Int J Digit Libr (2016) 17:239–256
DOI 10.1007/s00799-016-0185-3

Characteristics of social media stories

What makes a good story?

Yasmin AlNoamany[1] · **Michele C. Weigle**[1] · **Michael L. Nelson**[1]

Received: 11 January 2016 / Revised: 24 June 2016 / Accepted: 4 July 2016 / Published online: 21 July 2016

Abstract An emerging trend in social media is for users to create and publish "stories", or curated lists of Web resources, with the purpose of creating a particular narrative of interest to the user. While some stories on the Web are automatically generated, such as Facebook's "Year in Review", one of the most popular storytelling services is "Storify", which provides users with curation tools to select, arrange, and annotate stories with content from social media and the Web at large. We would like to use tools, such as Storify, to present (semi-)automatically created summaries of archival collections. To support automatic story creation, we need to better understand as a baseline the structural characteristics of popular (i.e., receiving the most views) human-generated stories. We investigated 14,568 stories from Storify, comprising 1,251,160 individual resources, and found that popular stories (i.e., top 25 % of views normalized by time available on the Web) have the following characteristics: 2/28/1950 elements (min/median/max), a median of 12 multimedia resources (e.g., images, video), 38 % receive continuing edits, and 11 % of their elements are missing from the live Web. We also checked the population of Archive-It collections (3109 collections comprising 305,522 seed URIs) for better understanding the characteristics of the collections that we intend to summarize. We found that the resources in human-generated stories are different from the resources in Archive-It collections. In summarizing a collection, we can only choose from what is archived (e.g., twitter.com is popular in Storify, but rare in Archive-It). However, some other characteristics of human-generated stories will be applicable, such as the number of resources.

Keywords Stories · Storify · Storytelling · Social media · Curation · Collections · Archive · Social networks · Archive-It

✉ Yasmin AlNoamany
yasmin@cs.odu.edu

Michele C. Weigle
mweigle@cs.odu.edu

Michael L. Nelson
mln@cs.odu.edu

[1] Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA

1 Introduction

Storify is a social networking service launched in 2010 that allows users to create a "story" of their own choosing, consisting of manually chosen Web resources, arranged with a visually attractive interface, and clustered together with a single URI and suitable for sharing. It provides a graphical interface for selecting uniform resource identifiers (URIs) of Web resources and arranging the resulting snippets and previews (see Fig. 1), with a special emphasis on social media (e.g., Twitter, Facebook, YouTube, and Instagram). We call these previews of Web resources "Web elements", and we call the annotations Storify allows on these previews "text elements".

We would like to use Storify to present the (semi-) automatically created summaries of collections of archived Web pages in a social media interface that is more familiar to users (as opposed to custom interfaces for summaries, e.g., [26]). Since the stories in Storify are created by humans, we model the structural characteristics of these stories, with particular emphasis on "popular" stories (i.e., the top 25 % of views, normalized by time available on the Web). Based on this paper's analysis of the structural characteristics of "good" stories, we have been able to automatically construct summarizing stories of Archive-It collections that are indis-

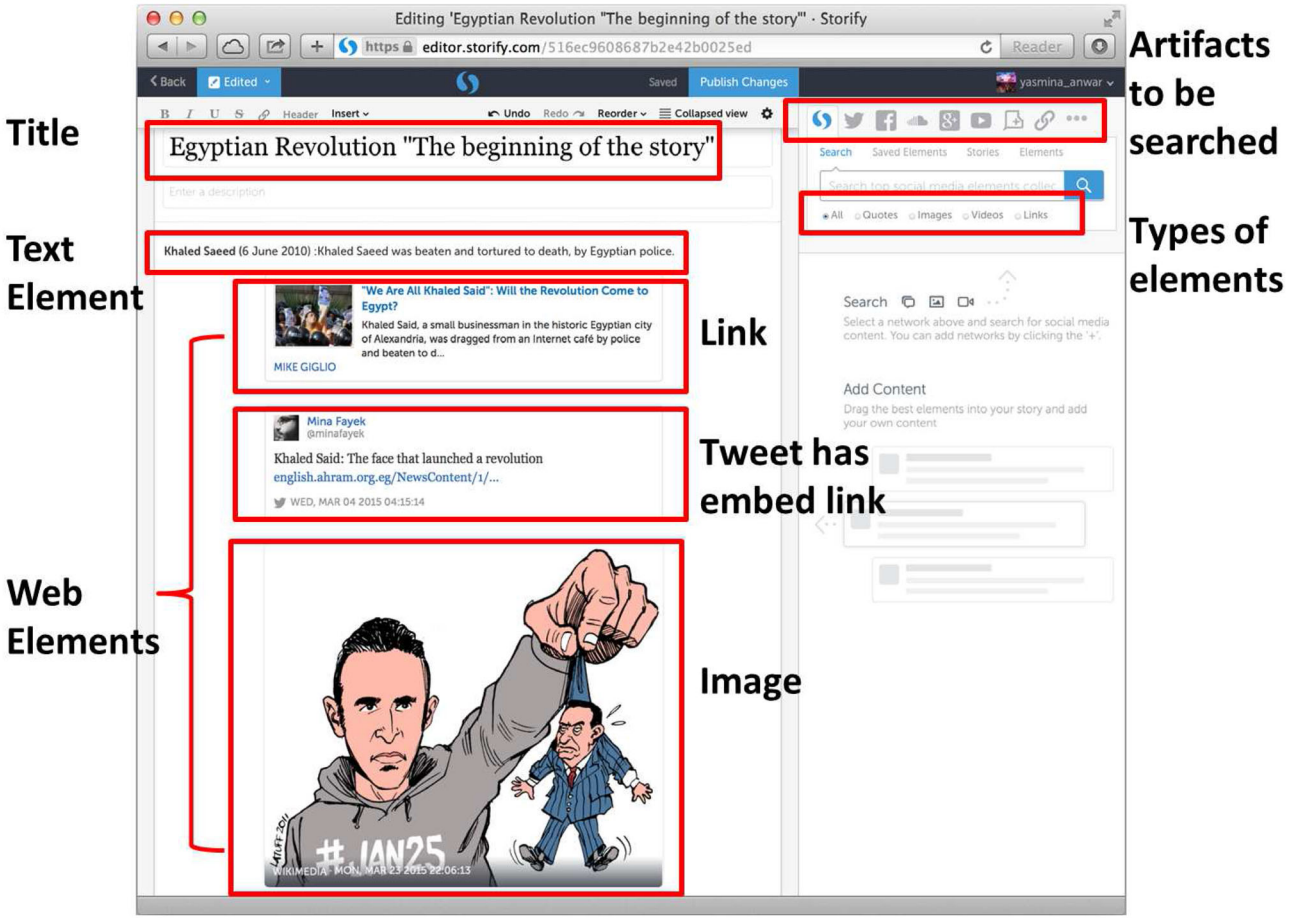


Fig. 1 Example of creating a story on Storify shows the Storify-defined categories for resources of the stories

tinguishable from those created by human subject domain experts, while at the same time, both kinds of stories (automatic and human) are easily distinguished from randomly generated stories [2].

In our previous work [4] and extended in this paper, we build a baseline for what human-generated stories look like and specify the characteristics of the popular stories. We answered the following questions: what is the length of the human-generated stories? What are the types of resources used in these stories? What are the most frequently used domains in the stories? What is the editing time of the stories? Is there a relationship between the editing time and features of the story? Is there a relationship between the popularity of the stories and the number of elements? Is there a relationship between the popularity and the number of subscribers of the authors? What differentiates the popular stories? How many of the resources in these stories disappear every year? Can we find these missing resources in the archives?

To answer these questions, we analyzed 14,568 stories from Storify comprising 1,251,160 elements. We found that popular stories have a min/median/max values of 2/28/1950 elements, with the unpopular stories having 2/21/2216 elements. Popular stories have a median of 12 multimedia resources (the unpopular stories have a median of 7), 38 % receiving continuing edits (as opposed to 35 %), and only 11 % of Web elements are missing on the live Web (as opposed to 13 %). The authors of popular stories have min/median/max values of 0/16/1,726,143 subscribers, while the authors of unpopular stories have 0/2/2469 subscribers. We found that there is a nearly linear relationship between the editing time of the story and the number of Web elements. We found that twitter.com dominates the Web resources of Storify stories. We also found that only 11 % of the missing resources could be found in public Web archives.

Since archived collections will be our source for creating stories, we want to understand the characteristics of these collections and what are the most used resources in these collections. In this paper, we extend our analysis to build a baseline of what is inside the archived collections then compare and contrast human-generated stories with curated archived collections of Web resources from Archive-It.[1] Based on the analysis of 3109 collections with 305,522

[1] http://www.archive-it.org/.

seed URIs, we found that the resources of archived collections are mostly dissimilar to the resources in Storify stories. Governmental and educational domains dominate the most frequent domains in the collections.

2 Background

2.1 Related Work

There have been many studies on how social media are being used in social curation [9,25,27,36,37]. Seitzinger defined social curation as "the discovery, selection, collection and sharing of digital artifacts by an individual for a social purpose such as learning, collaboration, identity expression or community participation [31]."

Duh et al. [9] studied how Togetter, a popular curation service in Japan, was being used for the social curation of microblogs, such as tweets. They studied the motivation of the curator by defining the topics being curated. They found that there is a diverse number of topics and a variety of social purposes for content curation, such as summarizing an event and discussing TV shows.

Many studies have been conducted to study data curation using data sets from Pinterest [11,30,37]. Zhong et al. [37] studied why and how people curate using data sets from Pinterest in January 2013 and the month of December 2012 from Last.fm. They found that curation tends to focus on items that may not be highly ranked in popularity and search rankings, which slightly contradicts our finding in Sect. 4.4. They also found that curation tends to be a personal activity more than being social.

Storify has been used in many studies by journalists [32] and also to explore how curation works in the classroom [17, 21]. Cohen et al. believe that Storify can be used to encourage students to become empowered storytellers and researchers [8]. Laire et al. [17] used Storify to study the effect of social media on teaching practices and writing activities.

Stanoevska-Slabeva et al. [32] sampled 450 stories from Storify about the Arab Spring from December 2010 to the end of August 2011. They found that social media curation is done by professionals as well as amateurs. They also found that the longer coverage stories use more resources. They also found that the stories created by both professionals and amateurs presented the primary gatewatching characteristics.

Kieu et al. [12] proposed a method for predicting the popularity of social curation content based on a data set from Storify. They specify the popularity of social curation using the number of views of the content. They used a machine learning approach based on curator and curation features (for example, the number of followers, the number of stories for the users, and the time that the user started using Storify) from stories. They found that combining the curator features (social features) and the curation features (content features) improves the performance of predicting the popularity.

In this paper, we also investigate if the number of views affects the features of the story, such as the number of elements. We consider the popularity of the story as a function of the number of views normalized by its time of existence on the live Web. Based on the popularity, we differentiate the characteristics of popular and unpopular stories in Storify.

Resources on the Web are known to disappear quickly [13–15,29], and the mean lifetime of a page is short (between 44 and 190 days) [7,18,20]. In a previous study, we found that most people come to the Web archives, because they did not find the pages on the live Web [3]. In this paper, we check the existence of the URIs in social networks and the seed URIs from Archive-It collections on the live Web. We also check the decay rate in Storify stories.

2.2 Memento terminology

In this section, we explain the terminology that we adopt in the rest of the paper. Memento [35] is an HTTP protocol extension which enables time travel on the Web by linking the current resources with their prior state. Memento defines the following terms:

- URI-R identifies the original resource. It is the resource, as it used to appear on the live Web. A URI-R may have 0 or more mementos (URI-Ms).
- URI-M identifies an archived snapshot, or a "memento", of the URI-R at a specific datetime, which is called Memento Datetime, e.g., URI-M_i = URI-R@t_i.
- URI-T identifies a TimeMap, a resource that provides a list of mementos (URI-Ms) for a URI-R with their Memento-Datetimes

3 Constructing the data set of storify stories

We created the data set by querying the Storify Search API[2] with the 1000 most frequent English keywords issued to Yahoo.[3] This set of available search keywords allowed us to gather sets of stories about many different topics. This was especially useful, since we do not know the ranking algorithm used by Storify search. We retrieved 400 results for each keyword, resulting in a total of 145,682 stories downloaded in the JavaScript Object Notation (JSON) format [33]. We created the data set in February 2015 and only considered stories authored in 2014 or earlier, resulting in 37,486 stories. We eliminated stories with zero or one elements or

[2] http://dev.storify.com/api/.

[3] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

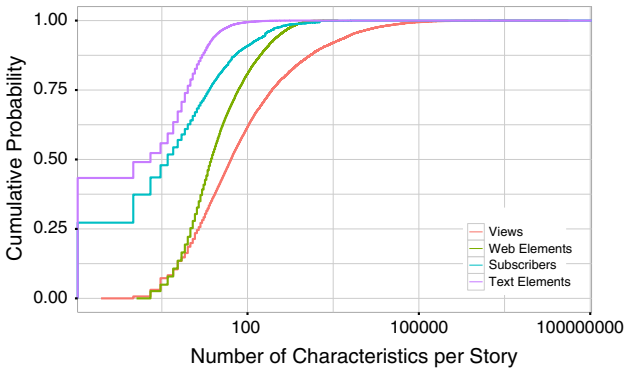


Fig. 2 Distribution of the characteristics of the 14,568 Storify stories analyzed

zero views, resulting in 14,568 unique stories authored by 10,199 unique users and containing a total of 1,251,160 Web and text elements.

4 Characteristics of human-generated stories

Figure 2 contains the distribution of the number of views of the stories, the number of Web elements, the number of text elements, and the number of subscribers. We notice that around 48 % of the stories do not have any text elements. This indicates that only about half of the stories are annotated with descriptive text.

For a closer look at the features of the stories, we present the distribution percentiles along with means of story views, Web and text elements, and number of subscribers for the story authors in Table 1. We show the distribution percentiles along with means because the distribution of the data is long-tailed. The editing time is the time interval (in h), in which users edit their stories and is calculated by taking the difference between the story creation date and the last-modified date. The median for all stories is 23 Web elements and 1 text element, and 44 % of the stories have no text elements at all. Due to the large range of values, we believe median is a better indicator of "typical values".

4.1 What kinds of resources are in stories?

Using the Storify-defined categories reflected in the Storify user interface (Fig. 1), the 1,251,160 elements consist of 70.8 % links, 18.4 % images, 8.1 % text, 2.0 % videos, and 0.7 % quotes. Text elements are relatively rare, meaning that few users choose to annotate the Web elements in their story.

4.2 What domains are used in stories?

The Web elements in Storify stories represent 91.95 % (1,150,399 out of 1,251,160) of all the resources. To analyze the distribution of domains in stories, we canonicalized the domains (e.g., www.cnn.com → cnn.com) and dereferenced all shortened URIs (e.g., t.co, bit.ly) to the URIs of the final locations. This resulted in 25,947 unique domains in the 14,568 unique stories.

Figure 3 shows the relationship between the frequency of the domains and the number of stories they appeared in. For example, the rightmost dot at the top of the graph represents the most frequent domain in the stories (twitter.com), which also appeared in the largest number of stories. This domain appears almost 1,000,000 times in over 10,000 dif-

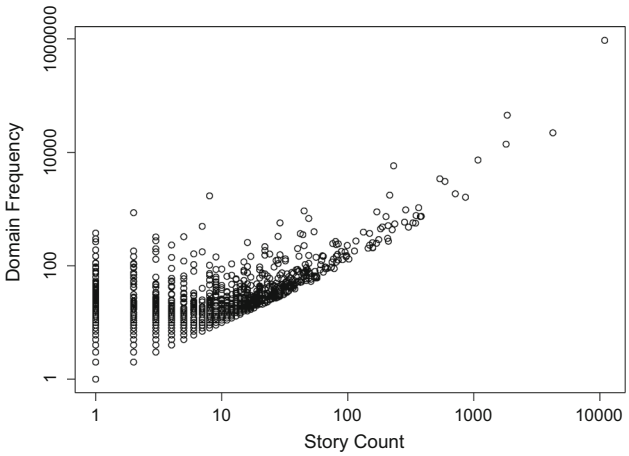


Fig. 3 Relationship between the frequency of the domains in Storify stories and the number of stories in which those domains appear

Table 1 Distribution of features of the stories in the data set

Features	Views	Web elements	Text elements	Subscribers	Editing Time
25th percentile	14	10	0	0	0.18
50th percentile	51	23	1	4	3
75th percentile	268	69	9	21	120
90th percentile	1949	210	19	85	1747
Maximum	11,284,896	2216	559	1,726,143	36,111
Mean	3790	80	8	286	855
SD	99,226	158	18	20,220	2982

Editing time is measured in hours

Table 2 Top 25 domains based on the frequency of appearance in Storify stories

Domain	Frequency	Percentage of Domains (%)	Story count	Alexa global rank as of 2015–03	Category
twitter.com	943,859	82.05	10,914	8	Social media
instagram.com	45,188	3.93	1841	25	Photos
youtube.com	22,076	1.92	4238	3	Videos
facebook.com	13,930	1.21	1802	2	Social media
flickr.com	7317	0.64	1079	126	Photos
patch.com	5783	0.50	231	2096	News
plus.google.com	3413	0.30	537	1	Social media
tumblr.com	3066	0.27	590	31	Blogs
blogspot.com	1857	0.16	713	18	Blogs
imgur.com	1756	0.15	215	36	Photos
coolpile.com	1706	0.15	8	149,281	Entertainment
wordpress.com	1615	0.14	859	33	Blogs
giphy.com	1055	0.09	365	1604	Photos
bbc.com	966	0.08	288	156	News
lastampa.it	927	0.08	45	2440	News
pinterest.com	892	0.08	170	32	Photos
softandapps.info	861	0.07	2	160,980	News
photobucket.com	768	0.07	348	341	Photos
nytimes.com	744	0.06	383	97	News
soundcloud.com	736	0.06	201	167	Audio
wikipedia.org	736	0.06	376	7	Encyclopedia
repubblica.it	682	0.06	49	439	News
theguardian.com	588	0.05	282	157	News
huffingtonpost.com	572	0.05	329	93	News
punto-informatico.it	570	0.05	29	42,955	News

ferent stories. We conclude from the graph that the most frequent domains are often used in the majority of stories.

Table 2 contains the top 25 domains of the resources ordered by their frequency. The list of top 25 domains represents 92.3 % of all resources. The table also contains the global rank of the domains according to Alexa[4] as of March 2015. We see from the table that Web elements from twitter.com appeared 943,859 times in 10,914 stories, comprising over 82 % of all Web elements. Note that plus.google.com has rank one, because Alexa does not differentiate plus.google.com from google.com. We manually categorized these domains in a more fine-grained manner than Storify provides with its "links, images, text, videos, quotes" descriptions.

Although the top 25 list of domains appearing in the stories is dominated by globally popular Web sites (e.g., Twitter, Instagram, YouTube, and Facebook), the long-tailed distribution results in the presence of many globally lesser known sites. In Sect. 4.4, we investigate the correlation between Alexa global rank and rank within Storify.

We also presented the list of top domains based on the count of stories in which they were used (Table 3). We notice that the two lists are similar. We also can see from Table 3 that storify.com appeared in the highly ranked domains across the stories, which means that many stories refer to other stories in Storify.

The embedded resources of twitter.com

Since Twitter is the most popular domain (>82 % of Web elements), we investigate if the tweets have embedded resources of their own. For example, Fig. 4 shows a tweet in a Storify story that contains an image from Twitter. Furthermore, other tweets may contain links or videos. This captures the behavior of users, including tweets in the stories, because the tweets are surrogates for embedded content. We randomly sampled 5 % of the Twitter resources (47,512 URIs). Of the sampled tweets in the stories, 32 % (15,217) have embedded resources, of which there are 14,616 unique URIs. Of 15,217,

[4] http://www.alexa.com/.

Table 3 Top 25 domains based on the number of stories that they appear in (Story Count)

Domain	Story count	Percentage of stories (%)	Frequency	Alexa global rank as of 2015–03	Category
twitter.com	10,914	74.92	943, 859	8	Social media
youtube.com	4238	29.09	22,076	3	Videos
instagram.com	1, 841	12.64	45,188	25	Photos
facebook.com	1802	12.37	13,930	2	Social media
flickr.com	1079	7.41	7317	126	Photos
wordpress.com	859	5.90	1615	33	Blogs
blogspot.com	713	4.89	1857	18	Blogs
tumblr.com	590	4.05	3066	31	Blogs
plus.google.com	537	3.69	3413	1	Social media
nytimes.com	383	2.63	744	97	News
wikipedia.org	376	2.58	736	7	Encyclopedia
giphy.com	365	2.51	1055	1604	Photos
photobucket.com	348	2.39	768	341	Photos
upload.wikimedia.org	345	2.37	564	200	Encyclopedia
huffingtonpost.com	329	2.26	572	93	News
cnn.com	303	2.08	480	76	News
bbc.com	288	1.98	966	156	News
theguardian.com	282	1.94	588	157	News
google.com	236	1.62	547	1	Search
patch.com	231	1.59	5783	2096	News
washingtonpost.com	225	1.54	432	218	News
imgur.com	215	1.48	1756	36	Photos
foxnews.com	210	1.44	271	215	News
storify.com	209	1.43	509	3237	Social network
forbes.com	207	1.42	304	164	News

The percentage of stories is out of 14,568

46 % are photos from twitter.com (hosted at twimg.com). Table 4 contains the ten most frequent domains for the embedded resources, which represent 61.6 % of all the URIs embedded in tweets. Again, we see that some Storify stories (0.49 %) point to other stories in Storify.

4.3 Classification of resources based on the TLD

Table 5 presents the distribution of the Top-Level Domains (TLDs) for the URIs that were used in Storify stories (only the top ten are shown). The table shows that the most used TLD is .com by far. Note that .cat is the TLD for a Catalan site (http://www.aragirona.cat/). The top ten list represents 98.92 % of all resources in Storify stories.

4.4 Correlation of global and storify popularity

We calculate Kendall's Tau correlation (τ_{sf}) between the top n domains in Storify stories based on their frequency (for example, the list of the top 25 domains in Table 2) and their Alexa global rank. We also checked Kendall's Tau correlation (τ_{sc}) between the top n domains used in the most number of stories (for example, the list of top 25 domains in Table 3) and their Alexa global rank.

The results are shown in Table 6. Statistically significant ($p \leq 0.05$) correlations are bolded. The highest correlation that we found between Alexa global rank and top domains based on frequency was 0.45 for the top 15 domains. The highest correlation between Alexa global rank and top domains based on the number of stories was 0.46 for the top 100 domains. From the results, we notice that most of the time, the highly ranked real-world resources, such as twitter.com, are correspondingly the most used in human-generated stories.

This is interestingly in contrast with Zhong et al. [37], which found that the most frequent sites on Pinterest had low Alexa global ranks. This is possibly due to the different natures of the usage of both sites. In Pinterest, users pin photos or videos of interest to create theme-based image/video collections, such as hobbies, fashion, and events. The most

Fig. 4 Tweet in Storify has an image as an embedded resource. Note that the text of the tweet includes the URI of the image

Table 4 Ten most frequent domains in the embedded resources of the tweets

Domain	Percentage	Category
twimg.com	46.17	Images
instagram.com	4.28	Images
youtube.com	2.82	Videos
linkis.com	2.04	Media sharing
facebook.com	1.40	Social media
wordpress.com	0.61	Blogs
vine.co	0.53	Videos
blogspot.com	0.52	Blogs
storify.com	0.49	Social network
bbc.com	0.44	News

used subject areas by Pinterest users are food and drinks, décor and design, and apparel and accessories [10]. Most of the pins on Pinterest come from blogs or are uploaded by users. In Storify, people tend to use social media and Web resources to create their narratives about events or news.

Table 6 Kendall's Tau correlation between the most frequent n domains in the stories and their Alexa global rank (τ_{sf}) and between the top n domains that have the most number of stories and Alexa global rank (τ_{sc})

n	10	15	25	50	100
τ_{sf}	0.1555	**0.4476**	**0.3372**	**0.3194**	**0.2485**
τ_{sc}	0.1556	0.3524	**0.4107**	**0.4260**	**0.4639**

4.5 What is the mean editing time for stories?

Table 7 shows the percentage of the stories with editing times in various time intervals. The table also shows the corresponding features of the stories, divided by their editing time. We normalized the number of views by the age of the story (dataset collection date − story creation date). The first two intervals (<1 h) represent stories that were created, modified, and then published with no continuing edits.

We see that the majority of the stories in the data set were created and edited in the span of one day. There are 14 % of stories that have been updated over a long period of time,

Table 5 Top ten TLDs of the resources

TLD	.com	.org	.it	.uk	.net	.de	.es	.info	.fr	.cat
Percentage	96.48	0.64	0.52	0.34	0.32	0.21	0.11	0.11	0.10	0.09

Table 7 Percentage of the stories based on the editing interval along with the median of Web elements, text elements, and views

Intervals	Percentage	Median Web elements	Median text elements	Median views
0–60 sec	14.0	15	0	23
1–60 min	26.7	19	0	53
1–24 h	23.4	25	5	110
1–7 days	13.5	26	7	78
1–4 weeks	8.4	26	9	80
1–12 months	10.9	38	2	129
1–4 years	3.1	56	15	156

The percentage is out of 15,568 stories

with the longest editing time in our data set covering more than four years and with more than 13,000 views. Curiously, this story had only 33 Web elements and 51 total elements. Although the story with the longest editing time did not have the largest number of elements, from Table 7, we can see that based on the median number of elements in each interval, there is a nearly linear relation between the editing time length of the story and the number of elements.

4.6 Decay of Web elements

In this section, we investigate how many resources in the stories are missing from the live Web and how many are available in public Web archives. We used Memento to check the existence in the archives. We checked the live Web and public Web archives for 265,181 URIs (202,452 URIs from the Web elements of stories + 47,512 randomly sampled tweet URIs + 15,217 URIs of embedded resources in those tweets), in which there are 253,978 unique URIs. Here, we further examine the results for the most frequent five domains in the stories: twitter.com, instagram.com, youtube.com, facebook.com, and flickr.com.

4.6.1 Existence on the live Web

We checked the existence of the 253,978 unique URIs on the live Web. We also checked the pages that give "soft 404s", which return HTTP 200, but do not actually exist [6]. The left two columns of Table 8 contain the results of checking the status of the Web pages on the live Web. Of all the unique URIs, 11.8 % are missing on the live Web. The table also contains the results of the five most frequent domains and all other URIs. We also included the results of checking the existence of Twitter embedded resources at the bottom of the table. From the table, we conclude that the decay rate of social media content is lower than the decay rate of the regular Web content and Web sites.

4.6.2 Existence on the live Web as a function of time

We measured the decay of the resources of Storify stories in time by measuring the percentage of the missing resources in the stories over time. For this experiment, we used the 249,964 (all the URIs excluding twitter embedded resources) resources in 14,513 stories to check the rate of the decay in the stories.

We found that 40.8 % of the stories contain missing resources with a mean value of 10.3 % of the elements missing per story. Figure 5 contains the distribution of the creation date of stories in our data set in each year and the percentage of the missing resources in each corresponding year. From the graph, we can infer a nearly linear decay rate of resources through time: the resources disappear at the rate of 30 % for the first year, 20 % for the second year, and then, the rate decreases steadily for the last three years until it reaches 9.6 % for the last year. This finding is close to the findings by SalahEldeen and Nelson [28], in which they found that there is a nearly linear relationship between time of sharing the resources and the percentage of resources lost from the live Web, with a rate of 11 % the first year and 7 % for each following year.

Table 8 Existence of the resources on the live Web (on the left) and in the archives (on the right)

Resources	Existence on live Web			Found in archives		
	Available (%)	Missing (%)	Total	Of the available (%)	Of the missing (%)	Total
Twitter	95.5	4.5	47,385	0.9	3.4	477
Instagram	86.6	13.4	43,396	0.3	0.07	103
Youtube	99.3	0.7	19,809	16.0	0.75	3140
Facebook	95.2	4.8	12,793	0.6	0.49	80
Flickr	95.6	4.4	6859	0.4	0.0	25
others	82.1	17.9	109,120	26.8	15.5	27,033
Twitter resources	90.1	9.9	14,616	8.0	14.1	1257

Available represents the requests which ultimately return HTTP 200, while missing represents the requests that return HTTP 4xx, HTTP 5xx, and HTTP 3xx to others except 200, timeouts, and soft 404s. Total is the total unique URIs from each domain

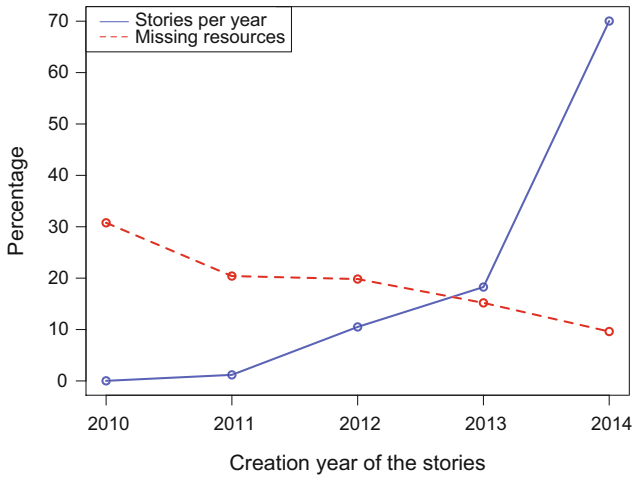


Fig. 5 Distribution of the stories per year and the decay rate of the resources in these stories through time

4.6.3 Existence in the archives

We checked the 253,978 resources for existence in general Web archives in March 2015. The existence in the Web archives was tested by querying a Memento Aggregator.[5]

The right-most columns of Table 8 contain the percentage of the URIs found in the Web archives out of the missing and the available URIs on the live Web. In total, 12.6 % of the URIs were found in the public Web archives. Of the missing resources (29,964), 11 % were found in public Web archives. In their study, Ainsworth et al. [1] estimated the coverage of Web resources in Web archives using 4000 URIs from the Open Directory Project (DMOZ), Delicious, Bitly, and search engines and measured their coverage in the public Web archives and the number and frequency of archived versions. They found that 35–90 % of the Web has at least one archived copy in the Web archives. Moreover, they found that DMOZ and Delicious samples had existence in the archives more than the samples of the Bitly and search engine. From Table 8, we notice that social media are not as well-archived as the regular Web. Facebook uses robots.txt to block Web archiving by the Internet Archive,[6] but the other sites do not have this restriction.

5 What does a popular story look like?

In this section, we establish structural features for what differentiates popular stories from normal stories for building a baseline for the stories that we will automatically create from the archives. We divided the stories into popular and unpopular stories based on their number of views, normalized by the amount of time that they were available on the Web. We consider as popular the top 25 % of stories (3642 stories) based on the number of views (over 377 views/year).

5.1 Features of the stories

We considered the distributions of several features of the stories: number of Web elements, the number of text elements, and the editing time. We also check if there is a relationship between the popular stories and the relative number of subscribers. Furthermore, we test if popular stories are different from the unpopular stories using the Kruskal–Wallis test [16], which allows comparing two or more samples that are independent and have different sample sizes.

We found that at the $p \leq 0.05$ significance level, the popular and the unpopular stories are different in terms of the following features: number of Web elements, text elements, editing time, and subscribers. Figure 6a shows that popular stories tend to have more Web elements (medians of 28 vs. 21) and a longer editing time (5 vs. 2 h) than the unpopular stories. The number of elements in the popular stories is between 2 and 1950 Web elements with $median = 28$ and $mode = 10$, and the number of text elements ranges from 0 to 559 with $median = 1$ and $mode = 0$. The popular stories tend to have longer editing time intervals than the unpopular stories. For the popular stories, 38 % have an editing time of at least one day, while only 35 % of the unpopular stories have this feature. The maximum editing time in the popular stories is 4.1 years, while it is 3.5 years for unpopular stories.

There is a large difference between the number of subscribers for the authors of popular stories than for those of unpopular stories. The authors of popular stories have min/median/max values of 0/16/1,726,143 subscribers, while the authors of unpopular stories have 0/2/2469 subscribers.

5.2 The type of elements

Figure 6b shows the distributions for the popular and unpopular stories for each element type. The popular stories tend to have more images than the unpopular stories. The median number of images in popular stories is 10, while it is 5 in the unpopular stories. For videos, the median is 2 for both popular and unpopular. We found that the median number of the links in popular stories (20 links) is higher than the unpopular stories (16 links). We also test if the types of elements used in popular stories are different from the unpopular sto-

[5] http://timetravel.mementoweb.org/guide/api/, which provided results from 12 different public web archives.

[6] See: https://archive.org/about/faqs.php#14.

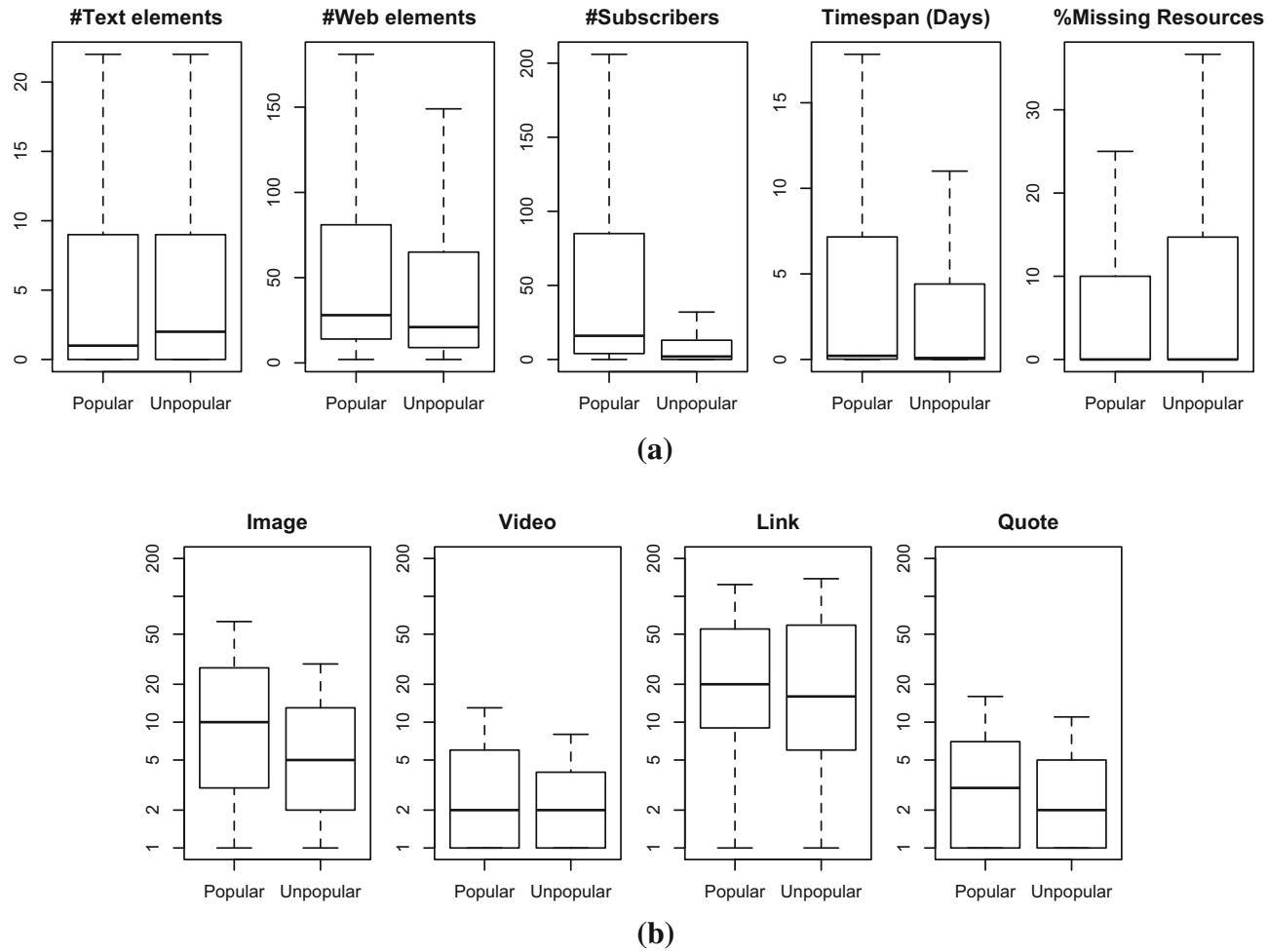


Fig. 6 Characteristics of popular and unpopular stories. **a** Distributions of the number of features per story. **b** Distributions of the number of elements per story

ries using the Kruskal–Wallis test and found that $p \leq 0.05$ for the distributions of each of the elements (images, videos, links, and quotes).

5.3 Do popular stories have a lower decay rate?

We checked the decay rate of the popular and the unpopular stories to investigate if there is a relationship between popularity and lower decay rate. We found that for the popular stories, 11.0 % of the resources were missing, while 12.8 % of the resources were missing for unpopular stories. Figure 6 contains the distribution of the percentage of missing resources per story in popular and unpopular stories. It shows that the resources of the popular stories have lower decay than the resources of the unpopular. A reason could be that the popular stories are edited, and edits could be fixing broken links. The 75th percentile of the decay rate per popular story is 10 % of the resources, while it is 15 % in the unpopular stories.

6 Characteristics of archived collections

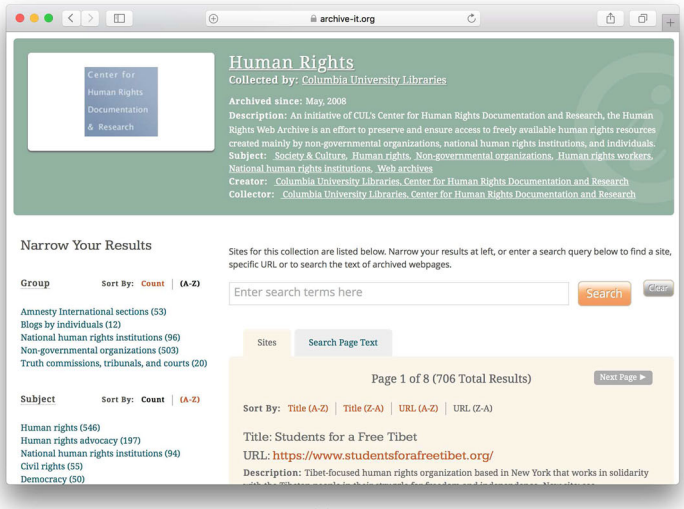
As the Web has become an integral part of our lives, shaping how we get news, shop, and communicate, preserving the resources of the Web is essential to facilitate research in history, sociology, political science, media, literature, and other related disciplines [24].

Web archives are institutions that preserve much of the cultural discourse by archiving the Web [19], such as the Internet Archive[7] (IA) and the UK Web Archive.[8] The Internet Archive is the most popular Web archive and it aims to maintain an archive of the entire Web by taking periodic snapshots of pages then providing an access to these snapshots via the Wayback Machine [23,34].

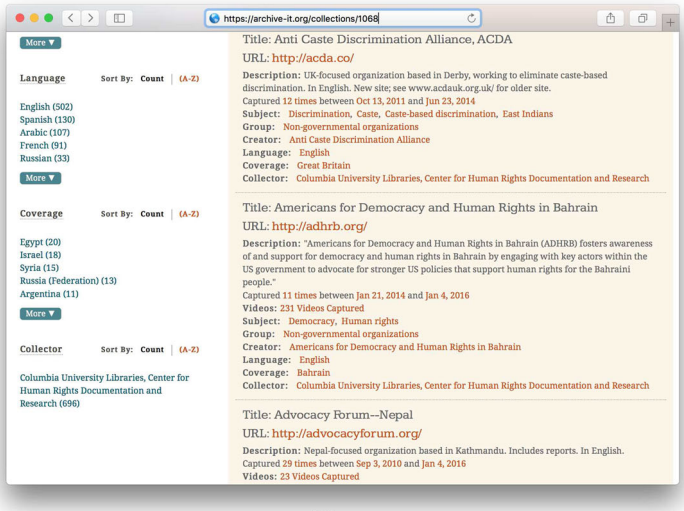
Archive-It is a collection development service operated by the Internet Archive since 2006. As of November 2015,

[7] http://archive.org/.

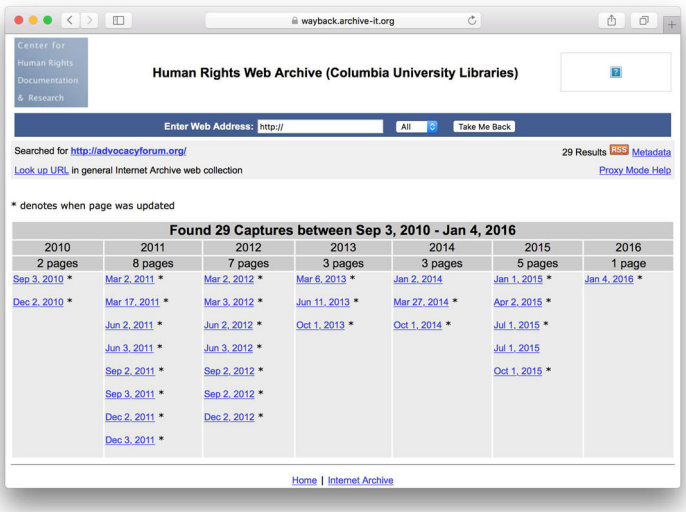
[8] http://www.webarchive.org.uk/.


archive-it.org
Human Rights
Collected by: Columbia University Libraries
Archived since: May, 2008
Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.
Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives
Creator: Columbia University Libraries, Center for Human Rights Documentation and Research
Collector: Columbia University Libraries, Center for Human Rights Documentation and Research
Narrow Your Results
Group
Amnesty International sections (53)
Blogs by individuals (12)
National human rights institutions (96)
Non-governmental organizations (503)
Truth commissions, tribunals, and courts (20)
Subject
Human rights (546)
Human rights advocacy (197)
National human rights institutions (94)
Civil rights (55)
Democracy (50)
Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.
Enter search terms here
Search
Sites
Search Page Text
Page 1 of 8 (706 Total Results)
Next Page
Title: Students for a Free Tibet
URL: https://www.studentsforafreetibet.org/

(a)


https://archive-it.org/collections/1068
Language
English (502)
Spanish (130)
Arabic (107)
French (91)
Russian (33)
Coverage
Egypt (20)
Israel (18)
Syria (15)
Russia (Federation) (13)
Argentina (11)
Collector
Columbia University Libraries, Center for Human Rights Documentation and Research (696)
Title: Anti Caste Discrimination Alliance, ACDA
URL: http://acda.co/
Title: Americans for Democracy and Human Rights in Bahrain
URL: http://adhrb.org/
Title: Advocacy Forum--Nepal
URL: http://advocacyforum.org/

(b)


wayback.archive-it.org
Human Rights Web Archive (Columbia University Libraries)
Enter Web Address: http://
Take Me Back
Searched for http://advocacyforum.org/
29 Results
Look up URL in general Internet Archive web collection
Proxy Mode Help
* denotes when page was updated
Found 29 Captures between Sep 3, 2010 - Jan 4, 2016
2010 2 pages
2011 8 pages
2012 7 pages
2013 3 pages
2014 3 pages
2015 5 pages
2016 1 page
Home | Internet Archive

(c)

Table 9 Distribution of features of Archive-It collections

Features	Seed URIs	Mementos	Timespan
25th percentile	1	1	0
50th percentile	5	3	154
75th percentile	21	9	973
90th percentile	73	26	1791
Maximum	123,600	3848	6945
Mean	98	17	628
SD	2260	106	921

Timespan is measured in days

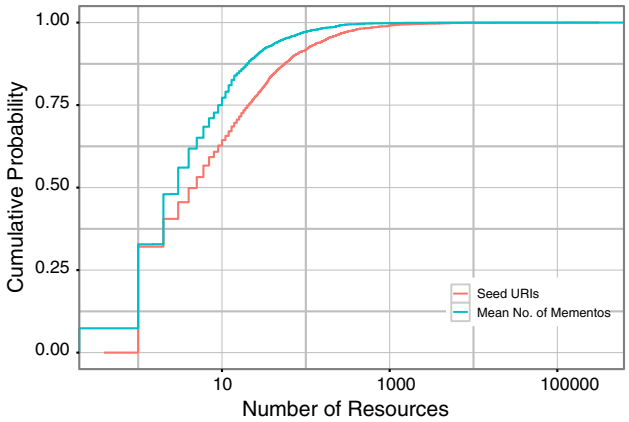


Fig. 8 Distribution of the number of seed URIs and the mean number of mementos per seed in Archive-It collections

Archive-It was used by over 340 institutions in 48 states, and featured over 9B archived Web pages in nearly 3200 separate collections.

Archive-It allows users to create, maintain, and view digital collections of Web resources hosted at Archive-It. This is done by the user manually specifying a set of seeds 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 [22] crawler at Archive-It crawls/recrawls that these seeds based on the predefined frequency and depth to build a collection of archived Web pages that the curator believes best exemplifies the topic of the collection. In a previous work [5], we presented multiple approaches for assisting the curator in identifying off-topic mementos in archived collections based on a data set from Archive-It as a first step for generating stories from archived collections. Archive-It provides a listing of all URIs in the collection along with the number of times and dates over which each site was archived, as well as a full-text search of archived sites. Figure 7 shows the current interfaces for a typical collection.[9] Figure 7a shows brief metadata about the collection. Figure 7b shows the Archive-It interface, which consists mainly of a list of seed URIs in alphabetical order in which the crawl information of each URI is available. Figure 7c shows a list of the times when a single URI was archived, which is called a TimeMap.

6.1 Archive-It collections

As of November 2015, we obtained the IDs of the whole population of Archive-It collections from the front-end interface of Archive-It. We excluded the collections that we knew were created automatically (the seed URIs have been extracted automatically from the Web), and also collections with no data. We kept collections with one URI, because they have mementos. The number of remaining collections is 3109, comprising 305,522 seed URIs. The total number of mementos for all the collections is 2,385,397. We downloaded the metadata of all seed URIs in November 2015. For each seed URI, we obtained its first crawling date, last crawling date, and number of mementos.

6.2 General characteristics

Table 9 shows the characteristics of Archive-It collections in terms of the number of seed URIs, the mean number of the mementos per seed, and timespan, which is the range of time period over which the Web pages have been archived. The mean number of seed URIs in Archive-It collections is 98 URIs, and the median is 5 URIs. The mean number of mementos is 17 mementos per seed URI, and the mean timespan is 21 months. Figure 8 contains the distribution of the number of seed URIs and the mean number of mementos per seed in each collection.

The largest collection in terms of the number of seed URIs is the "Government of Canada Publications"[10] collection that archives Canadian governmental pages, created by the Canadian Government Information PLN Web Archive.[11] It contains 123,647 URIs with a span of 2 years (2013–2015) and a mean of 2 mementos for each URI. The largest timespan in the collections is 19 years (from 1996 until 2015) for only 21 seed URIs. The start date of crawling for multiple collections is before the existence of Archive-It in 2006. This is possible, because some organizations imported previously archived pages to initialize their collections.

6.3 What domains are used in collections?

Canonicalizing the domains of 305,522 URIs resulted in 57,640 unique domains in the 3109 collections. Figure 9 shows the relationship between the frequency of domains

[9] https://archive-it.org/collections/1068/.

[10] https://archive-it.org/collections/3572/.

[11] https://archive-it.org/organizations/700/.

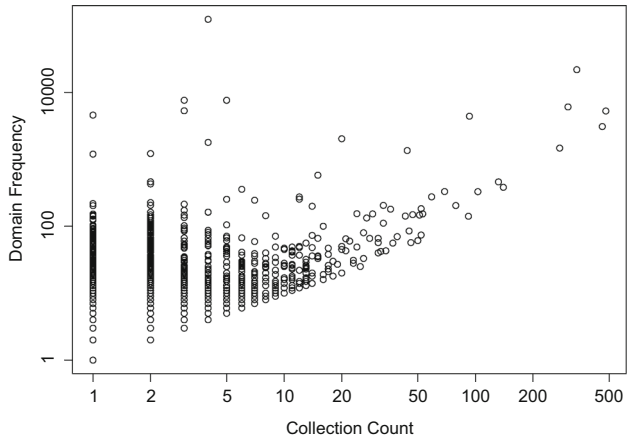


Fig. 9 Relationship between the frequency of the domains in Archive-It collections and the number of collections in which those domains appear

and the number of collections that they appeared in. For example, the dot at the top of the graph represents the most frequent domain, which appears over 100,000 times in only four different collections. We notice that multiple domains in Archive-It collections have a high frequency, but appear in only a few collections. This is because some collections are devoted to archiving specific domains.

Table 10 contains the top 25 domains of the resources ordered by their frequency. The list of top 25 domains represents 66.1 % of all the resources. The table also contains the global rank of the domains according to Alexa as of March 2015. We also added our manual categorization for the domains. We notice that the most used domain is publications.gc.ca from the "Government of Canada Publications" collection, which contains the largest number of URI-Rs. We added the collection counts to the table to reflect the global rank of the domains across the collections. We notice that the first ranked domain based on the frequency of the domains appeared in only four collections. The table also shows that most of domains in the top list are for government and education Web sites. There are also blogs and social media Web sites, such as facebook.com and twitter.com. Table 10 also shows that some collections use archived URIs in their seed list. The domain wayback.archive-it.org

Table 10 Top 25 domains based on the frequency of appearance in Archive-It

Domain	Frequency	Percentage	Collection count	Alexa global rank as of 2015–11	Category
publications.gc.ca	123,604	40.46	4	192,814	Government
youtube.com	21,838	7.15	337	3	Videos
mtholyoke.edu	7632	2.50	3	34,718	Education
nsa.gov	7625	2.50	5	49,313	Government
blogspot.com	6072	1.99	305	38	Blogs
nsf.gov	5312	1.74	3	15,613	Government
facebook.com	5268	1.72	480	2	Social media
hem.bredband.net	4582	1.50	1	367,103	Company
wikipedia.org	4405	1.44	93	7	Encyclopedia
twitter.com	3089	1.01	460	9	Social media
nlm.nih.gov	2030	0.66	20	196	Government
wayback.archive-it.org	1791	0.59	4	133,005	Archive
wordpress.com	1471	0.48	276	36	Blogs
vimeo.com	1354	0.44	44	186	Blogs
uwrf.edu	1218	0.40	2	157,000	Education
pubs.pembina.org	1196	0.39	1	709,328	Education
hhs.gov	579	0.19	15	8641	Government
globe.gov	462	0.15	2	559,353	Government
flickr.com	460	0.15	132	159	Education
netfiles.uiuc.edu	429	0.14	2	17,442	Education
orgsync.com	356	0.12	6	12,450	Company
nytimes.com	330	0.11	69	97	News
tumblr.com	328	0.11	102	43	Blogs
baylor.edu	274	0.09	12	17,643	Education
rochester.edu	254	0.08	12	9093	Education

The percentage is the frequency of the domain out of 305,522

Table 11 Top 25 domains based on the number of Archive-It collections that they appear in

Domain	Collection count	Percentage	Frequency	Alexa global rank as of 2015–11	Category
facebook.com	480	15.44	5268	2	Social media
twitter.com	460	14.80	3089	9	Social media
youtube.com	337	10.84	21,838	3	Videos
blogspot.com	305	9.81	6072	38	Blogs
wordpress.com	276	8.88	1471	36	Blogs
flickr.com	132	4.25	460	159	Photos
tumblr.com	102	3.28	328	43	Blogs
wikipedia.org	93	2.99	4405	7	Encyclopedia
ok.gov	92	2.96	141	24,315	Government
instagram.com	78	2.51	203	24	Photos
nytimes.com	69	2.22	330	97	News
sites.google.com	69	2.22	194	1	Wikipedia
tn.gov	53	1.70	153	13,494	Government
bbc.com	52	1.67	183	100	News
slco.org	52	1.67	74	100,152	Government
cnn.com	51	1.64	147	75	News
sfgov.org	50	1.61	61	101,777	Government
huffingtonpost.com	47	1.51	149	122	News
tennessee.gov	46	1.48	57	175,859	Government
yahoo.com	45	1.45	85	5	Search
vimeo.com	44	1.42	1354	186	Videos
weebly.com	43	1.38	142	252	Company
typepad.com	39	1.25	70	1126	Blogs
washingtonpost.com	36	1.16	180	198	News
pinterest.com	36	1.16	55	30	Photos

The percentage is the number of collections the domain appeared in out of 3109

is ranked 12th based on its frequency and appeared in four collections.

Table 11 shows the top 25 domains based on the number of collections that they appeared in. It is clear from the table that the top list of domains based on the number of collections that they appeared in is different from the top domains based on the frequency. Note that sites.google.com has rank one, because Alexa does not differentiate sites.google.com from google.com. In Sect. 6.5, we investigate the correlation between the rank of the domains within Archive-It collections and their Alexa global rank.

6.4 Classification of the seed URIs based on the TLD

Table 12 presents the distribution of TLDs for the seed URIs in Archive-It collections (only the top ten are shown). The top ten list represents 97.8 % of the TLDs in the collections. It can be noticed that most of the URIs are for the .ca, .com, .edu, .org, .gov, .net, .us, .uk, and .de domains. The .ca comes from the publications.gc.ca, which dominates the top 25 most frequent domains. We notice that there are many governmental, organizational, and educational sites in the collections.

6.5 Correlation of global and Archive-It popularity

Table 13 shows Kendall's Tau correlation τ_{af} for the most frequent n domains in Archive-It collections and their Alexa global rank. It also shows Kendall's Tau correlation τ_{ac} for the top n domains based on the largest number of collections and their Alexa global rank. Statistically significant ($p \leq 0.05$) correlations are bolded. The table shows that the correlation between the most frequent n domains and their Alexa global rank is very low. The highest correlation between the most frequent n domains and the Alexa global rank is 0.17 for the

Table 12 Top ten TLDs of the resources

TLD	.ca	.com	.edu	.org	.gov	.net	.us	.uk	.de	.fr
Percentage	41.96	23.73	9.77	8.50	8.21	2.24	0.70	0.61	0.38	0.31

Table 13 Kendall's Tau between the most frequent n domains in the stories and their Alexa global rank (τ_{af}) and between the top n domains that have the most number of collections and Alexa global rank (τ_{ac})

n	10	15	25	50	100
τ_{af}	−0.2000	0.0286	−0.0467	0.0008	**0.1741**
τ_{ac}	0.4222	**0.4857**	**0.4174**	**0.4399**	**0.3180**

list of the 100 domains. This may be due to the nature of the collections and the purpose for which they are created. Most of the collections are explicitly centered around topics. Furthermore, some collections archive specific domains (e.g., publications.gc.ca). Many of these domains are not high ranked globally, but the collections that they appeared in have a large number of seed URIs, which results in high frequency for these domains.

Although the frequency of domains does not correlate with the globally high-ranked domains, the top list of the domains based on the number of collections that they appeared in highly correlates with the global rank of these domains. For most of the top n domains across Archive-It collections, $\tau_{ac} > 0.4$. The highest correlation is 0.49 for the list of 15 domains.

6.6 What is the mean timespan for digital collections?

Table 14 shows the percentage of the collections that have been crawled in each time interval. The table also shows the corresponding features of the collection in terms of the number of seed URIs and the mean number of mementos per seed. Note that the timespan of the collection is different from the editing time of Storify stories.

The first row contains collections with 0 mementos as of November 2015. About 20 % of these collections have been created recently, and their crawling date started after we captured the metadata of the collections. Among these collections, the collection with the largest number of URIs in this category ("Cal Poly University Web Archive"[12]) has 412 seed URIs.

We see that the majority of the collections have a long timespan, meaning that they have been crawled over the span of years. There are 17 % of the collections with a span of more than 4 years. The collection with the longest timespan of 19 years has URIs that were crawled before Archive-It existed.

From Table 14, we notice that there is a linear relationship between the mean number of mementos per seed in the collection and the timespan of the collection. The mean number of mementos per seed URIs increases with an increase in the timespan of the collection. The mean number of mementos in the span of 4 years (or more) is 60 mementos per seed, and goes down 70 % to be 17 mementos per seed in the span of 1–4 years.

[12] https://archive-it.org/collections/6191/.

6.7 The decay rate in Archive-It collections

We extracted 293,883 unique seed URIs from Archive-It collections and checked their existence on the live Web. We found that 8.3 % (24,521 out of 293,883) of the seed URIs in Archive-It collections are missing from the live Web. Missing represents the requests that return HTTP 4xx, HTTP 5xx, and HTTP 3xx to others except 200, timeouts, and soft 404s. Note that 42 % of the seed URIs belong to the "Government of Canada Publications" collection, which is devoted to archiving governmental publication documents that are well preserved by the Canadian government. We measured the loss for this collection and found that only 0.1 % (102 URIs out of 122,948 unique URIs) of the documents are missing. For these kinds of collections, we expect that if the domain is lost or unavailable for any reason [6,20], all the 122,948 URIs might disappear. Excluding the "Government of Canada Publications" collection, the decay rate for the rest of the collections is 14.3 % (24,419 out of 170,935 unique URIs).

We also found that 58.7 % (1825 out of 3109) of the collections contain seed URIs that had disappeared from the live Web. Of these, 22.5 % (410 out of 1825) have 100 % loss of their seed URIs from the live Web.

7 Storify stories versus Archive-It collections

In this section, we contrast the general characteristics of human-generated stories from Storify and human-curated collections from Archive-It. For example, the most frequent domain in Storify (twitter.com), which is represented in Fig. 3 by the right-most dot, appeared almost 1,000,000 times in the largest number of stories (over 10,000 stories). On the other hand, the most frequent domain in Archive-It collections (publications.gc.ca), which is represented by the dot on the top left of Fig. 9, appeared over 100,000 times in only four collections. The difference in the nature of the domains could be due to the difference of who is creating the collection: regular users (Storify), or librarians employed by government, museums, etc. (Archive-It).

In addition, the most frequent domains in the stories have a higher correlation with the Alexa global rank than the most frequent domains in the archived collection, as shown in Tables 6 and 13. For most of the n values in Table 6, there is a high correlation between the most frequent n domains in the stories and their Alexa global Rank (τ_{sf}). The τ_{sf} at $n = 15$ is 0.45, while in Archive-It collections, the list of the most frequent 15 domains and their Alexa global rank (τ_{af})

Table 14 The distributions of the number of collections in each time interval

Intervals	*Percentage*	Seed URIs			Mean no. of mementos/seed		
		Mean	Median	Maximum	Mean	Median	Maximum
URI with no captures	6.80	7	1	412	0	0	0
<1 day	21.00	24	1	7619	1.1	1	4.8
1–7 days	4.90	101	5	7590	2.6	2	10.1
1–4 weeks	4.60	28	12	495	3.7	2.7	29.8
1–12 months	19.90	66	10	5309	10.9	3.4	277.6
1–4 years	25.40	242	6	123,648	16.6	5	594.5
>4 years	17.30	69	10	2365	59.5	13.7	3848

are statistically independent (Table 13). The largest value of τ_{af} is 0.17 at $n = 100$.

In addition, Tables 5 and 12 show that the list of TLDs in Storify is dominated by .com, which represents 96.5 % of the resources, while it represents only 23% in Archive-It collections. The list of TLDs in Archive-It collections contains a significant existence for .gov and .edu domains. That is because many collections are devoted to archiving governmental pages (e.g., all Web pages published by the state of California) and memory organizations, such as libraries and museums, but many of the collections are explicitly centered around topics in arts and humanities, politics, spontaneous events, and blogs and social media.

For the decay rate, 11.8 % of Storify resources do not exist on the live Web, while 8.3 % of Archive-It URIs are missing. Although the decay rate in Storify stories is larger than the decay rate of Archive-It collections, the percentage of the affected collections (58.7 %) is larger than the percentage of the affected stories (40.8 %). Furthermore, the mean value of the missing elements per story is 10.3 %, although the mean value of the missing seed URIs per collection is 42 %.

To conclude, the resources that are used in Storify stories are different from the resources in Archive-It collections. In summarizing a collection, we can only choose from what is archived. Therefore, if there are no tweets in the collection, twitter.com will not be the most common domain in the generated stories. Although some content in Storify stories will not be applicable (e.g., twitter.com is popular in Storify, but mostly missing in Archive-It collections), some other characteristics will be applicable, such as the number of resources. Accordingly, our choices of what to select from the collection needs to be informed by what constitutes a "popular" story.

8 Future work and conclusions

In this paper, we presented the structural characteristics of human-generated stories on Storify, with particular emphasis on "popular" stories (i.e., the top 25 % of views, normalized by time available on the Web). Upon analyzing 14,568 stories, the popular stories have a median value of 28 elements, while the unpopular stories have 21. The median value of multimedia elements in popular stories is 12, with only 7 in unpopular stories. Of the popular stories, 38 % receive continuing edits (as opposed to 35 %), and only 11% of Web elements are missing on the live Web (as opposed to 13 %). We found that there is nearly a linear relationship between the timespan of the story and the number of Web elements. There were 11.8 % of the resources missing from the live Web, in which 11 % were found in the archives. The percentage of the missing resources is proportional with the age of the stories.

Studying human-generated stories in Storify helped us to profile different kinds of stories by examining the typical length (in terms of the number of resources included), time frames covered, structural metadata (e.g., PageRank, images and video, social media vs. news), and other features. We model the structural characteristics of these stories, with particular emphasis on "popular" stories. For example, we generate stories (semi-)automatically from archived collections with a typical length close to 28 (more or less based on the collection size). Based on the analysis in this paper of the structural characteristics of "good" stories, we have been able to automatically construct summarizing stories of Archive-It collections that are indistinguishable from those created by human subject domain experts, while at the same time, both kinds of stories (automatic and human) are easily distinguished from randomly generated stories [2].

We checked the characteristics of archived collections using data set from Archive-It for specifying what can be applied in our intended framework of generating stories from these collections. We found that some characteristics of human-generated stories may not be possible to apply, because the nature of the resources in the stories is different from what compose the collections. For example, we found that twitter.com is popular in Storify, but mostly is missing in Archive-It. Some other structural characteristics of human-generated stories, such as the number of elements and the distribution of domains, will provide us with a

template with which to evaluate our automatically generated stories. Future work also will include investigating if the structural characteristics of stories hold for other social media storytelling services, such as Paper.li, Scoop.it!, and Pinterest.

Acknowledgments This work was supported in part by IMLS LG-71-15-0077-15. We thank Kristine Hanna and Jefferson Bailey of the Internet Archive for the Archive-It data and baseline story summaries.

References

1. Ainsworth, S.G., AlSum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proceedings of the 11th ACM/IEEE-CS joint conference on digital libraries, JCDL '11, pp. 133–136. ACM Press, New York (2011). doi:10.1145/1998076.1998100
2. AlNoamany, Y.: Using web archives to enrich the live web experience through storytelling. Ph.D. thesis, Old Dominion University (2016)
3. AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L.: Who and what links to the internet archive. Int. J. Digit. Libr. **14**(3–4), 101–115 (2014). doi:10.1007/s00799-014-0111-5
4. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Characteristics of social media stories. In: Proceedings of the 19th International conference on theory and practice of digital libraries, TPDL '15, pp. 267–279. Springer International Publishing, Cham (2015). doi:10.1007/978-3-319-24592-8_20
5. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Detecting off-topic pages in web archives. In: Proceedings of the 19th international conference on theory and practice of digital libraries,TPDL'15, vol. 9316, pp. 225–237. Springer International Publishing (2015). doi:10.1007/978-3-319-24592-8_17
6. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic transit gloria telae: towards an understanding of the web's decay. In: Proceedings of the 13th international conference on World Wide Web, WWW '04, pp. 328–337 (2004). doi:10.1145/988672.988716
7. Brewington, B., Cybenko, G.: Keeping up with the changing web. Computer **33**(5), 52–58 (2000). doi:10.1109/2.841784
8. Cohen, J., Mihailidis, P.: Storify and news curation: teaching and learning about digital storytelling. In: Second annual social media technology conference & workshop, vol. 1, pp. 27–31 (2012)
9. Duh, K., Hirao, T., Kimura, A., Ishiguro, K., Iwata, T., Yeung, C.M.A.: Creating stories: social curation of twitter messages. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, ICWSM' 12 (2012)
10. Hall, C., Zarro, M.: Social curation on the website pinterest.com. Am. Soc. Inf. Sci. Technol. **49**(1), 1–9 (2012)
11. Han, J., Choi, D., Choi, A.Y., Choi, J., Chung, T., Kwon, T.T., Rha, J.Y., Chuah, C.N.: Sharing topics in pinterest: understanding content creation and diffusion behaviors. In: Proceedings of the 2015 ACM on conference on online social networks, COSN '15, pp. 245–255. ACM, New York (2015). doi:10.1145/2817946.2817961
12. Kieu, B.T., Ichise, R., Pham, S.B.: Predicting the popularity of social curation. In: Knowledge and systems engineering, pp. 413–424. Springer, Cham (2015)
13. Klein, M., Nelson, M.L.: Find, new, copy, web, page—tagging for the (re-)discovery of web pages. In: Proceedings of the 15th international conference on theory and practice of digital libraries, TPDL'11, pp. 27–39. Springer, Berlin, Heidelberg (2011). doi:10.1007/978-3-642-24469-8_5
14. Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., Tobin, R.: Scholarly context not found: one in five articles suffers from reference rot. PloS One **9**(12), e115,253 (2014). doi:10.1371/journal.pone.0115253
15. Koehler, W.: Web page change and persistence-a four-year longitudinal study. J. Am. Soc. Inf. Sci. Technol. **53**(2), 162–171 (2002)
16. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**(260), 583–621 (1952). doi:10.1080/01621459.1952.10483441
17. Laire, D., Casteleyn, J., Mottart, A.: Social media's learning outcomes within writing instruction in the EFL classroom: exploring, implementing and analyzing storify. Proc. Soc. Behav. Sci. **69**, 442–448 (2012)
18. Lawrence, S., Pennock, D.M., Flake, G.W., Krovetz, R., Coetzee, F.M., Glover, E., Nielsen, F.A., Kruger, A., Giles, C.L.: Persistence of web references in scientific research. Computer **34**(2), 26–31 (2001). doi:10.1109/2.901164
19. Lyman, P.: Archiving the world wide web. building a national strategy for digital preservation: issues in digital media archiving, pp. 38–51 (2002)
20. Marshall, C., McCown, F., Nelson, M.: Evaluating personal archiving strategies for internet-based information. Proc. Archiv. **2007**(1), 151–156 (2007)
21. Mihailidis, P., Cohen, J.N.: Exploring curation as a core competency in digital and media literacy education. J. Interact. Media Educ. **2013**, 1–19 (2013). doi:10.5334/2013-02
22. Mohr, G., Stack, M., Ranitovic, I., Avery, D., Kimpton, M.: An introduction to heritrix an open source archival quality web crawler. In: Proceedings of the 4th international web archiving workshop, IWAW '04, pp. 43–49 (2004)
23. Negulescu, K.C.: Web archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIIPP072110FinalIA.ppt (2010)
24. Nelson, M.L.: A plan for curating "Obsolete Data or Resources". Tech. Rep. arXiv:1209.2664 (2012)
25. Ottoni, R., Las Casas, D., Pesce, J.P., Meira Jr, W., Wilson, C., Mislove, A., Almeida, V.: Of pins and tweets: investigating how users behave across image-and text-based social networks. In: Proceedings of the 8th international AAAI conference on weblogs and social media, ICWSM' 14, pp. 386–395 (2014)
26. Padia, K., AlNoamany, Y., Weigle, M.C.: Visualizing digital collections at archive-It. In: Proceedings of the 12th annual international ACM/IEEE joint conference on digital libraries, JCDL '12, pp. 15–18 (2012). doi:10.1145/2232817.2232821
27. Palomo, B., Palomo, B.: New information narratives: the case of storify. Hipertext.net 12 (2014). doi:10.2436/20.8050.01.6
28. SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: How many resources shared on social media have been lost? In: Proceedings of the 16th international conference on theory and practice of digital libraries, TPDL'12, pp. 125–137. Springer-Verlag, Cham (2012). doi:10.1007/978-3-642-33290-6_14
29. SalahEldeen, H.M., Nelson, M.L.: Carbon dating the web: estimating the age of web resources. In: Proceedings of 3rd temporal web analytics workshop, TempWeb '13, pp. 1075–1082 (2013)
30. Sastry, N.: Predicting pinterest: organising the world's images with human-machine collaboration. In: Proceedings of the 24th international conference on world wide web, WWW '15 Companion, pp. 1065–1065. International World Wide Web Conferences Steering Committee (2015). doi:10.1145/2740908.2744719
31. Seitzinger, J.: Curate me! exploring online identity through social curation in networked learning. In: Proceedings of the 9th international conference on networked learning, pp. 7–9 (2014)
32. Stanoevska-Slabeva, K., Sacco, V., Giardina, M.: Content curation: a new form of gatewatching for social media? In: Proceedings of the

12th international symposium on online journalism (2012). http://online.journalism.utexas.edu/2012/papers/Katarina.pdf
33. Taylor, M.: Introduction to javascript object notation: a to-the-point guide to JSON. CreateSpace Independent Publishing Platform, USA (2014)
34. Tofel, B.: Wayback for accessing web archives. In: Proceedings of international web archiving workshop. IWAW (2007). http://iwaw.europarchive.org/07/IWAW2007_tofel.pdf
35. Van de Sompel, H., Nelson, M.L., Sanderson, R.: RFC 7089—HTTP framework for time-based access to resource states–Memento (2013). http://tools.ietf.org/html/rfc7089
36. Zhong, C., Salehi, M., Shah, S., Cobzarenco, M., Sastry, N., Cha, M.: Social bootstrapping: how pinterest and last.fm social communities benefit by borrowing links from facebook. In: Proceedings of the 23rd international conference on World Wide Web, WWW '14, pp. 305–314. ACM, New York (2014). doi:10.1145/2566486.2568031
37. Zhong, C., Shah, S., Sundaravadivelan, K., Sastry, N.: Sharing the loves: understanding the how and why of online content curation. In: Proceedings of the 7th international AAAI conference on weblogs and social media, ICWSM' 13, pp. 659–667 (2013)