 250 recommendation specialists who review all of the incoming materials and determine whether or not the library will collect.
- An ongoing project is e-deposit, where we're learning how to acquire born digital materials for the library. The Library had problems with the first two titles that were submitted. The first one was a 172 page PDF from University of Florida. They sent over 500 files, which were determined to be production files. The Library does not need these types of files and is determining what to do with the extras. The second journal had different names on the XML and the PDF files. They finally determined that the journal republished materials.
- The Library of Congress is the largest library in the world. It contains 35-36 million print items and about 142 million items total. It's a complex task to keep track of all these materials.

Q & A:

- Will the recommendation officer for audio now be the digital audio person? Yes.

- For solicited donations, they are collecting electronic materials? Yes. We are getting digital materials in prints and photographs and manuscripts.

- Is e-deposit the same as NARA's ERA system? No, NARA is collecting government materials and the Library is collecting materials through copyright. They are completely separate, legally authorized systems. Web archiving is not covered in copyright yet. A few years ago web archiving met with the Copyright office, and the files that web sites sent were a mess. Sometimes they sent executable files that we couldn't use.

- How are you dealing with non-copyrighted items? The Library's entire collection of genealogy is offered voluntarily. They are not copyrighted. For the web archives, the lawyers only recently allowed automatic archiving of Creative Commons licensed websites.

- Are you getting the metadata from the Creative Commons licensed sites? The Library does take advantage of data that comes with websites. They try to leverage as much as possible and run scripts to extract it and then it goes to cataloging. Most of the info in the catalog is automatic and then catalogers fill in the blanks.

Mark Sweeney: Chief of Serial and Publications Division (newspaper collection); it's one of the most heavily used collections at the library.

Sweeney gave a brief history of how the newspaper collection has been developed over time and the current collection policy statement the Library has for newspapers. The collection policy covers newspapers, not journalism. Maybe this change could be an outcome of this conference.

The library has a very small collection of all the newspapers ever published in the United States. It's all valuable, key and important that it be preserved. No one library can collect all newspapers; no one library can collect all citizen journalists.

Deciding what to collect was originally given to Congress. They generally chose their local papers. The Library lost all the newspapers in the fire of 1814 and worked hard to reconstitute the collection. In 1851, another fire wiped out the collection. By 1874, they were presented with the problem of volume.

The Library must collect at least two newspapers from each state, representing different political views. The volume problem continues to plague the Library of Congress. Beginning in the 20th century, they confront the problem of loss due to poor quality of ink and paper. Preservation microfilming comes along, and the Library plays a huge role in developing that technology and then goes back and collects all the things it couldn't before.

The Library collects newspapers from every state and territory capital and the top 100 newspapers based on circulation. The national collection is distributed throughout the entire United States. It's not all kept at the library.

The current project is to digitize national historic newspapers. It's called Chronicling America. They're working with the states and asking them to determine what's best to submit to the national collection. The Library knows the challenges in capturing all these born digital newspapers.

Every two years people camp out in the newspaper room to research electoral candidates. Sweeney isn't sure how they're going to do that in the future if we're not archiving these born digital websites.

If outside companies crawl the sites, do they then own the material? There are concerns in libraries regarding that declared ownership of things. It's not automatic that the company that digitizes automatically owns it. The Association of Research Libraries (ARL) has discussed this. ARL recommends an explicit and short exclusive use period. Getting a restart on copyright because you've imaged something is a very shaky law. Public/private partnerships are tricky.

Cathy Hartman: Hartman told about her journey to understanding the importance of community newspapers.

A few years ago she came up with an idea for a collaborative project for the Portal to Texas History. Hartman traveled all over Texas to meet with people in libraries and museums. The first thing they wanted to collect was community newspapers. The Portal launched mainly with photos, letters, journals, scrapbooks, and books, but wanted to eventually have newspapers.

The Portal received funding in 2007 to digitize newspapers. They started working with local communities to help them get funding to put their own newspapers online. There is a strong interest in community news. Small publishers wanted it and historians wanted to use these materials for data mining. Researchers and genealogy organizations are also getting involved. They can't collect everything, so they decided to scope the collection on the community news level.

The Portal shared responsibility for collecting. They collect both historical and current community news. They contacted publishers and offer to take their born digital content. How do we provide unified access across different formats? This is the real problem and is not easy to solve. Some digital newspapers require a login to view material. So do they capture both the print and online version? They have to work with the publishers to crawl that site.

Q & A:

- Do they have a login because they charge people? They still have paid subscriptions, so that login is for subscribers. The technology on some of the sites is not archive friendly.

- Do you have any tips? Basically we need a seed list so that we know where you are and how to access the material. The list of URLs for entry points into your site. For selective crawling we have to tell the crawler where to go. A better way is to have a master RSS feed that can capture all the new content. Robots also cause problems.

Devon Akmon: The Arab American National Museum is in Dearborn, MI, which has the highest concentration of Arab Americans inside the United States.

The Arab American National Museum felt its community was being excluded. They were built by the community and went out into the community to gain content and support. They collect and preserve Arab American materials. They have a public library that people can use for research as well as traditional museum type materials. They tell stories through the personal vignette and believe every object has a personal story attached to it.

The Museum is looking at simple things like building a wiki that's populated by the community. They're also looking at a memory builder collection where people can share their personal stories. (See The Electronic Intifada for an example) They're trying to collect newsletters of Arab American institutions, Flickr site photos, and Arab American newspapers. There is a long history of Arab American journalism, but much of it has been lost. There are changing demographics in the U.S., and Akmon feel that their stories are being excluded and/or being incorrectly reported by major institutions and main stream media.

The Arab American National Museum has regional institutions and maintains a dialog with them. Partnering with institutions such as the Library of Congress would help them deal with a lack of expertise and resources. They're a medium size museum and don't really know much about web archiving; they just hired an archivist.

Q & A:

- I understand the sensitivities around current conversations. I wonder if your plan is to archive oral histories. That is our plan, but we need the resources. Right now our major drive in collecting oral histories is centered around particular exhibits. We've been working on a ten-year collection of oral histories about 9/11. We have a pretty aggressive oral history project, but it's centered around particular themes.

- How can libraries reach bloggers? Start by going through the people who provide the platforms for blogs. WordPress, Blogger, etc. Bloggers are pretty good at spreading the word if they think something interesting is being asked/done. You can also find a good blogger in your area and then look at the blogs they're following. Also look at PlaceBlogger.

External Environment: Factors with Preservation Implications (Clifford Lynch, Coalition for Networked Information)

It was interesting to note that the Library has a collection policy governing newspapers, not journalism.

Understanding that news flows in various ways, we hear one discussion that's about the history of local newspapers and that it's very tidy. If we step back and talk about what's going on in the circulation of news, the local paper is dying. A lot of local TV news stations are dying too because it's just a luxury that people can't afford and so news is becoming more centralized.

Looking at the traditional news media, there's a growing set of orthodoxies. But on the net, there are a huge set of news sources. Not just about places, but about hobbies, activities, commerce. This is taking place through blogs. Lynch thinks we need to be mindful in how we think about journalism and how we capture the news. In the future, those independent voices are going to be very important subjects.

Documentation of the big financial crisis in 2008-09 was extensively covered by sources outside the traditional news media.

The second comment is on the nature of platforms. This distinction between strategic and tactical archiving is very compelling. Maybe it's more about topic-driven vs. a large sample of the public mind. Now we're in pretty good shape on the tactical stuff. Bloggers are a hospitable group to archive. We might get into trouble when we try to go into the more complex social media. Once you try to understand these complex environments as transaction and transmission environments, you need to understand and track the various social relationships in there. Really understanding what's going on there means you need to understand who's following who and not just the text in the tweets. We will need to collaborate with the environment operators. We need to recognize that capturing what's going on the in the latest generation of social mediums will involve collaboration with platform providers. There are plenty of institutions that have the technical capabilities to encourage those types of collaborations. One of the strategic roles for those leader institutions is in building those relationships.

The last point is that investing in common tools and having site and platform operators use those tools is good. We need best practices for site developers, bloggers, etc. to make tactical collections as intellectually driven as possible. The toolkit becomes routine and the collection is driven by intellect. He's intrigued by non-geographic topics. For some reason they don't call this citizen journalism. When we talk about those non-geographic communities, we wonder who will step up and collect them. Larger research libraries may pick up the mantle there, where public and state libraries take on the geographic topics.

Q & A

- To what degree do you see this happening, where something like the Chicago Tribune hires people to cover suburban sports? If you can go to one place to make your archival arrangements, it's easier. There's an interest in aggregating violent crime. I'm not sure the economics will work out for what ESPN is doing though.
- A lot of us think the Internet is the archive. Just an observation. I think there's a need for education, for example people who think Flickr is a preservation site instead of a distribution site.
- We can now have news that's very timely in fields that couldn't afford to publish more often than weekly in paper. Does p equal np? Case study. Someone said they think they solved it. It didn't make the news for at least a week, but the ripples in the blogosphere started immediately.

Roles and Responsibilities of Research Institutions and Libraries (Don Waters, The Andrew W. Mellon Foundation)

We haven't talked about journalism much in this meeting. The criterion should be high quality vs. low quality rather than amateur vs. professional. Are the professional tools of journalists now so easily available by amateurs that it's hard to tell the difference? The ability to compose and distribute access to sources, fact-checking review, direct response, objectivity… These are all issues that need attention.

If we focus on the factors that the online environment provides, it's easy to morph the whole issue away from journalism to something else. It becomes an attempt to redefine journalism as conversations among people. Has the change in media given us a new category, and is it a journalistic category? Community is a word that gets thrown around. Is the use of this term in the context of journalism meant to be loose or is there a need for more precision? Websites can be community makers. They give rise to policy and governance questions. They become a place of convening. It becomes an object of journalism, not a source of journalism.

What we're dealing with is an emergent area. The emergence itself is important. The dynamic of this emerging area is formation. We're in the early stages of all that. Dave's idea of a un-conference of stakeholders is brilliant and necessary for identity formation. Because it's in formation, you have to just do it. You can't make any rules. We can formulate hypotheses, but that's all they are. The discontinuities are just as important as continuities.

We crawl the websites of our grantees because we wouldn't have all a complete picture of the work they are doing otherwise. Think about how web archiving affects your business in the library. If the aspiration is to stay current and influence practice, you have to think in terms of entrepreneurial activity. You can't embed this in an existing organization. There needs to be a middle ground between top-down control and laissez-faire when it comes to how libraries cooperate with each other and decide who will collect what. With respect to producers, everybody is looking for the easy button. The blog space is fairly mature and there has been increasing software and platforms.

Second issue is access to data on the point of the large aggregators. Content has really become devalued. There is a lot of it and its value is in aggregation.

The final point is the critical issue of engaging at the practice level is that is does build a community. How you form that community is extremely important and helps you deal with the most vexing issue which is how to deal with intellectual property issues.

Q&A

- Talk a little about the entrepreneurial issue. If you're thinking in the long view, that view has to embrace a short view of that business. Include the early stages as an effective part of how to deal with the long view.

- Comment not question: the series of questions rose about the role of journalism and they raised one more for me. It's not clear what the relationship is between news and journalism. You can have news without journalists. One other aspect is to not forget the citizen part. We take the citizen aspect very seriously. Communities need to communicate news and by communicating the news they become communities. There are so many exciting pilots and steps that will be spectacular failures where we'll learn what not to do and pass that along to others.

Identify Concrete Next Steps, Recommendations to the Library of Congress (Martha Anderson, Library of Congress)

Anderson appreciated that Don Waters painted out that this is an emergent area. We want to work collaboratively and those things just don't happen, especially since there is an expectation that we're just going to construct something and tell people what to do. When NDIIPP was formed in 2000, there was no Facebook, Twitter or Google. E-journals were the big thing. The ability to just do it, learn from failure, and have that entrepreneurial spirit is important.

Suggestions for next steps:

- Provide a Mechanism for Participation
    - Make it easy and second nature
    - Clarify incentives – "why should I worry about this?" Explain the value to the content producers so they have a reason to flag to archive
- Provide Tools to Support Archiving
    - Use RSS
    - Flag as archivable, then archives select from that and keep what they want (the Library's Recommending officers could provide subject expertise, decide what the Library keeps)
    - Use Robots.txt
    - Plug-in linked to a platform
    - API key and infrastructure
    - Push or pull?
    - How to deselect?
    - Like/Dislike/Archive button!
- Best Practices
    - Reach out to local institutions
    - Document existing best practices
- Partnership Networks
    - Community as nominator/editor
    - Leverage existing communities
    - Distribute stewardship (ex. IMLS digitization grantees)
    - Ask for funding
- Conferences
    - Vermont!
    - Unconference (producers & Archivists)