Combining Social Media Storytelling With Web Archives

Michael L. Nelson, Michele C. Weigle, Kristine Hanna
{mln, mweigle}@cs.odu.edu, kristine@archive.org

In this project, Old Dominion University is the lead applicant and is collaborating with Archive-It (a branch of the Internet Archive) to develop tools and techniques for integrating "storytelling" social media and web archiving. The project will run from May 2015 to the end of April 2018.

Much of our cultural discourse occurs primarily on the Web and its preservation is a fundamental precondition for research in history, sociology, political science, media, literature, and other related disciplines. Archiving web pages into themed collections is a method for ensuring these resources are available for posterity. Services such as Archive-It (`archive-it.org`) exist to allow institutions to develop, curate, and preserve collections of web resources. This is done by specifying a set of seeds, Uniform Resource Identifiers (URIs), that should be crawled periodically. At the same time, storytelling is becoming a popular technique in social media for selecting representative tweets, videos, web pages, etc. and arranging them in chronological order to support a particular narrative or "story". Tools such as Storify (`storify.com`) provide an easy interface for users to arrange web resources to create a story.

We will address two main research thrusts: 1) summarize existing collections, and 2) bootstrapping new collections. When an archivist creates a collection, it can include 1000s of seed URIs. Over time, each of these URIs can be crawled 100s or 1000s of times, resulting in a collection having thousands to millions of archived web pages. Understanding the contents and boundaries of a collection is then difficult for most people, resulting in the paradox of the larger the collection, the harder it is to use. We will develop techniques to automatically (with optional human review and "steering") sample pages from a collection that summarize and describe the collection. For example, given a collection of 1000s of pages, our tool will automatically select 20–30 representative pages that will then be linked in storytelling web applications, such as Storify. Although page selection is not dependent on tools such as Storify, we are committed to the approach of using existing tools instead of developing new ones. Similarly, archivists wanting to create collections about breaking events (e.g., Ebola, enterovirus D68) need to have domain knowledge about the event to create a quality collection. Since users around the world are already creating stories in places like Storify about these events (but without an archival component), we will develop tools that allow archivists to mine existing public stories to quickly generate seed URLs for a collection.

We will be creating two main software products. First, we will create a web-based interface for curators to analyze their collections and create stories that can be uploaded to a service such as Storify. Second, we will create a tool that can recommend new seed URIs for existing collections based on already-created stories on Storify. The audience for our tools will be collection curators, initially at Archive-It, but later this could be any curator whose collection is stored in the standard WARC (Web ARChive) format. The curators will use these tools to enhance their collections and create interesting and compelling stories from the collections. These stories will be available for viewing by the general public through Storify or similar services.

In addition to project's contributions in information retrieval, text mining, and web archiving, we will also demonstrate that web archives can increase discovery, access, and the quality of user experience by using tools (e.g., Storify) with which users are already familiar as a bridge between the current and past web. Our partners at Archive-It will help to evaluate the impact of the developed stories in terms of numbers of users and web referrals to the archive.

1 Statement of Need

Much of our cultural discourse occurs primarily on the Web and its preservation is a fundamental precondition for research in history, sociology, political science, media, literature, and other related disciplines [30]. For example `sonicmemorial.com` was constructed to be an archive of digital memorials and shared media from 9/11 [11], but that site itself has since been lost and is only partially archived[1]. Archiving web pages into themed *collections* is a method for ensuring these resources are available for posterity.

Perhaps the largest and most influential actor in web preservation is the Internet Archive (IA). Archive-It (`Archive-It.org`) is a collection development service deployed by the Internet Archive in 2006. Archive-It is currently used by over 300 institutions in 47 states and 16 countries, and features over 4B archived web pages in over 2500 separate collections. Archive-It partners receive an account at Archive-It and build themed collections of archived web pages hosted on Archive-It's machines. This is done by the user specifying a set of *seeds*, Uniform Resource Identifiers (URIs), that should be crawled periodically (the frequency is tunable by the user), and to what depth (e.g., follow the pages linked to from the seeds two-levels out). The Heritrix [28] crawler at Archive-It then recrawls these seeds at the specified frequency and depth to build a collection of archived web pages that the curator believes best exemplifies the topic or theme of the collection. Archive-It provides faceted browsing and search services on the resulting collection (Figure 1 shows the current interfaces for a typical collection[2]).

Of the over 2500 collections in Archive-It[3], many are devoted to archiving governmental pages (e.g., all web pages published by the state of California) and memory organizations like libraries and museums, but many of the collections are explicitly centered around topics in arts and humanities (331 collections), politics (183), spontaneous events (134), and blogs and social media (234). Choosing seed URIs for a collection, especially collections not centered around an organization or governmental entity, is currently more art than science: too few seeds and you fail to capture the zeitgeist of the topic you wish to archive ("low recall" in information retrieval terms), but too many seeds (or too much crawling depth) and you risk introducing web pages that bloat the collection with off-topic material (known as "low precision"). Furthermore, judging precision and especially recall requires a great deal of domain knowledge about the collection's topic, making it difficult for non-specialists to quickly create effective collections for rapidly evolving and occasionally multi-part topics, such as disasters (e.g., 2011 Deep Water Horizon) and political events (e.g., Arab Spring).

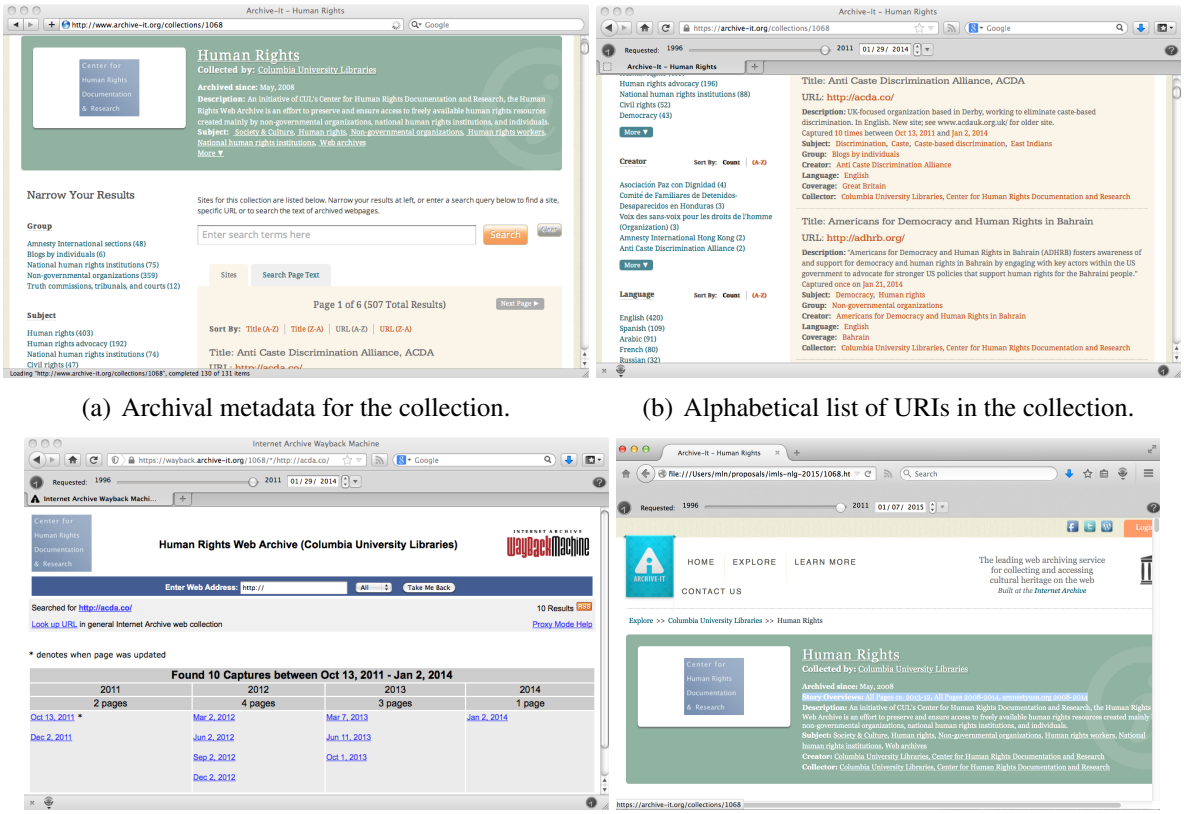
Even if we set aside the problem of selecting the right seeds to create a collection, there is still the problem of *collection understanding*. It is difficult for users arriving at the page shown in Figure 1 to understand what is in this collection and how it differs from the 17 other collections in Archive-It that are also about "human rights", albeit each with their own specialization. Aside from the brief metadata about the collection (Figure 1(a)), the interface mainly consists of a list of seed URIs in alphabetical order (Figure 1(b)), and for each of these URIs a list of the times when the page was archived (Figure 1(c)).

At the same time, "storytelling" is becoming a popular technique in social media for selecting

[1] It became spam in 2006: `wayback.archive.org/web/*/http://www.sonicmemorial.com/`

[2] Yes, the screen shots are small but links to all the web pages shown here can be found in the supplemental information document.

[3] For a complete listing, visit: `www.archive-it.org/explore?show=Collections`.



(a) Archival metadata for the collection.

(b) Alphabetical list of URIs in the collection.

(c) Archived copies of the first URI in the collection.

(d) Proposed "Story Overviews" (highlighted) added to collection metadata.

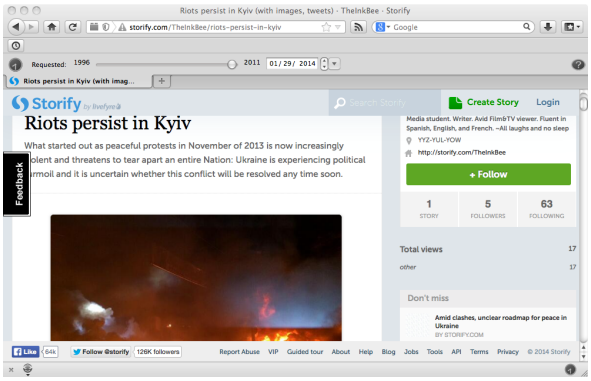
Figure 1: Browsing and searching services for Archive-It collections (Current: a–c, Proposed: d).

representative tweets, videos, web pages, etc. and arranging them in chronological order to support a particular narrative or "story"[4]. There is a lot of interest in this (e.g., both Twitter and Facebook support "Timeline" and "Year in Review" applications) and is best typified by the company Storify[5] which provides an easy interface for people to arrange web resources to create a "story". For example, Figures 2(a) and 2(b) show two actual stories on Storify from January 2014 about the riots in Kiev, the Ukrainian capital. This is before the crisis came to a head in late February, 2014 when Russia began the annexation of the Crimean Peninsula. Both stories were created by interested individuals (not necessarily professional journalists or archivists) and provide a sampling of social media and conventional news stories about the riots, along with commentary from the user to provide further context for the story. This is in stark contrast with the collections about the Ukraine in Archive-It as of January 2014 (Figure 2(c)), the most recent of which began in 2009. Archive-It did not create a collection about the Ukrainian Conflict until the annexation of Crimea, approximately one month later (Figure 2(d), even then most of the seed URIs were collected in the summer of 2014). There is a good chance that collection in Figure 2(d) will be missing many of the prelusive contents of the stories shown in figures 2(a) and 2(b), which we believe can be used

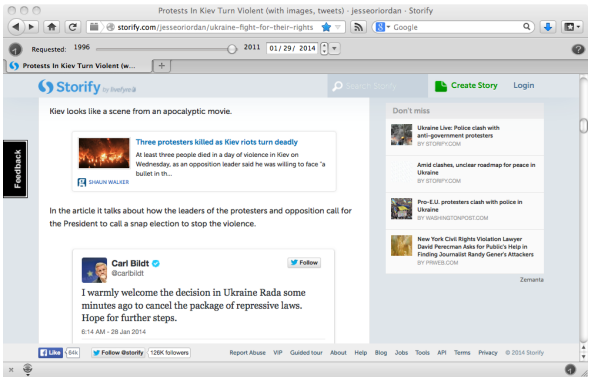
[4]We use "story" in its current, loose context of social media, which is sometimes missing elements from the more formal literary tradition of dramatic structure, morality, humor, improvisation, etc.

[5]`storify.com`; other similar sites include `pinterest.com`, `scoop.it`, and `paper.li`

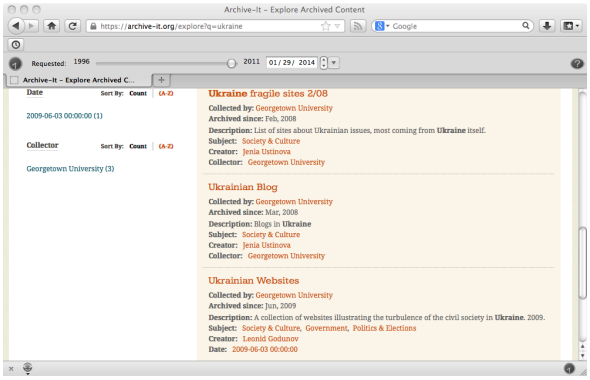
to either *augment* existing collections or to *bootstrap* the creation of a collection.



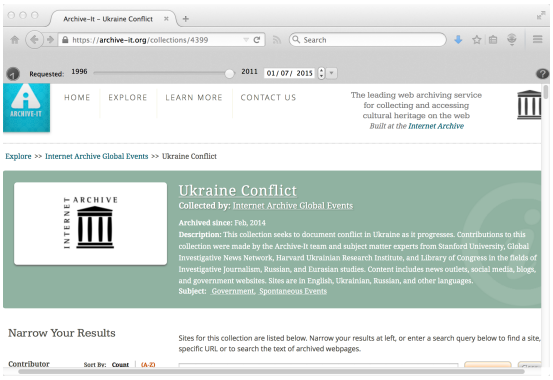
(a) Storify: "Riots Persist in Kyiv".



(b) Storify: "Protests In Kiev Turn Violent".



(c) Archive-It: Late Jan 2014, Only Four General Collections About Ukraine (all ca. 2009).



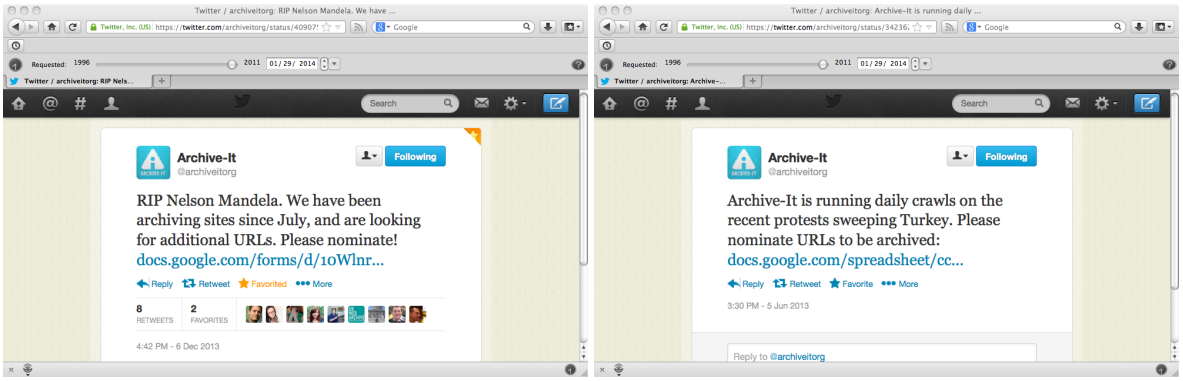
(d) Archive-It: Late Feb 2014, A Collection About the Ukrainian Conflict Added.

Figure 2: Coverage about Ukraine in Storify and Archive-It.

One reason why the collection shown in Figure 2(d) was not created until after the conflict had become well-advanced is that creating collections is arduous, and often requires manual input from subject experts. Figure 3 shows Archive-It asking the community at large for seed URI nominations to augment two new collections. Nelson Mandela (Figure 3(a)) was an international super star whose failing health was widely known, but why did protests in Turkey receive a collection (Figure 3(b)) and protests in Ukraine did not? This is a function of domain knowledge more than a judgment about importance; our intuition is that while the archiving community might have been aware of the Ukrainian protests, no one in the Archive-It community happened to have the knowledge of the the actors, history, and details of this particular event. In short: creating good collections is hard.

Figures 1, 2, and 3 illustrate two needs that we will address through the integration of "storytelling" social media and web archiving. First, we can use (semi-)automatically[6] created stories, similar to those in Figures 2(a) and 2(b), to summarize the holdings in large collections, like the

[6]We use the term "semi-automatically" because all our proposed tools will allow for manual insertion and deletion at the curator's discretion. In our experience, people are willing to edit automatically generated results as long as they are very close to expectations, but past a threshold people will consider them too broken to fix.



(a) Nominating seed URIs about Nelson Mandela. (b) Nominating seed URIs about Protests in Turkey.

Figure 3: Beginning collections in Archive-It.

one shown in Figure 1. Collections are already a topically-based, proper subset of the entire web, so similarly we can sample from the collections and provide one or more summaries or abstracts of the entire collection. The sampled web pages are then placed in an interface that users are already familiar with, such as Storify. A large collection will likely have many different "good" stories that summarize it, so we will consider a story to be "good" if a person considers it to be indistinguishable from a human-generated story – like a Turing Test for stories.

Second, to bootstrap a web archive collection, especially on fast-moving current events or topics in which the curator has limited domain knowledge, we can leverage one or more stories as the source for seed URIs to either create new collections or augment existing collections. For example, if Archive-It members sought to create an archive about the situation in the Ukraine, then the stories shown in Figures 2(a) and 2(b), as well as others (note the "Don't Miss" topical recommendations in the side bar) available at Storify, are likely good places to start collecting seed URIs.

Summarizing the content of web archive collections is difficult because there are two dimensions that need to be summarized: the URIs that comprise the collection (e.g., `cnn.com`, `bbc.co.uk`) and the archived copies (called "mementos") of those URIs at different times (e.g., `cnn.com`@t_1, `cnn.com`@t_5, `bbc.co.uk`@t_3). Either dimension by itself is difficult, but combined they present a number of challenges, and are hard to adapt to most conventional visualization techniques. We have explored applying well-known, advanced visual interfaces (e.g., timelines, Treemaps, wordles and tagclouds, bubble charts) for Archive-It collections and the results are sufficient for those already with an understanding of what is in the collection, but they do not facilitate an understanding to those unfamiliar with collection (see the supporting documents for examples) [32, 31].

One problem with the above approaches is there is not an emphasis on ignoring content: there is often an implicit assumption that everything in a collection is equally valuable and should be visualized. Some of the web pages change frequently and some are near-duplicates. Some go off-topic and no longer contribute to the collection. Collections grow quickly: the Human Rights collection in Figure 1 has nearly 600 seed URIs, and each URI has between one and 43 mementos. Visualization techniques with an emphasis on recall (i.e., "here's everything in the collection") do not scale. We believe increased textual metadata (e.g., Encoded Archival Description (EAD) [33])

added to the interface in Figure 1(a) is not the solution. Instead, we are informed by emerging trends in social media storytelling, which focus on a small number of exemplary pages (i.e., high precision) as chosen by a human. At the same time, services like Storify are not sufficient; in previous work we have measured that resources linked to in social media disappear (i.e., HTTP 404) at the rate of 11% per year [38, 39, 40].

This proposal supports the national digital platform in three ways. First, it will *increase access* to existing archives as well as facilitate creation of new collections within archives. The contents of the stories and the entire collections that they represent will be available as DPLA service hubs. Second, we will *increase the discoverability* by using popular Web 2.0 tools such as Storify. This is similar to the DPLA model of daily tweeting about randomly selected holdings with the hashtag "#dplafinds" – people are already familiar with Twitter, so placing content there attracts attention to the entire collection. Third, we will *improve the user experience* via additional UIs at the web archive that align with current interaction motifs.

2 Impact

We will combine two existing tools in an intelligent way. The goal of Archive-It is not necessarily crafting a story, but preserving content. The goal of Storify is not necessarily preserving content, but crafting a story. By combining Archive-It and Storify we can do both. Our research focus is to *explore information retrieval techniques to (semi-)automatically generate stories summarizing a collection that will approximate what a knowledgeable human would generate, as well as use stories to create or augment collections*.

One of the concerns in the web archiving world is how to generate more interest in and use of web archives. In a study last year, we showed that although the Internet Archive receives a lot of traffic[7], robots outnumber humans $10:1$ in accessing the Wayback Machine [3]. Furthermore, the humans that visit the Internet Archive's Wayback Machine typically visit a single page and then leave; depending on the source this can be as often as 64% of the time. In web analytics terminology, this is known as an undesirably high "bounce rate" [2]. In short, web archives are not well-known by the general web population (and are not indexed by search engines), and those who do know about web archives consider them difficult to use. We have worked on enriching APIs for web archives [5], but better APIs do not directly support increased archive exploration by humans. Rather than develop custom exploration interfaces for web archives, we plan to utilize existing interfaces, such as Storify, with which the public is already familiar. The collections will contain links to the story overviews (see Figure 1(d)), and the stories in Storify will link to the Archive-It collections.

Working with Archive-It is a perfect opportunity for ODU to disseminate our research results. We would have access to their knowledgeable engineers, and even more importantly to the entire Archive-It partner community, all of whom have an interest in increasing the usage and visibility of their collections. Furthermore, Archive-It staff members acknowledge that their collections are limited to the areas in which their partners have specific interest and expertise. We will have the benefit of their input throughout the project, not only in our regular communication with Archive-It but with the user community at the annual Archive-It partner meetings as well.

After the initial design and evaluation, we will work with Archive-It staff to integrate the story and collection creation tools into their partner's tool suite. The resulting one or more stories that

[7] `http://www.alexa.com/siteinfo/archive.org`

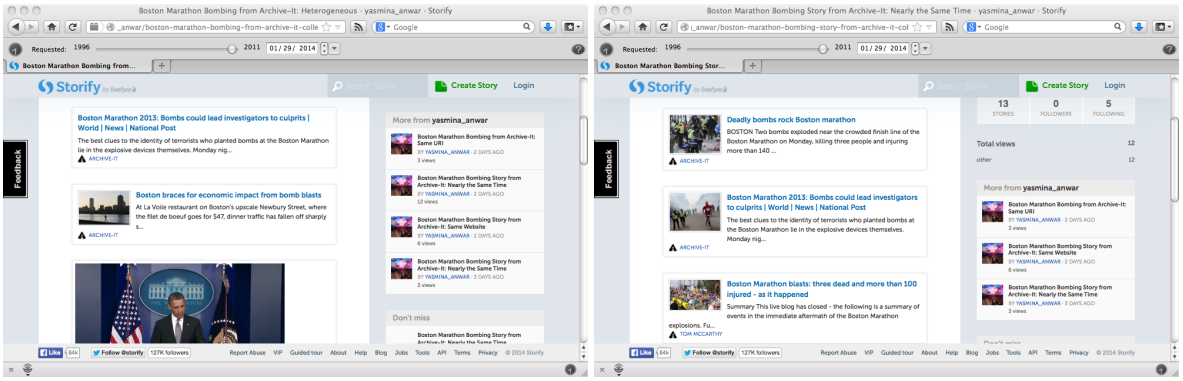
summarize the collection (whether ultimately in Storify or additional interfaces) will be linked from the collection page (Figure 1(a)). If the curators approve of the resulting stories, they will link them from the collection metadata pages, and if users find the stories useful the collections will enjoy increased traffic. Similarly, if the curators find existing stories (like those in Figures 2(a) and 2(b)) useful starting points for collections, then we should see an increase in the number and breadth of collections.

We will develop open source tools that *increase the use of archive collections as well as ease the creation of new collections*. The tools will be applicable for all Open Wayback Machine (the de facto standard in web archiving) users, as well as a variety of popular tools such as Storify. We will begin working within the Archive-It community because of its mature and highly-collaborative member environment. Although the tools themselves are targeted towards archivists and collection curators, the output of the tools (i.e., the stories) will be created by the curators and available to the general public.

3 Project Design

The detailed research plan along with milestones appears in the "schedule of completion", but there are two primary research goals for this project:

- Sample from existing collections to generate stories that summarize the collection (i.e., large → small).
- Use existing stories to generate or augment collections (i.e., small → large).



(a) Different URIs, different times. (b) Different URIs, same time.

Figure 4: Different kinds of stories that we created by manually selecting URIs from the Archive-It collection about the 2013 Boston Marathon bombing (collection 3649).

		Time: same	Time: different
URIs: same		differences in GeoIP, mobile, etc. ([12])	evolution of a single page (or domain) through time
URIs: different		different perspectives on a point in time	broadest possible coverage of a collection

Table 1: Four basic story types (others may be possible).

Figure 4 gives two examples of the kinds of stories that can be generated from an Archive-It collection[8]. These stories about the Boston Marathon Bombing were generated manually, but are representative of the kinds of stories that we will generate automatically. They simultaneously invite exploration and convey a sense of the "aboutness" of a collection better than what is found in Figure 1. In Figure 4(a), there is a broadly defined story that samples from different URIs and different times, while Figure 4(b) shows different URIs at approximately the same time. The former gives a summary of the entire collection, while the latter gives different perspectives around a particular point in time (in this case, shortly after the bombings were first reported). We will explore many different types of stories, some of which are listed in Table 1. It will be possible for collections to be summarized with more than one kind of story (depending on the nature of the collection as well as the curators preference). It is also possible that there are additional types of stories beyond those in Table 1 that we will discover when we create a quantitative baseline of story characteristics.

3.1 Creating Stories From Collections

Before we can create a story about a collection, we must compute the "aboutness" of individual pages in a collection as well as the collection itself. For discussion here, we will define the aboutness of a page P at time t to be a list of k terms ordered by decreasing weights[9], where k is some threshold (e.g., 5, 10, 20):

$$aboutness(P@t) = \{term_1, term_2, term_3, ..., term_k\} \quad (1)$$

We can compute $aboutness(P)$ for all times t in a similar fashion. Also, if collection C consists of various pages at various times:

$$C = \{\{P_1@t_1, P_1@t_5, ...\} + \{P_2@t_1, P_2@t_5, ...\} + ... + \{P_i@t_n, P_i@t_m, ...\}\} \quad (2)$$

We can then compute $aboutness(C)$ and $aboutness(C@t)$ according to equation 1. Given these measurable quantities, we can track how aboutness changes over time for both pages and collections. It is possible that a page's aboutness can change over time, but still fit within the larger theme of the collection. Some pages might be on topic for a period of time, but then go off topic for various reasons: changing interests on the part of the site owner, the domain registration is lost, the site is hacked, etc. In this case, the collection owner might choose to exclude off-topic pages from the summarization.

If we compute $aboutness(C)$ and then compute $aboutness(P) \forall P$, we can begin to select from candidate pages. The selection process will be influenced first by the kind of story the curator would like. For example, if we wish to emphasize how various networks initially reported on the Boston Marathon Bombing, then we select different URIs with a publishing date or modification date of approximately 3pm on April 15, 2013. If the goal is to summarize the entire collection, then the time frame would be expanded to the boundaries of the collection, and we might have a preference for page diversity, where $aboutness(P_1) \not\approx aboutness(P_2)$. We can imagine other criteria for picking candidate pages, such as: archival completeness (e.g., not missing embedded

[8]See our now famous Hurricane Katrina story `http://slidesha.re/1tPf9GE`.

[9]The terms will be discovered and weighted using standard Topic Modeling techniques, including Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), that address the term mismatch problem (e.g., "Kiev" and "Kyiv" are synonyms, "Turkey" is a homonym); see [8] for a review.

images, videos, or style sheets – see our work in automatically measuring archive "damage" [13, 9, 14, 1]), page popularity, generates a nice link preview, language preferences (e.g., do you separate or combine English and Arabic content in stories summarizing collections about the Egyptian Revolution?), and possibly many others.

We will perform the collection → story computations at ODU, on the ODU mirror of Archive-It's collections. Although it has not been publicly announced, ODU has created a dark archive of Archive-It's contents (ca. late 2013) and will in the future set up an update mechanism. Although we will not be running an interface open to the public, we will have it available to interested researchers. This will allow us to direct access to actual collections without interfering with Archive-It's production service.

We will be running an open source Wayback Machine[10]. ODU has received the Web Archive (WARC) [22, 15] files[11] from Archive-It and we will operate on a test set of collections from Archive-It. We anticipate that the tool will operate directly on the WARC files themselves, or possibly using the ODU-developed "ArcLink" Wayback extension software [5]; in no case will we be screen scraping the HTML pages hosted at `archive-it.org`.

3.2 Creating Collections From Stories

In going from story → collection, many of the same concepts will be used. In this case, we will compute $aboutness(S)$, where S is composed of the URIs in the story and optionally the resources linked to from the URIs in the story. For example, if a Archive-It curator happened upon stories similar to those in Figures 2(a) and 2(b), we would take the list of URIs in the story, and possibly the list of URIs linked in the story (e.g., news articles linked to in a tweet), as well as any descriptive metadata the story author has generated to determine what the story is about.

From this story, we can use the resulting URIs as a seed list for the beginnings of a collection. We will explore increasing the seed list by querying search engines with terms from $aboutness(S)$, looking in sites such as `trends.google.com` and `wikinews.org` to discover new terms, URIs, etc., looking for back-links in Google and Twitter to find what pages link to the pages in our story, and other techniques for expanding the story to become a full-fledged set of seed URIs from which we can generate a collection.

We will also study the quantitative characteristics of stories created by humans: What are the mean and median numbers of URIs that comprise a story: 10, 20, 50? How many URIs before "tl;dr" (too long; didn't read)? How many resources are linked to from the story, and what kinds are they (e.g., videos, news, social media)? What time frame do stories cover? Short-lived, topical events like the Boston Marathon Bombing, or on-going themed collections like Human Rights? How variable is the $aboutness(P)$ between each page in the story – are they all similar to each other, or do they cover a broad range of topics? How "archivable" are the resources in a story (in preliminary work we discovered that URIs shared via Twitter are less archivable than the typical seed URI in Archive-It [10])? In short, we have to better understand how humans create their stories so we can both create stories that look like what humans would create, and also understand any possible limitations of using stories to seed collections.

[10] See: `https://github.com/internetarchive/wayback`

[11] WARC files are a de facto standard in the web archiving community, they are similar to "tar" or "zip" files but are generated from web crawlers.

3.3 Prior Work

As mentioned above, we have explored visualizing Archive-It collections using standard visualization techniques, but feel they are ill-suited for the scale and temporal nature of web archives [32, 31]. In our "Just-in-time" preservation research we discovered new locations and alternate versions of web pages that are missing in the current web [16] by determining their aboutness using a variety of techniques, including using page titles [20], tags [19], and lexical signatures computed from archived versions [21, 18, 17], all of which could be used as queries to search engines to find replacement copies of the missing web page.

We also have significant experience with the subtle, multiple definitions of time that occur in web archiving. There is the creation date of the web page (which typically can only be estimated) [37], the modification time (which is sometimes available), the time the page was archived [29], and the time the page is about (e.g., a page published today about events in the past or future). We have begun development on a tool for identifying off-topic web pages in collections [4].

4 Project Resources: Personnel, Time, Budget

PI Dr. Michael L. Nelson has a distinguished career in digital libraries and web preservation. Since creating the first web-based digital library for NASA, he has moved to academia and co-authored a number of key digital library specifications such as OAI-PMH, OAI-ORE, Memento, and ResourceSync. At both NASA and ODU his research has always been involved augmenting the scholarly communications process with web technologies. His research has been supported by the NSF, NEH, NASA, Library of Congress, Sloan Foundation, and the Andrew Mellon Foundation.

Dr. Michele Weigle began her career studying computer networking and simulation and, with funding from the NSF, she helped to build the Intelligent Networking and Systems (INetS) research group at ODU. Because of an interest in web science and information visualization, she began collaborating with the Web Sciences and Digital Libraries research group at ODU in 2010. This collaboration has led to publications at the ACM/IEEE Joint Conference on Digital Libraries related to the coverage of web archives, visualization tools for Archive-It collections, and tools for personal web archiving. Dr. Weigle has also recently taught several well-received graduate courses in the area of information visualization.

Dr. Nelson and Dr. Weigle have graduated seven Ph.D. students total in the last five years, and have 12 current Ph.D. students at various stages of completion. More information about the Web Science and Digital Libraries Group can be found at `ws-dl.cs.odu.edu`.

Kristine Hanna is the Director of Archiving Services at the Internet Archive. She works with institutions and foundations to fund and build projects and programs; including Archive-It. In recent years, Kristine has served as the lead on numerous grant funded programs for the Internet Archive, working with partners including Virginia Tech on "Crisis Tragedy and Recovery Network" and University of Massachusetts, Amherst on the "Million Books Project". Kristine also serves on several steering committees for international organizations including the International Conference on Asian Digital Libraries (ICADL), and is actively involved in the newly launched "National Digital Stewardship Alliance" (NDSA), heading up the group identifying "at risk" selection for preservation and access.

The "Schedule of Completion" gives a more complete description of the tasks as well as their evaluation and deliverables. In high-level terms, ODU will take the lead on the development of software and initial evaluation of the resulting stories and collections, and Archive-It will take the

lead on their deployment and integration into their production services, as well as evaluation within their community of partners (i.e., subscribers).

The budget is primarily for ODU and Archive-It staff members, and will be managed by the ODU Research Foundation. For ODU, we will support two graduate students as well as a portion of faculty summer months (although faculty time is charged in terms of summer months, we will work on the project year-round). Archive-It charges based on subscription costs, but these costs include staff time, covering their deployment and evaluation activities. Existing computer hardware at ODU and Archive-It will be sufficient for development.

5 Communications Plan

As we have always done, we will make the resulting software and systems available as open source; some of our well-known tools include ArchiveFacebook, WARCreate, and Carbon Date[12]. We will also use our standing in the digital library community to support the research results in the form of peer reviewed publications, tutorials (typically at digital library conferences like JCDL and TPDL), and software. We have had great success with this in the past with OAI-PMH, OAI-ORE, Memento, and now ResourceSync.

Complementing the above, the Internet Archive is the primary player in the development of large-scale web archiving tools, developing open-source solutions such as the Heritrix crawler and the Wayback Machine archiving software. In addition to the public services the Internet Archive and Archive-It provide, their code also forms the basis for most of the web crawling activities performed by various national libraries and archives, via the International Internet Preservation Coalition (IIPC).

Archive-It already has an annual partners meeting where users meet to share lessons learned and discuss future directions. ODU has attended and presented at the last three of these meetings[13], so we are well-integrated into the user community. Archive-It is especially well-suited for training and dissemination, and they will greatly assist in ensuring our successes are rolled out to the collection curators, and then to the general public. We have had good results in the past using the Share-alike Creative Commons license and the GNU General Public License (GPL) and placing the code in places like SourceForge, Google Code, and GitHub. We have also used archived email lists, most recently using Google Groups, to establish a place for community discussion. We use a variety of tools, such as blogs and wikis for project coordination and dissemination. We will also use our standing in the digital library community to demonstrate and present the resulting work at workshops such as the annual Library of Congress NDIIPP meeting, the bi-annual Coalition for Networked Information (CNI) meetings, and the annual IIPC meetings.

Tools, projects, evaluations and other supporting activities will be incorporated in our graduate classes, including CS 751/851 “Introduction to Digital Libraries” and CS 725/825 “Information Visualization”. We will also introduce some of the concepts of preserving and exploring the social context of the past in the undergraduate classes of CS 312 “Internet Concepts”. Furthermore, this project will support two Ph.D students, and we will count it as a tremendous dissemination success if we are able to graduate two Ph.D. students with experience in designing, implementing, and evaluating the integration of storytelling techniques with web archiving technologies.

[12]For a review of 2014 software, popular media coverage, and other events from our research group, please see: `http://ws-dl.blogspot.com/2015/01/2015-01-03-review-of-ws-dls-2014.html`.

[13]See our trip reports: `bit.ly/1wsCyJz`, `bit.ly/1Atb4r1`, and `bit.ly/1FpgGIu`

6 Schedule of Completion

Table 2 provides the Gannt chart for the six primary tasks. The work is scheduled to begin in May 2015 (beginning of year one) and complete in April 2018 (end of year three).

Task	Year 1	Year 2	Year 3
Task 1: Baseline Storify, Archive-It	X		
Task 2: Ongoing Interface Review	X	X	X
Task 3: Generating Stories From Collections	X	X	X
Task 4: Generating Collections From Stories	X	X	X
Task 5: Metadata and Serialization		X	X
Task 6: Dissemination and Training		X	X

Table 2: Research Task Gantt Chart.

Task 1: Baseline Storify, Archive-It (lead: Nelson). We will begin by quantifying stories in Storify and collections in Archive-It. We need to understand the measurables of both stories and collections, as generated by humans, before we can generate them with our tools. For example, we will sample stories from Storify on a variety of topics:

- Mean and median length of resources in the stories. For example, the story shown in Figure 2(a) has 27 resources – is this too many? Too few? Just right?
- The nature of the resources. For example, the story shown in Figure 2(a) has four YouTube videos, five tweets, seven images, and the rest are from various international news sources. For this kind of story, is this a typical distribution of resources?
- How quickly do the resources linked to from stories become unavailable (HTTP 404)? In the past, we have measured the loss rate of linked resources in social media to be 11% per year [39, 40], but does that rate hold in Storify? Or do Storify users intuitively pick more stable resources for their stories?
- Are resources linked to from stories to "popular" sites (e.g., `cnn.com`)? Or are they to little-known outlets, blogs, and other sites that one would not typically discover on page one of a search engine result page?
- Are the resources "deep links" within a site (i.e., to a specific story or post)? Or are they to top-level sites?
- Does Table 1 account for all forms of stories? If so, what is the frequency distribution? Are there additional types of stories that commonly appear?

For Archive-It, we will perform a similar baseline but since we have access to all of the Archive-It collections (at least for our late-2013 snapshot), these measurements will be for the population instead of the sample:

- What is the mean and median number of URIs in a collection? What is the typical crawl depth and breadth for example, do the crawls stay at just the top-level site like cnn.com, or do they go deep into the sites as well?
- Using our "archivability" metric [10], are pages in Archive-It more or less archivable than those found in Storify? Our intuition is that they are less archivable (e.g., YouTube videos, which are difficult to archive), but we will quantify this assumption.
- Archive-It collections seem to consist of three different high-level categories: institutions (e.g., Alabama State Government web sites), topics (e.g., Human Rights), and events (e.g., Boston Marathon Bombing). How do these types of collections influence collection characteristics,

such as the types of pages chosen, the time frame of the crawl, and inter-page aboutness? Do different categories lend themselves to different kinds of stories (Table 1)?

Evaluation: In this task, we will measure stories in Storify (and any other similar sites) and collections in Archive-It and build a quantitative, descriptive model of them. Not only will we compare them with each other, but we will compare them to other sources, such as URIs sampled from Twitter, Wikipedia articles, etc.

Deliverables: We will publish the description in conventional publications as well as blog posts. Although this task is expected to be accomplished mostly in the first year, we will revisit the characteristics of stories and collections each year and update our findings accordingly.

Task 2: Ongoing Interface Review (lead: Weigle). Throughout this proposal we have primarily talked about Storify, but as was noted in the introduction, Facebook and Twitter have timeline services, and other services like `scoop.it` and `paper.li` can be used in a storytelling function as well. Task 1 will begin with stories from Storify, but we will continually survey the community for additional services that will likely emerge during the duration of this research. This fits within our goal of not defining new interfaces for web archives, but rather adapting popular, existing interface for web archives. As the public adapts and adopts, so will we.

Evaluation: This task is relatively simple: we will monitor developments in the communities of social media, web archiving, and storytelling. A Google search on "social media storytelling" is indicative of the interest in this area.

Deliverables: One of our PhD students has already performed an exhaustive review of services and tools available ca. early 2014 as part of her candidacy exam. We will make this available on our research group's blog, as well as new developments that occur in the duration of this project.

Task 3: Generating Stories From Collections (lead: Nelson). The key element of this task is choosing the "best" representative k samples, where k is specifiable by the collection curator, but is much smaller than the number of seed URIs and mementos in the collection. Suggested values of k will be determined by the result of Task 1, and other tunable parameters will include the timeline of the desired story (which may exclude some portions of the collection), page popularity (e.g., current and/or historical values of PageRank), memento quality (incomplete pages are not desirable candidates), aboutness spread (i.e., a range of topics or a narrowly defined topic), story type (cf. Table 1), etc.

Evaluation: We evaluate a variety of existing topic modeling software packages, such as MALLET [27], TMT [35, 34], and gensim [36], or adapt textbook methods in Python on our own as necessary. In year one, evaluation of the selected k URIs will be done within our research group, and we will compare them with those manually created stories of those with domain expertise. In years two and three will we gather feedback of the resulting stories from the curators, as well as crowdsource (e.g., Mechanical Turk) to see if the resulting stories are distinguishable from human generated stories. It is during this stage that we will determine if collections that are for institutions (e.g., state government) or broad topics (human rights) are suitable for storytelling techniques. If they are not suitable, we will restrict our focus to event-based stories and collections.

Deliverables: In year one we will have a command line tool that uses standard topic modeling libraries and analyzes local (ODU) WARC files for a collection and selects pages according to the parameters and story type specified. In years two and three, we will convert this into a web-based tool that can be hosted at Archive-It and can be made part of their suite of tools for their partners.

Task 4: Generating Collections From Stories (lead: Weigle). In a sense, a story selected from

Storify can be treated as a small collection C with no mementos (i.e., only live web URIs). Using this small set of URIs, we will augment them via search engine queries with terms from the computed aboutness, as well as crawling outward from the links that appear in the pages in the story (i.e., a breadth-first crawl). Performing a crawl with a particular aboutness in mind is known as a "focused crawl" [6, 7], and although focused crawls have been largely supplanted by search engine queries, we will see if there is trade-off for novelty vs. popularity (i.e., focused crawls might find less popular pages than search engine queries). In short, the purpose of this task is to automate the nomination function seen in Figure 3.

Evaluation: In year one, we will use our own tool to generate a handful of collections based on stories for which our research group feels competent to evaluate the output. We will generate collections that are similar to existing collections in Archive-It and measure the overlap, both in terms of aboutness and URIs. We establish a point of diminishing returns on crawl depth relative to aboutness (i.e., the more hops away from a seed URI, the more likely you are to go off-topic).

Deliverables: The tool in year one will be a command line tool that takes as arguments URIs of stories and extracts the first- and second-order pages that comprise the story. In year one, the resulting seed URIs will be manually transferred into the existing Archive-It collection creation procedure, but in year two and three we will work on a tighter integration into a web-based tool that can be hosted at Archive-It and can be made part of their suite of tools for their partners. In year three, we will focus on using this tool to augment existing collections in addition to creating new collections.

Task 5: Metadata and Serialization (lead: Nelson). We are unaware of a standard metadata format for expressing seed URIs (and URIs discovered from seed URIs) for a collection. Similarly, we are not aware of a machine-readable serialization for URIs that comprise a Storify story. Formats may emerge during this research, but if they do not we will investigate expressing collections of URIs as Open Archives Initiative Object Reuse and Exchange (OAI-ORE) "Aggregations" [23, 24, 25, 26], possibly with a new serialization of ORE in JSON-LD [42, 41].

Evaluation: Our goal will be to find a widely adopted metadata standard for serializing and transferring chronologically sorted lists of URIs and associated metadata, not necessarily to come up with the *best* such metadata format. If a winner emerges, we will adopt it. If not, we will propose one to the community.

Deliverables: We will publish the details of our adopted serialization format, both on our blog and in appropriate workshops like the Archive-It partners meeting. This task is reserved for years two and three because do not wish to prematurely adopt or promote any particular format.

Task 6: Dissemination and Training (lead: Hanna). The first year will be spent internally evaluating the tools listed in Tasks 3 and 4, and then in year two we will work with Archive-It staff with the goal for receiving feedback from the general partner community in the last half of year 2 and year 3. We will work with the Archive-It engineers to ensure a simple and smooth transition from the functionality of research software to something they can offer to their partners.

We will also encourage adoption of the tools in any Wayback Machine installation via tutorials at conferences like JCDL and additional communities like the IIPC.

Evaluation: If the tools are successful, they will be adopted not just by Archive-It and embraced by their partners, but other Wayback Machine users as well.

Deliverables: In addition to having the software available on `github.com` under a GPL license, we will also make the associated tutorial and training products available in the same manner.

References

[1] S. G. Ainsworth, M. L. Nelson, and H. V. de Sompel. A framework for evaluation of composite memento temporal coherence. Technical Report arXiv:1402.0928, 2014.

[2] Y. AlNoamany, A. AlSum, M. C. Weigle, and M. L. Nelson. Who and what links to the Internet Archive. In *Proceedings of Theory and Practice of Digital Libraries (TPDL)*, pages 346–357, 2013.

[3] Y. AlNoamany, M. C. Weigle, and M. L. Nelson. Access patterns for robots and humans in web archives. In *JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Lib raries*, pages 339–348, 2013.

[4] Y. AlNoamany, M. C. Weigle, and M. L. Nelson. Detecting off-topic pages in web archives. In *Submitted for publication*, 2015.

[5] A. AlSum and M. L. Nelson. ArcLink: Optimization techniques to build and retrieve the temporal web graph. In *JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Lib raries*, pages 377–378, 2013.

[6] D. Bergmark. Collection synthesis. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 253–262, 2002.

[7] D. Bergmark, C. Lagoze, and A. Sbityakov. Focused crawls, tunneling, and digital libraries. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 91–106, 2002.

[8] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[9] J. F. Brunelle, M. Kelly, H. S. M. C. Weigle, and M. L. Nelson. Not all mementos are created equal: Measuring the impact of missing resources. In *JCDL '14: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 321–330, 2014.

[10] J. F. Brunelle, M. Kelly, M. C. Weigle, and M. L. Nelson. The impact of JavaScript on archivability. *International Journal on Digital Libraries (accepted for publication)*, 2015.

[11] E. Cohen and C. Willis. One nation under radio: Digital and public memory after September 11. *New Media & Society*, 6(5):591–610, 2004.

[12] M. Kelly, J. F. Brunelle, M. C. Weigle, and M. L. Nelson. A method for identifying personalized representations in web archives. *D-Lib Magazine*, 19(11/12), 2013.

[13] M. Kelly, J. F. Brunelle, M. C. Weigle, and M. L. Nelson. On the change in archivability of websites over time. In *Proceedings of Theory and Practice of Digital Libraries (TPDL)*, pages 35–47, 2013.

[14] M. Kelly, M. L. Nelson, and M. C. Weigle. The archival acid test: Evaluating archive performance on advanced HTML and JavaScript. In *JCDL '14: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 25–28, 2014.

[15] M. Kelly and M. C. Weigle. WARCreate - create Wayback-consumable WARC files from any webpage. In *Proceeding of the 12th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pages 15–18, 2012.

[16] M. Klein. *Using the Web Infrastructure for Real Time Recovery of Missing Web Pages*. PhD thesis, Old Dominion University Department of Computer Science, 2011.

[17] M. Klein and M. L. Nelson. Revisiting lexical signatures to (re-)discover web pages. In *ECDL '08: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, pages 371 – 382, 2008.

[18] M. Klein and M. L. Nelson. Evaluating methods to rediscover missing web pages from the web infrastructure. In *JCDL '10: Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 59–68, 2010.

[19] M. Klein and M. L. Nelson. Find, new, copy, web, page - tagging for the (re-)discovery of web pages. In *Proceedings of TPDL*, pages 27–39, 2011.

[20] M. Klein, J. L. Shipman, and M. L. Nelson. Is This a Good Title? In *HT '10: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 3–12, 2010.

[21] M. Klein, J. Ware, and M. L. Nelson. Rediscovering missing web pages using link neighborhood lexical signatures. In *Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 137–140, 2011.

[22] J. Kunze. WARC: An Archiving format for the Web. In *5th International Web Archiving Workshop (IWAW'05)*, September 2005.

[23] C. Lagoze, H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, and S. Warner. ORE Specification and User Guide - Table of Contents, 2008. http://www.openarchives.org/ore/1.0/toc.

[24] C. Lagoze, H. Van de Sompel, P. Johnston, M. L. Nelson, R. Sanderson, and S. Warner. Object Re-Use & Exchange: A Resource-Centric Approach. Technical Report arXiv:0804.2273, 2008.

[25] C. Lagoze, H. Van de Sompel, P. Johnston, M. L. Nelson, R. Sanderson, and S. Warner. Adding eScience Assets to the Data Web. In *Proceedings of the Linked Data on the Web Workshop (LDOW 2009)*, 2009.

[26] C. Lagoze, H. Van de Sompel, M. L. Nelson, S. Warner, R. Sanderson, and P. Johnston. A web-based resource model for scholarship 2.0: object reuse & exchange. *Concurrency and Computation: Practice and Experience*, 24(18):2221–2240, 2012.

[27] A. K. McCallum. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002.

[28] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, September 2004.

[29] M. L. Nelson. Memento-Datetime is not Last-Modified. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html`, 2011.

[30] M. L. Nelson. A plan for curating "obsolete data or resources". Technical Report arXiv:1209.2664, 2012.

[31] K. Padia. Visualizing digital collections at archive-it. Master's thesis, Old Dominion University Department of Computer Science, 2012.

[32] K. Padia, Y. AlNoamany, and M. C. Weigle. Visualizing digital collections at Archive-It. In *Proceeding of the 12th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pages 437–438, 2012.

[33] D. V. Pitti. Encoded archival description: An introduction and overview. *D-Lib Magazine*, 5(11), 1999.

[34] D. Ramage and E. Rosen. Stanford Topic Modeling Toolbox. `http://nlp.stanford.edu/software/tmt/tmt-0.4/`, 2014.

[35] D. Ramage, E. Rosen, J. Chuang, C. D. Manning, and D. A. McFarland. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, volume 5, 2009.

[36] R. Rehurek. gensim: topic modeling for humans. `http://radimrehurek.com/gensim/`, 2014.

[37] H. SalahEldeen and M. L. Nelson. Carbon dating the web: estimating the age of web resources. In *Proceedings of TempWeb 2013*, 2013.

[38] H. M. SalahEldeen. Losing my revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone. `http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html`, 2012.

[39] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: How much social media content has been lost? In *Proceedings of TPDL*, pages 125–137, 2012.

[40] H. M. SalahEldeen and M. L. Nelson. Resurrecting my revolution: Using social link neighborhood in bringing context to the disappearing web. In *Proceedings of Theory and Practice of Digital Libraries (TPDL)*, pages 333–345, 2013.

[41] S. Soiland-Reyes, M. Gamble, and O. Corcho. ORE User Guide - Resource Map Implementation in JSON-LD, 2014. http://www.openarchives.org/ore/0.9/jsonld.

[42] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindstrom. JSON-LD 1.0: A JSON-based serialization for linked data. http://www.w3.org/TR/json-ld/, 2014.

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:
IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:
If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to complete this form.

Please indicate which of the following digital products you will create or collect during your project.
Check all that apply:

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☐	Digital content	Part II
☑	New software tools or applications	Part III
☐	A digital research dataset	Part IV

PART I.

A. Copyright and Intellectual Property Rights

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? http://us.creativecommons.org/

We will release our software under the GNU General Public License (GPL) https://www.gnu.org/licenses/gpl.html. We use this because Creative Commons does not recommend using a CC license for software.

The rest of our material is covered under the existing Archive-It operating policy (see: http://archive-it.org/).

A.2 What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.

We will not be producing any new digital content, only software.

We will publish the results of our research in the standard conferences (e.g., JCDL, TPDL), community meetings (e.g., Archive-It partner meetings, Library of Congress Digital Preservation workshop), and our research group's blog (http://ws-dl.blogspot.com/).

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

N/A

Part II: Projects Creating Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and the quantities of each type and format you will use.

A.2 List the equipment and software that you will use to create the content or the name of the service provider who will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, pixel dimensions).

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during your project and after the grant period.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created during your project (e.g., an Advanced Programming Interface, contributions to the DPLA or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide URL(s) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating New Software Tools or Applications

A. General Information

A.1 Describe the software tool or electronic system you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) the system or tool will serve.

We will be creating two main products:
1) A web-based interface for collection curators to analyze their collections and create stories that can be uploaded to a service such as Storify.
2) A tool inside Archive-It that can recommend new seed URIs for archiving based on already-created stories on Storify.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

To the best of our knowledge, no tools of this type exist.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

Python, HTML/CSS/JavaScript/JQuery
Linux
WARC, Heritrix, Wayback Machine
wget

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

The story-creating software will read WARCs from Archive-It collections, analyze the features of the collections, create metadata, and produce candidate URIs for creating stories in Storify. It will interoperate with both Archive-It and Storify.

The collection-enhancing software will read URIs from Storify and select candidates for seed URIs to be archived by Archive-It. It will interoperate with both Storify and Archive-It.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software or system you will create.

It will be designed to operate in a Heritrix and Wayback Machine environment (typically Linux/Unix), and thus we will design our tool within this environment

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software or system.

We maintain a private wiki for internal notes and discussions. This will be used for development documentation and recording the processes used in developing the tool.

In the past we have used Google Code and GitHub for hosting our source code as well as facilitating the release of technical documentation for our software. We have also used archived email lists, most recently using Google Groups, to establish a place for community discussion.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

A partial list is available at http://ws-dl.cs.odu.edu/Main/Software/
We have released several software tools. Direct URLs to the repositories for some of our software are below:
* Warrick - http://code.google.com/p/warrick/
* WAIL - http://github.com/machawk1/wail/
* mcurl - http://github.com/aalsum/mcurl/
* CarbonDate - https://github.com/HanySalahEldeen/CarbonDate
* HTTPMailbox - https://github.com/ibnesayeed/HTTPMailbox
* Memento Plugin for MediaWiki - https://github.com/hariharshankar/mediawiki/

C. Access and Use

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

As we have done with previous software, we will place our source code on Google Code or GitHub with a GNU GPL license. This license is available at https://www.gnu.org/licenses/gpl.html

C.2 Describe how you will make the software or system available to the public and/or its intended users.

The source code will be hosted on a public site such as Google Code or GitHub. We will provide a link to the software from our research group's website at http://ws-dl.cs.odu.edu/Main/Software

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

2. Does the proposed research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity already been approved? If not, what is your plan for securing approval?

3. Will you collect any personally identifiable information (PII) about individuals or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release (e.g. data anonymization, suppression of personally identifiable information, synthetic data).

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

6. What documentation will you capture or create along with the dataset(s)? What standards or schema will you use? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

7. What is the plan for archiving, managing, and disseminating data after the completion of research activity?

8. Identify where you will be publicly depositing dataset(s):

Name of repository: _____

URL: _____

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

6 Schedule of Completion

Table 2 provides the Gannt chart for the six primary tasks. The work is scheduled to begin in May 2015 (beginning of year one) and complete in April 2018 (end of year three).

Task	Year 1	Year 2	Year 3
Task 1: Baseline Storify, Archive-It	X		
Task 2: Ongoing Interface Review	X	X	X
Task 3: Generating Stories From Collections	X	X	X
Task 4: Generating Collections From Stories	X	X	X
Task 5: Metadata and Serialization		X	X
Task 6: Dissemination and Training		X	X

Table 2: Research Task Gantt Chart.

Task 1: Baseline Storify, Archive-It (lead: Nelson). We will begin by quantifying stories in Storify and collections in Archive-It. We need to understand the measurables of both stories and collections, as generated by humans, before we can generate them with our tools. For example, we will sample stories from Storify on a variety of topics:

- Mean and median length of resources in the stories. For example, the story shown in Figure 2(a) has 27 resources – is this too many? Too few? Just right?
- The nature of the resources. For example, the story shown in Figure 2(a) has four YouTube videos, five tweets, seven images, and the rest are from various international news sources. For this kind of story, is this a typical distribution of resources?
- How quickly do the resources linked to from stories become unavailable (HTTP 404)? In the past, we have measured the loss rate of linked resources in social media to be 11% per year [39, 40], but does that rate hold in Storify? Or do Storify users intuitively pick more stable resources for their stories?
- Are resources linked to from stories to "popular" sites (e.g., `cnn.com`)? Or are they to little-known outlets, blogs, and other sites that one would not typically discover on page one of a search engine result page?
- Are the resources "deep links" within a site (i.e., to a specific story or post)? Or are they to top-level sites?
- Does Table 1 account for all forms of stories? If so, what is the frequency distribution? Are there additional types of stories that commonly appear?

For Archive-It, we will perform a similar baseline but since we have access to all of the Archive-It collections (at least for our late-2013 snapshot), these measurements will be for the population instead of the sample:

- What is the mean and median number of URIs in a collection? What is the typical crawl depth and breadth for example, do the crawls stay at just the top-level site like cnn.com, or do they go deep into the sites as well?
- Using our "archivability" metric [10], are pages in Archive-It more or less archivable than those found in Storify? Our intuition is that they are less archivable (e.g., YouTube videos, which are difficult to archive), but we will quantify this assumption.
- Archive-It collections seem to consist of three different high-level categories: institutions (e.g., Alabama State Government web sites), topics (e.g., Human Rights), and events (e.g., Boston Marathon Bombing). How do these types of collections influence collection characteristics,

such as the types of pages chosen, the time frame of the crawl, and inter-page aboutness? Do different categories lend themselves to different kinds of stories (Table 1)?

Evaluation: In this task, we will measure stories in Storify (and any other similar sites) and collections in Archive-It and build a quantitative, descriptive model of them. Not only will we compare them with each other, but we will compare them to other sources, such as URIs sampled from Twitter, Wikipedia articles, etc.

Deliverables: We will publish the description in conventional publications as well as blog posts. Although this task is expected to be accomplished mostly in the first year, we will revisit the characteristics of stories and collections each year and update our findings accordingly.

Task 2: Ongoing Interface Review (lead: Weigle). Throughout this proposal we have primarily talked about Storify, but as was noted in the introduction, Facebook and Twitter have timeline services, and other services like `scoop.it` and `paper.li` can be used in a storytelling function as well. Task 1 will begin with stories from Storify, but we will continually survey the community for additional services that will likely emerge during the duration of this research. This fits within our goal of not defining new interfaces for web archives, but rather adapting popular, existing interface for web archives. As the public adapts and adopts, so will we.

Evaluation: This task is relatively simple: we will monitor developments in the communities of social media, web archiving, and storytelling. A Google search on "social media storytelling" is indicative of the interest in this area.

Deliverables: One of our PhD students has already performed an exhaustive review of services and tools available ca. early 2014 as part of her candidacy exam. We will make this available on our research group's blog, as well as new developments that occur in the duration of this project.

Task 3: Generating Stories From Collections (lead: Nelson). The key element of this task is choosing the "best" representative k samples, where k is specifiable by the collection curator, but is much smaller than the number of seed URIs and mementos in the collection. Suggested values of k will be determined by the result of Task 1, and other tunable parameters will include the timeline of the desired story (which may exclude some portions of the collection), page popularity (e.g., current and/or historical values of PageRank), memento quality (incomplete pages are not desirable candidates), aboutness spread (i.e., a range of topics or a narrowly defined topic), story type (cf. Table 1), etc.

Evaluation: We evaluate a variety of existing topic modeling software packages, such as MALLET [27], TMT [35, 34], and gensim [36], or adapt textbook methods in Python on our own as necessary. In year one, evaluation of the selected k URIs will be done within our research group, and we will compare them with those manually created stories of those with domain expertise. In years two and three will we gather feedback of the resulting stories from the curators, as well as crowdsource (e.g., Mechanical Turk) to see if the resulting stories are distinguishable from human generated stories. It is during this stage that we will determine if collections that are for institutions (e.g., state government) or broad topics (human rights) are suitable for storytelling techniques. If they are not suitable, we will restrict our focus to event-based stories and collections.

Deliverables: In year one we will have a command line tool that uses standard topic modeling libraries and analyzes local (ODU) WARC files for a collection and selects pages according to the parameters and story type specified. In years two and three, we will convert this into a web-based tool that can be hosted at Archive-It and can be made part of their suite of tools for their partners.

Task 4: Generating Collections From Stories (lead: Weigle). In a sense, a story selected from

Storify can be treated as a small collection C with no mementos (i.e., only live web URIs). Using this small set of URIs, we will augment them via search engine queries with terms from the computed aboutness, as well as crawling outward from the links that appear in the pages in the story (i.e., a breadth-first crawl). Performing a crawl with a particular aboutness in mind is known as a "focused crawl" [6, 7], and although focused crawls have been largely supplanted by search engine queries, we will see if there is trade-off for novelty vs. popularity (i.e., focused crawls might find less popular pages than search engine queries). In short, the purpose of this task is to automate the nomination function seen in Figure 3.

Evaluation: In year one, we will use our own tool to generate a handful of collections based on stories for which our research group feels competent to evaluate the output. We will generate collections that are similar to existing collections in Archive-It and measure the overlap, both in terms of aboutness and URIs. We establish a point of diminishing returns on crawl depth relative to aboutness (i.e., the more hops away from a seed URI, the more likely you are to go off-topic).

Deliverables: The tool in year one will be a command line tool that takes as arguments URIs of stories and extracts the first- and second-order pages that comprise the story. In year one, the resulting seed URIs will be manually transferred into the existing Archive-It collection creation procedure, but in year two and three we will work on a tighter integration into a web-based tool that can be hosted at Archive-It and can be made part of their suite of tools for their partners. In year three, we will focus on using this tool to augment existing collections in addition to creating new collections.

Task 5: Metadata and Serialization (lead: Nelson). We are unaware of a standard metadata format for expressing seed URIs (and URIs discovered from seed URIs) for a collection. Similarly, we are not aware of a machine-readable serialization for URIs that comprise a Storify story. Formats may emerge during this research, but if they do not we will investigate expressing collections of URIs as Open Archives Initiative Object Reuse and Exchange (OAI-ORE) "Aggregations" [23, 24, 25, 26], possibly with a new serialization of ORE in JSON-LD [42, 41].

Evaluation: Our goal will be to find a widely adopted metadata standard for serializing and transferring chronologically sorted lists of URIs and associated metadata, not necessarily to come up with the *best* such metadata format. If a winner emerges, we will adopt it. If not, we will propose one to the community.

Deliverables: We will publish the details of our adopted serialization format, both on our blog and in appropriate workshops like the Archive-It partners meeting. This task is reserved for years two and three because do not wish to prematurely adopt or promote any particular format.

Task 6: Dissemination and Training (lead: Hanna). The first year will be spent internally evaluating the tools listed in Tasks 3 and 4, and then in year two we will work with Archive-It staff with the goal for receiving feedback from the general partner community in the last half of year 2 and year 3. We will work with the Archive-It engineers to ensure a simple and smooth transition from the functionality of research software to something they can offer to their partners.

We will also encourage adoption of the tools in any Wayback Machine installation via tutorials at conferences like JCDL and additional communities like the IIPC.

Evaluation: If the tools are successful, they will be adopted not just by Archive-It and embraced by their partners, but other Wayback Machine users as well.

Deliverables: In addition to having the software available on `github.com` under a GPL license, we will also make the associated tutorial and training products available in the same manner.

Original Preliminary Proposal

1 Project Directors

Michael L. Nelson (www.cs.odu.edu/~mln), Associate Professor of Computer Science at Old Dominion University (ODU) will be the PI. The Co-PIs will be Michele C. Weigle (www.cs.odu.edu/~mweigle), Associate Professor of Computer Science and Kristine Hanna (www.linkedin.com/pub/kristine-hanna/0/46/198), Director of Archiving Services at the Internet Archive (IA). This project brings together leading researchers and practitioners in the field of web archiving: Nelson is renown for the Memento Framework, OAI-PMH, and other infrastructure-level contributions to archives and repositories, Weigle is an expert in web-based visualization and UIs, and Hanna directs the IA's "Archive-It" subscription service with over 300 international members. The PIs have a number of successful past and current collaborations, including the NEH-funded research in tools for personal archiving (bit.ly/odu-dhig-2014) and form a uniquely qualified team for supporting a national digital platform.

2 Proposed Work Plan

This Research Grant proposal focuses on investigating how *storytelling* techniques and technologies can be used to increase the discoverability of collections of archived web pages. Currently, organizations and individuals around the world create collections of archived web pages to preserve the discourse surrounding news, events, organizations, and other culturally significant actors. While there are mature tools for collecting and preserving web pages, tools for post-archiving use, discovery, and exploration remain limited. Our goals are to use storytelling techniques to better summarize existing collections and to create new thematic collections of archived web pages.

We will address two main research thrusts: 1) summarize existing collections, and 2) bootstrapping new collections. When an archivist creates a collection, it can include 1000s of "seed" URLs. Over time, each of these URIs can be crawled 100s or 1000s of times, resulting in a collection having thousands to millions of archived web pages. Understanding the contents and boundaries of a collection is then difficult for most people, resulting in the paradox of the larger the collection, the harder it is to use. We will develop techniques to automatically (with optional human review and "steering") sample pages from a collection that summarize and describe the collection at large. For example, given a collection of many 1000s of pages, our tool will automatically select 20–30 representative pages that will then be linked in storytelling web applications, such as Storify or Paper.li. Although page selection is not dependent on tools such as Storify, we are committed to the approach of using existing tools and applications instead of developing new ones. Collection curators will also have the ability to export the sampled URIs and import them into whatever storytelling tool they prefer.

Similarly, archivists wanting to create collections about breaking events (e.g., Ebola, enterovirus D68) need to have domain knowledge about the event to create a quality collection. Since users around the world are already creating stories in places like Storify about these events (but without an archival component), we will develop tools that allow archivists to mine existing public stories to quickly generate seed URLs for a collection.

3 Relevance to the National Digital Platform

This proposal supports the national digital platform in three ways. First, it will *increase access* to existing archives as well as facilitate creation of new collections within archives. The contents of

the stories and the entire collections that they represent will be available as DPLA service hubs. Second, we will *increase the discoverability* by using popular Web 2.0 tools such as Storify and Paper.li. This is similar to the DPLA model of daily tweeting about randomly selected holdings with the hashtag "#dplafinds" – people are already familiar with Twitter, so placing content there attracts attention to the entire collection. Third, we will *improve the user experience* via additional UIs at the web archive that align with current interaction motifs.

4 Potential Impact

Web archives currently represent a significant investment, both in terms of hardware but more importantly in terms of archivist time. Despite the fact we know these archives contain valuable cultural information, access remains below its potential because tools remain experimental and isolated. Drawing on our past experiences, we will develop open source tools that *increase the use of archive collection as well as ease the creation of new collections*. The tools will be applicable for all Open Wayback Machine (the de facto standard in web archiving) users, as well as a variety of popular tools such as Storify and Paper.li. We will begin working within the Archive-It community because of its mature and highly-collaborative member environment. Although the tools themselves are targeted towards archivists and collection curators, the output of the tools (i.e., the stories) will be available to the general public.

5 Performance Goals and Outcomes

What makes a good story is a matter of human judgement and difficult to evaluate. Inspired by the Turing Test, one of our performance goals will be our automatically created stories being indistinguishable to general users from hand-crafted stories by expert archivists. Similarly, our automatically created collections should be indistinguishable from carefully selected collections by experts.

By studying existing user-created stories in places like Storify, we will be able to profile different kinds of stories by examining the typical length (in terms of pages included), topical diversity, timeframes covered, structural metadata (e.g., Pagerank, images and video, social media vs. news) and other features that will inform our creation and mining of stories. In this way, we expect to be able to intelligently guide our tools to also generate popular stories.

In addition to project's contributions in information retrieval, text mining, and web archiving, we will also demonstrate that web archives can increase discovery, access, and the quality of user experience by using tools and methods (e.g., Twitter, Storify) with which users are already familiar as a bridge between the current and past web. Our partners at Archive-It will help to evaluate the impact of the developed stories in terms of numbers of users and web referrals to the archive.

6 Estimated Budget

For this three year project the total budget for IMLS will be approximately $470,000. Although not required for research projects, ODU will cost share an additional total of approximately $107,000. These funds will cover the time of the ODU PI and Co-PI, the Archive-It Co-PI, two graduate students, and the associated travel and related expenses for engagement with the Archive-It community specifically, as well as the web archiving community at large via additional organizations like the International Internet Preservation Consortium (IIPC, of which both ODU and the Internet Archive are members) and conferences like iPres, Open Repositories, and JCDL.