Issues with archiving community data

Lydia Spotts and Andrea Copeland

Introduction

Transportation in Indianapolis is evolving. The bicycle, two-wheeled agitator of a similar transportation revolution across the United States in the 1890s, is back. The city landscape, overwhelmingly distinguished by auto-centric design, is increasingly being reshaped to support cycling as the economic impact of these alterations changes perceptions and the cycling movement gains momentum. How to document the impact of an urban landscape in flux from the perspective of a loosely codified community centered on cycling is a considerable challenge worthy of consideration by archivists and information professionals in general.

Strategic efforts to improve liveability and multimodal connectivity in Indianapolis are traceable back to early 1990s joint efforts of the city's parks system and metropolitan development (Indy Greenways Master Plan 2013-2023, 2014). However, widespread support for and growth in cycling infrastructure has become increasingly prominent in the last decade, spurred by the success of pilot greenways initiatives and public demand. The now nationally renowned and heavily used Monon Trail "rails-to-trails" conversion (transformation of disused rail lines to recreational corridors) completed in 2003 has had tangible economic impact, as well as extreme popularity, that provided a strong case for further private and public initiatives (Indy Greenways, 2014, p. 33; Indpls. Cultural Trail, History). Contributing factors behind public demand include returning population density to the urban core and growing governmental and public support for diversifying mobility options due to environmental, health, and quality of life concerns (Indianapolis Bicycle Master Plan, 2012).

According to the June 2012 Indianapolis Bicycle Master Plan, between 2008 and 2012 the city's overall bicycling infrastructure has seen a mileage increase of 130 percent. Multiple agencies are now involved in development and operation of the expanding trails and bikeways network. Since its debut in 1994, the Indianapolis parks system initiative, Indy Greenways, has succeeded in developing nearly 70 of the 155 miles of greenways and conservation corridors first defined in the thrice updated comprehensive plan (Indy Greenways Master Plan, 2014, p. 36). The city began developing bike trails and on-street lanes in 2007, and a dedicated Bikeways division, under the Department of Public Works Office of Sustainability and Mayor's Bicycle Advisory Council, guide development and educational efforts. Currently, approximately 95 miles of on-street bicycle facilities have been constructed, with plans to reach over 200 by 2020 (City of Indpls. Office of Sustainability). The non-profit developed and managed eight-mile Cultural Trail, constructed between 2007 and 2013 with private and federal transportation grant funding, has connected cultural districts downtown, gained international acclaim, and has

---

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). *Participatory heritage*. London: Facet Publishing.

proven to be a viable recreational and transportation pathway (Indpls. Cultural Trail, Recognition & Awards).

Despite the extensive investment from agencies and organizations in fostering the culture of cycling in the city, there is not yet a significant formal mechanism for documenting and analyzing the effects of these rapid changes. While the work and decisions of government agencies and organizations will be documented in board minutes and public notices, and the physical reshaping of the city can be traced through maps and plans of engineers, how will the experience of the public and use of the changing space be captured and preserved? The official perspectives of politicians, media organizations, and commercial entities are recorded, but the public discourse taking place in comments sections following blogs and online newspaper articles, or the personal snapshots and reflections published via varying social media platforms, are of a troubling ephemeral nature.

The bicycle movement in Indianapolis presents an ideal issue around which to develop a community heritage collection, as the geographic and mobile nature of the phenomenon will expose the challenges of capturing both place-bound and digital history as it is happening. Information regarding the movement is current and thus is mostly in a digital form, published to the web and maintained on personal devices. Much like changes to the physical landscape of a city, current digital information can be difficult to grasp all at once as it is widely-distributed. This chapter will discuss challenges related to the capture and preservation of born-digital community documents and possible approaches to address them, which will be useful to information professionals and researchers alike, who will inevitably contend with multi-creator ever evolving born-digital collections as facets of the cultural record.

1. Social challenges

Capturing a representative portrait of the individual and community experience of this rapidly changing infrastructure and cycling culture is difficult because of the diverse and dispersed nature of the content, expressing praise and dissent within the broader community, and the informal nature of the complex and diverse cycling community. Navigating these issues to build a collection practically demands collaboration between heritage professionals/academics (librarians, archivists) acting in a facilitator role, and community members, using a participatory or community archiving model incorporating shared expertise (Cook, 2013).

1.1 Collection conceptualization and infrastructure

The multi-creator generated, dispersed textual, visual, and audio output of such an amorphous community lack a straightforward architecture and extent. Content cannot be randomly collected, nor collections constructed, by non-members without context and parameters for selection and appraisal. Community member expertise would be critical to determine a meaningful organizational structure. A collection scope, what to gather of the diverse and

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

plentiful output, in terms of content subject, extent/chronological range, creator, platform, and format, requires definition by representative community members and heritage professionals.

The lack of a unified, stable space or archive for the digital records of the loosely codified "bicycling community" could be remedied by building an institutionally-facilitated and hosted community heritage collection. After addressing technical and legal issues, dispersed content could be gathered and deposited in one place, then described and annotated by community members. Connecting communities to heritage or academic institutions with preservation infrastructure, as opposed to collections existing solely on corporate platforms with no promise of sustainability, is critical.

1.2 Community/creator identity and engagement

In order to begin conceptualizing a collection, however, some boundaries must be drawn and members must be identified so that professionals working for the hosting institution can solicit engagement. Expressions of identity and reactions to the changes in bicycling infrastructure are as diverse as they are dispersed. Conceivably, any person that utilizes public streets and bike paths is a member of this broad, loosely codified, informal community. Co-creating parameters, rather than imposing boundaries and identity on a community, enable the inclusion of diverse experiences–not just that of "model" cycling enthusiasts; However, this would be most difficult. The perspectives of an individual who may self-identify as a member of the cycling community in Indianapolis, but that is not a dues-paying member in any organization, does not participate in online discourse, and only occasionally signs a petition, can be equally valuable.

While fairly amorphous, the bicycling community does have some structured components and flagship events that attract a range of participants, and could be used as starting points for engagement. Bicycling community members may formally identify through alignment with cycling advocacy and educational organizations (IndyCOG, indycog.org; Central Indiana Bicycling Assoc., cibaride.org ; Freewheelin' Comunity Bikes, freewheelinbikes.org) and event participation (NITE Ride, niteride.org ; City of Indpls. Mayor's Bike Ride). The growing range of cycling organizations in Indianapolis with varying missions and constituencies–youth, commuters, competitive cyclists, history enthusiasts– and the diversity of events, from family-friendly to elite and niche, increase the possibility of connecting with individuals from casual cyclists to aficionados.

A community collection must also document pushback and dissent. While widespread, the support for transportation evolution in Indianapolis is not unanimous and there are many who would prefer uninterrupted continuation of the dominant car culture. As creative reconfigurations are imposed on a city sprawled by post-war suburbanization and carved-up by interstate construction, there is ample overall dissent as well as criticism regarding quality of infrastructure changes and implementation. As some residents increasingly integrate modes of transportation beyond the automobile into daily life, while others do not, the expression of

---

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

commendation and displeasure over how public space is used and how it is physically altered is sometimes a fleeting verbal clash on the road. However, individuals also express displeasure in the realm of online media via editorials and comments following articles. Many pro-cycling organizations exist, but there is not evidence of an organized opposition, making documentation of these perspectives more challenging and dependent on capturing reaction to and commentary on mostly pro-cycling coverage.

2. Technical challenges

The rise and diversification of social media has created great opportunity for more voices to be heard and has transformed human communication, while presenting many challenges for long-term preservation. Forum posting, microblogging, and sharing images permeates our everyday 21st century lives, and much more is created than what can possibly be methodically captured or curated. Given the ubiquity of multifunctional recording and communication devices at our fingertips, both purposeful and serendipitous documentation and interaction is ongoing. Collective memory creation is ceaseless.

Unfortunately, the creation of born-digital content via dynamic social media platforms far outpaces the development of digital preservation techniques. Standards and best practices for collecting and preserving content from widely-utilized, established platforms that have played integral roles in significant political and social movements can still be considered underdeveloped after a decade (Arnold & Sampson, 2014). Social media's legitimacy can no longer be dismissed as it gains acceptance as an official communication stream of institutions, and means of personal expression and network curation, usurping traditional communication methods. It has been acknowledged that the "record of histories and cultures will increasingly rely on the ability of researchers and archivists to document this fast-paced and dynamic form of data"--there is urgency to the work of heritage professionals using case studies and research to refine strategies and recommendations (Thomson, 2016).

2.1 Preservation in practice

Evolving practices and strategies for both small- and large-scale web and social media archiving efforts would need to be surveyed to determine what will best support the co-creation of a cycling community archive. While a portion of the proposed community collection could be built via direct web-form upload or other transfer of video, image, text, and audio files submitted by community members, or oral histories captured by heritage professionals, much of the content will be networked communication on web platforms, likely needing some form of selection, retrieval and curation–manual, automated, or a combination of both.

Archiving web content, "Web Archiving," in the most rudimentary sense, can mean manually capturing discrete content meeting predetermined criteria and identified by manual browsing or crawling, with screen capture or PDF-creation. This static capture, however, does not preserve any of the context or functionality of the original digital environment. Both

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

commercial and open source web-archiving software for websites can capture entire sites in their native file formats and preserve most functionality using the WARC archival format. However, deciding where to draw boundaries with complex and interconnected websites, both temporally and in terms of linked and embedded content, can be challenging (Pennock, 2013). Selecting and archiving even thematic portions of content on social media platforms, with innumerable users endlessly generating new content, is even more challenging. Since content is dynamic, it cannot be captured by web crawlers, but requires use of platform-specific APIs. Most commercial products available for social media archiving are geared toward institutions needing to retain their own records for legal purposes, and thus can be used only for one's own accounts. The "backup" version may not be identical to the native objects and suffer from other technical limitations (National Archives and Records Administration, 2013).

Taking a broader view, institutions and researchers have experimented with large-scale methods of social media capture that involve running processes to extract machine-readable data generated by user interaction on platforms. However, harvesting primary and secondary social media data via APIs, data which likely requires professional skills to interpret and may have reuse limitations set by the commercial company, preventing deposit in an archive, is not a good fit for building a public, easily navigable community heritage collection.

Like many institutional self-documentation efforts, a co-created community archive might be predominantly subject- or event-based and built around more traditional Web 1.0 content, like static sites, and methods, with selected Web 2.0 social media pieces. Instead of extracting data from a platform, this could mean capturing only the results of isolated searches on selected platforms, working with account owners to self-archive their streams and the responses, as is possible with Twitter (https://support.twitter.com/articles/20170160), or capturing comments on specific blogs or news sites.

Whatever capture methods are used, given the necessity to create collection parameters, the current lack of standards for social media preservation, and the limitations of tools, it is likely the resultant preserved digital objects will be different from how they existed online. An invaluable part of preservation efforts in building a community collection will be documenting the decisions, methods, and tools used.

2.2 Technical aspects of collection parameters

In the cycling community of Indianapolis, multiple creators are generating and altering dynamic web content in real time, across myriad communication streams, but unified by geography and common themes. Positive and negative on-the-road encounters are captured with a GoPro camera and shared, or simply relayed in personal blogs. Opposing viewpoints may be presented on radio programs or podcasts. News coverage of infrastructure changes is in the local paper, editorials in the business journal, and analysis on blogs, with accompanying comments. Making selections from this format-diverse and dispersed deluge of content is necessary to enable the

creation of a navigable, cohesive collection, able to be preserved and migrated to new storage locations over time.

While some co-created parameters for collection scope, may be predominantly social-cultural considerations–community members would identify important subjects, events, and opinion leaders based on their expert knowledge of the community–many aspects of making selections from the deluge of potential collection acquisitions have technical components, on which heritage and information professionals could advise.

Key platforms will likely be identified by community members based on where the critical conversations are happening, but not all platforms allow for easy archiving of user-generated content due to individual account privacy settings and some, such as SnapChat, are designed explicitly to allow content to expire. Depending on the platform, the desired extent or chronological range may not be available for capture. For example, past Tweets cannot be searched through available APIs (Thomson, 2016, p. 7). It may not be possible to capture content in its native format, and some formats may have degradation issues that would hinder future access, making them less than ideal acquisitions for an archive.

Whether content is uploaded through a web form, independently collected from the open web, or collected by archiving specific accounts in cooperation with creators, a technical collection strategy, capture protocol/policies, and minimum preservation standards for each format type will need to be determined, documented, and promoted to community members creating content.

3. Legal challenges

Much of the born-digital content created and shared by community members that would make critical contributions to a community heritage collection are not necessarily easily transferred. There are extensive legal challenges to collecting and preserving digital content hosted by social media platforms, accompanying online news articles, or residing on personal devices.

3.1 Terms of service

While corporate entities such as Facebook have created a low barrier to interacting, constructing identity, documenting, and creating in the digital world, much of this content that is our collective social history is often no longer solely "ours." The terms of service (TOS) for most social media platforms gives the corporation a permanent license to any content contributed to their sites, a license that is transferable to a third party, such as a data broker or company wishing to market to users, but also potentially to a memory organization. Some terms of use/service, such as Yahoo, lay claim to submitted ideas or intellectual property provided via feedback functions as well (Lipinski & Copeland, 2013, p. 189).

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

Republication and preservation outside of a commercial platform is not straightforward, a troubling situation when corporate entities could abruptly cease to exist or shut down platforms without notice. Further, corporations have no legal obligation to preserve our collective social heritage and have often demonstrated a lack of concern for users and user-created content (Hicks, 2016). Much user data has been lost because of the lack of corporate commitment to preservation planning. Last-minute "emergency recovery" efforts of archivists have been rare and serendipitous, with resultant incomplete data capture barely sufficient when compared to the native environment. The cultural record as created via these currently ubiquitous platforms will be completely absent or no more interpretable than the Voynich Manuscript when future researchers attempt to analyze it if commercial entities persist without preservation plans or deposit agreements (Thomson, 2016, p. 17).

One option would be for memory organizations working with community-created archives to develop their own terms of service to ensure they can legally retrieve and host content originally published via social media platforms. Institutions and individuals could digitally sign the terms, allowing the memory organization to capture content posted from their specified accounts to platforms. Since many terms of service for social media platforms give the corporation a permanent license to any content contributed to their sites, arrangements could be made with corporations to have that license transferred to the memory organization (Comer & Copeland, 2015; Comer & Copeland, 2015 Oct. 23). Unfortunately, "there is very little precedent for relationships between social media platforms and institutions who specialize in archiving and long-term preservation;" so far, only Twitter has shown concern for the preservation and research of user data, notably, by depositing it with the Library of Congress, as well as working with some research institutions (Thomson, 2016, p. 17).

If social media sites such as Facebook and Flickr would extend their permanent license of user-generated content to the cycling community archive, for example, then web-harvesting protocols could be established for those sites. Alternatively, a registry could be created for individuals wishing to contribute regularly to the archive. Registry participants would authorize the capture and preservation of their content under certain circumstances. For example, a contributor could indicate anything tagged with "#CycleIndy" be harvested for the archive.

3.2 Communicating reproduction rights

The right to capture and preserve content, covered by the terms of service, will not necessarily equate to the memory organization owning copyright to all contributed works or being able to provide researchers accessing the collection permission to reproduce. As part of agreeing to terms of service, a contributor might simply retain ownership of their work or be given the option to select a license.

Creators wishing to submit discrete objects to the archive via email transfer or a web form upload could also either select a Creative Commons license

---

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

(https://creativecommons.org/licenses/) for their content, or sign a transfer of copyright similar to a "Deed of Gift" traditionally used for donations to archives. No matter the copyright status, standardized rights statements, with concise versions for collection end-users, would need to be developed to overcome the confusion, fear, and hesitation that might hinder creative reuse, ensuring a living cultural heritage collection, or engagement with the collection from the community that built it. The joint initiative of digital collection aggregators Europeana (http://europeana.eu/) and the Digital Public Library of America (DPLA, http://dp.la/), Rightsstatements.org, would serve as an excellect model.

Conclusion

Biking in Indianapolis is a drastically different experience than just a few years ago, but without a central archive of information to revisit, how will city planners, sociologists, and cultural historians of the future examine this transformation? The perspectives of the community, captured with quick photos on personal devices, spontaneous textual expression broadcasted via their platform of choice, or purposeful videos posted to YouTube, can provide valuable documentation. There is a need to capture and sustain this born-digital documentation that will provide a comprehensive, accurate, and authentic portrait of a city in transition. The co-creation of an online repository for community-created content, while challenging, would begin to address this need and would provide long-term access to information that would not otherwise be aggregated and preserved in a meaningful, re-discoverable manner.

Arnold, T., & Sampson, W. (2014). Preserving the voices of revolution: examining the creation and preservation of a subject-centered collection of Tweets from the eighteen days in Egypt. *The American Archivist*, *77*(2), 510-533.

City of Indianapolis Office of Sustainability. *Bikeways*, http://www.indy.gov/eGov/City/DPW/SustainIndy/Bikeways/Pages/Bicycling%20in%20Indy.aspx

Comer, R. S. & Copeland, A. J. (2015). Methods for capture of social media content for preservation in memory organizations. Bulletin of IEEE Technical Committee on Digital Libraries, *11(2). http://www.ieee-tcdl.org/Bulletin/v11n2/papers/comer.pdf*

Comer, R. S. & Copeland, A. J. (2015, October 23). Methods for capture of social media content for preservation in memory organizations. [Video File]. Retrieved from: http://disted.informatics.iupui.edu/event/lis/Copeland.php

Cook, T. (2013). Evidence, memory, identity, and community: four shifting archival paradigms. *Archival Science, 13*(2-3), 95-120.

Hicks. J. (2016 May 8). The guerrilla archivists saving the Internet's dying websites from oblivion. *The Kernel*, *Archive Fever Issue.*

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.

http://kernelmag.dailydot.com/issue-sections/features-issue-sections/16616/archive-team-saving-the-web/

Indy Greenways Master Plan 2013-2023. (2014). *Full Circle Master Plan*, Chapter 2, https://indygreenwaysmasterplan.files.wordpress.com/2014/03/ch2-indygreenways-overview.pdf
Indianapolis Bicycle Master Plan. (2012). *Executive Summary,* www.indy.gov/eGov/City/DPW/SustainIndy/Bikeways/Documents/Indpls%20Bike%20Plan%20FINAL%20June%202012.pdf
Indianapolis Cultural Trail: A Legacy of Eugene and Marilyn Glick. *History*, http://indyculturaltrail.org/about/history/
Indianapolis Cultural Trail: A Legacy of Eugene and Marilyn Glick. *Recognition & Awards*, http://indyculturaltrail.org/about/recognition-awards/

Lipinski, T. A., & Copeland, A. J. (2013). Look before you license: the use of public sharing websites in building co-created community repositories. *Preservation, Digital Technology & Culture*, *42*(4), 174-198. doi 10.1515/ pdtc-2013-0028
National Archives and Records Administration. (2013 May). *Best Practices for the Capture of Social Media Records*, www.archives.gov/records-mgmt/resources/socialmediacapture.pdf

Pennock, M. (2013). Web-Archiving. *DPC Technology Watch Report*, 13(01). doi: http://dx.doi.org/10.7207/twr13-01
Thomson, S. (2016). Preserving social media. *DPC Technology Watch Report*, 16(01). doi: http://dx.doi.org/10.7207/twr16-01

This is the author's manuscript of the chapter published in final edited form as:

Spotts, L., & Copeland, A. Issues with archiving community data. In Roued-Cunliffe, H., & Copeland, A. (2017). Participatory heritage. London: Facet Publishing.