Submitted on: 17/08/2014

We are All Catalogers Now: Leveraging Do It Yourself Metadata for Research Collections

Rebecca L. Lubas

Claremont Colleges Library, Claremont, California, USA
rebecca_lubas@cuc.claremont.edu

Abstract:

Today's cultural heritage professionals face an embarrassment of riches and possibilities. Those possibilities are accompanied by new questions about collections and metadata. Content is being produced at the grassroots level, 24 hours a day, 365 days a year due to the portability and accessibility of digital recording devices and proliferation of content management sites. Events are not only captured in real time, but they are archived in real time. Personal recording devices and archiving methods are becoming more sophisticated, capturing metadata automatically.

This paper will examine how organizations can harvest content to support their research missions. It will include examples, practical advice, and discussion of the long-term implications. This paper examines how socially produced content can be leveraged by libraries and research collections for historical benefit. Society produces the content, and we as librarians are positioned to preserve it and place it in cultural context for researchers to discover.

Keywords: Metadata, tagging, user-contributed content, digital collections

We are All Catalogers Now: Leveraging Do It Yourself Metadata for Research Collections

The Current Context

In the pre-Web, pre-social media, pre-smart phone world, libraries and archives gathered materials on local history and events as best they could, based on the focus of a given

collection. Newspaper clippings, pamphlets, posters, and photographs created a picture of events that passed from current news into history. There was also much less of it then there is today, as not everyone had access to printing or photography equipment. Tangible content may enter collections years after an event, when memories have faded and the creators of the content have passed into history themselves.

Content describing and documenting events has transformed and multiplied due to the widely used tools available today. Every citizen can be a reporter, and advanced technological skills are not required. The content creators are becoming more sophisticated – they may understand and use Creative Commons licenses, they may add descriptive metadata to the content, and they may back up their data, to name some examples. How can cultural heritage professionals tap into the mountains of content? Should they even try? Do libraries and archives have a responsibility to do this? (Lampert 75) The challenge is to find methods for coping with the volume of material available. There is risk in not considering this material. Decades from now donors will not have boxes of photographs to give to an archive – the images and other content may or may not have been have been uploaded to the cloud, may or may not have been curated, and may or may not have been backed up. Although there are easily accessible tools now for archiving one's own content, that doesn't mean that individuals will diligently curate their material over time any better than they cared for old photo albums, especially if they need to migrate to new devices and platforms.

Libraries are uniquely positioned to address this dilemma. Librarians have the organization skills and tools to collect, curate, and make available the material in a way enriches their value beyond if the content remained "in the wild" on the web. We come from a culture of standardization, a skill that can be utilized to address this dual dilemma/opportunity. The fact that users can create content and interact with it more easily should not only mean we have more material to collect, but also that we have more opportunities to engage users of the collection with the content as technology breaks down barriers to use and makes the users feel more personally connected. The work no longer ends with collection and preservation, collections are now dynamic (Lampert 81). Collections have a life of their own, rather that remaining frozen in time.

This potential creates a new tension for librarians and archivists. Cultural heritage institutions have two missions: preserving resources for the future and making them discoverable for current users (Van Hooland 717). Using user-created metadata is becoming more accepted in the archival community (Gorzalski 20). The possibilities for the former increase with the volume of content as the potential for the latter increases with technology enabling user-added metadata.

Preparing for a Rich Harvest

Perhaps the most difficult question is where to start. What does your library already collect, and what is the special collections focus? Then, with the whole Internet to choose from, how do you narrow it down?

The issue becomes more manageable if you think in terms of what you would want to capture if the content had been in print. So, for example, if your library collects material on local politics (campaign posters, photos of political rallies, etc.) that is a good place to begin.

Searching established services for content can help with identifying content owners and material that is already grouped by subject. The IMLS Digital Collections and Content project tested this model using Flickr groups (Jett). Flickr and similar photo sharing and archiving services can be good starting places because of how the members can use photo sets, albums and groups to collate content based on geography and subject matter. Using an established service will also give you a chance to do what librarians do well: learn the way a system is organized to maximize your discovery.

Consider using social media methods as you would if you were looking for something personally – especially if you are trying to harvest material from current events. Knowing the trending hashtags in Twitter can be a key to finding the right content and the right communities for the topic you are attempting to curate. Users have a tendency towards describing collections (Stvilia 64). Therefore, it may be more efficient to seek batches of relevant content rather than attempt to uncover a single gem at a time.

When you create your harvesting process, make it deliberately structured, yet flexible (Lampert 74). As technology evolves rapidly, you want to be able to respond to new trends that will increase the ease of content creation and distribution.

Once you have established paths to discovering candidate content, then you can develop a proactive collection deposit strategy. Learning where the content creators are is the first step, and then you can create paths for the content creators to actively come to your collection.

Rights

Once you identify a body of materials that you would like to collect, whether or not you can proceed will depend on the rights. This can be a very time consuming step in the process that may not always yield usable results. The rights issue alone is a good reason to begin small. Rights are a good reason to target services that will capture and display the rights data for you up front, especially if it's in a community where Creative Commons licenses are encouraged. This can perhaps save you the contact time with content creators. You will also need to consider at what level your institution is comfortable with reusing – for example, do you only want to collect material with the most permissive license (the Attribution license, allowing anyone to reuse and change the content) or are you comfortable taking on Attribution-NoDerivs material, which requires the content be unchanged.

Create guidelines for what your library or institution is willing to tolerate, or, in other words, how much monitoring you are comfortable with needing to do in relation to how much you want the material in your collection. If you think it is likely there will be material worth your time to contact the author with a request to open up the rights, contact them for permission, but have a communication plan in place. Your institution can have a template letter with standard language for such contact, and then you can customize as needed. Keep in mind that the time invested chasing permissions should be in proportion to the priority of the subject matter for your collection. A popular event that is annual may not need the same investment as a once in a lifetime event. If one content owner doesn't grant permission, there may be plenty of alternative, similar content.

Deposit & Delivery

You will need to make decisions about how to store the material, which will be connected to the discovery of the material. Define formats you are willing to collect and maintain and chose a system that you have the technical knowledge to manage and the resources to support.

As we know how important the metadata is to the discovery of resources, chose a system that will capture the metadata you need.

Balance the quality of the system with the resources you will need to manage it. Consider that you may be able to scale larger numbers of community built collections by putting the control in the hands of the creators. For example, the Omeka platform allows content creators to easily create online exhibits of digital collections with a modest amount of support. You may be able to provide communities with advice and consultation so they can create do-it-yourself online collections that are a level of sophistication up from social media.

Descriptive Metadata

Harvested content will come with some of its own metadata. Naturally you will want to use the standards your institution has chosen; our professional values and standards for consistency supporting organization and discovery all apply. But what to do with the metadata that comes with the object?

The technical metadata may be machine supplied (such as directly from the camera used) however the description will likely be far more subjective and personalized.

A content creator may label an image with words that provide a value judgement, or subjective view of the event or object in the image. How much is that subject view a part of the content? Does an archive or library using the content have an intellectual responsibility to preserve this view, as well as a responsibility to provide an objective metadata and analysis? What are the ethical concerns if an "objective" view may run contrary to those of the content creator? The point of view provided by the content creator may add research value to the information object.

There is utility in opening up tagging to the content creator, or not stripping out user-created metadata when harvesting content – the user will know when, where, and who is in an image, for example. The situation gets trickier to navigate if you have images of not just Main Street in 1910 – but a current political event. The user comments may relate to the content creator's opinions. In libraries we have a tradition of being carefully neutral in our metadata, but there can be historical value in keeping the subjective commentary as well. The new culture of instant commentary can capture history. Therefore, setting policy on how subjective metadata is handled is critical before you begin to collect and has a role in the purpose of the collection. Also remember that the content creator may not want their content altered, and has possibly put this in their license.

Your potential audience has much to offer – even to the level of pinpointing content in a video – taggers may make "deep links" to specific moments in a video (Madden 699). Content can be described to a granular level that librarians do not have the resources to create.

Permanence

You have identified the content, resolved the issue of rights, contacted the creator if needed, and "collected" it in your system – now what? You need to consider the issue of your responsibility for the permanent preservation of this content, or if some of the ownership of responsibility on the shoulders of the content creators. Do the traditional questions of collection development apply in the same what to this content, or is a new approach required?

Defining what formats you collect before you begin will help guide you in answering this question. Keep in mind that if you collect a wide variety of formats you may need to be concerned about making software available to read the format – which is another resource to conserve. An alternative is to convert selected content to a standard format to extend its life, even if it somewhat compromises the original object.

Finding content on established services helps as many already have standards in place for format and file size. In other words – this is where you can take advantage of the some of the work that the services and content creators have already done.

A Checklist for Collecting Community Content

This checklist is a tool to get you thinking about the guidelines you will create to lay the foundation for a successful collection of user-generated content with user-generated metadata.

Harvesting

- Select subjects or events to monitor and seek
- Target open content management sites and social media
- Determine what formats you will collect

Deposit

- Select a system
- Know the capabilities and limits of your systems

Rights Questions

- Decide what rights restrictions you can manage
- Decide if you will pursue contacting the creator, or use content only if rights are readily explicit
- Create a content creator agreement

Metadata

- Define what metadata you will have on every resource

- Consider what resources and processes you need to edit and add metadata
- Define an editing policy for the user-created metadata

Engagement

- Decide if you will include a tool/membership on your site for adding tags and links
- Decide if you will edit user comments
- Create a Boilerplate letter for commenters to sign rights for using and editing their contributions
- Collect use data

Permanence

- Determine how long will you maintain commentary feature
- Determine if you will archive comments

Conclusion

The possibilities are very rich for using metadata as a doorway to more current, vibrant, and comprehensive collections. Libraries and archives can provide a new level of context for resources (Bullen 31). User engagement and technology opens as never before, and the advantage is for current user to increase interaction with archives to record history and for future scholars to have access to in-the-moment recording of events. We can now engage users and create a rich record for scholars. Metadata was once ours and ours alone. Now that that the masses have metadata creation capability, let us use that to better record history.

References

Bullen, Andrew. "The 'Long Tale': Using Web 2.0 Concepts to Enhance Digital Collections." *Computer in Libraries* [Online], 28.9 (2008): 31-35. Web. 27 Apr. 2014

Gorzalski, Matt. "Examining User-Created Description in the Archival Profession." *Journal of Archival Organization* [Online] 11.1-2 (2013): 1-22. Web. 27 April 2014

Jett, Jacob, Megan Senseney, & Carole L. Palmer. "A Model for Providing Web 2.0 Services to Cultural Heritage Institutions: The IMLS DCC Flickr Feasibility Study." *D-Lib Magazine* [Online], 19.5/6 (2013). Web. 27 Apr. 2014

Lampert, Cory, & Su Kim Chung. "Strategic Planning for Sustaining User-Generated Content in Digital Collections." *Journal of Library Innovation* [Online], 2.2 (2011): 74-93. Web. 27 Apr. 2014

Madden, Amy, Ian Ruthven, & David McMenemy. "A classification scheme for content analyses of YouTube video comments." *Journal of Documentation* [Online], 69.5 (2013):.693 – 714. Web. 27 Apr. 2014

Stvilia, Besiki & Corinne Jörgensen. “User-generated collection level metadata in an online photo-sharing system.” *Library & Information Science Research* [Online], 31.1 (2009): 54-65. Web. 27 Apr. 2014

Van Hooland, Seth, Eva Mendez Rodriguez & Isabelle Boydens. “Between Commodification and Engagement: On the Double-Edged Impact of user-Generated Metadata within the Cultural Heritage Sector.” *Library Trends* [Online], 59.4 (2011): 707-720. Web. 27 Apr. 2014