Characteristics of Social Media Stories

Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA 23529, USA
{yasmin,mweigle,mln}@cs.odu.edu

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 like Storify to present 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 the elements are missing from the live web.

Keywords: Stories, Storify, Storytelling, Social Media, Curation

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, clustered together with a single URI and suitable for sharing. It provides a graphical interface for selecting URIs of web resources and arranging the resulting snippets and previews (see Figure 1), with a special emphasis on social media (e.g., Twitter, Facebook, Youtube, Instagram). We call these previews of web resources "web elements", and the annotations Storify allows on these previews we call "text elements".

We would like to use Storify to present automaticaly created summaries of collections of archived web pages in a social media interface that is more familiar users (as opposed to custom interfaces for summaries, e.g. [11]). 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).

Fig. 1. Example of creating a story on Storify.

In this paper, we will build a baseline for what human-generated stories look like and specify the characteristics of the popular stories. We answer 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 timespan (editing time) of the stories? Is there a relation between the timespan and the features of the story? Is there a relation between the popularity of the stories and the number of elements? Is there a relation 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 value of 2/28/1950 elements, with the unpopular stories having 2/21/2216. 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 value 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 relation between the timespan of the story and the number of web elements. We also found that only 11% of the missing resources could be found in public web archives.

2 Related Work

There have been many studies on how the social media is being used in social curation [12, 4, 20, 10, 19]. 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 [15].

Duh et al. [4] 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 through defining the topics being curated. They found that there are a diverse number of topics and a variety of social purposes for the content curation, such as summarizing an event and discussing TV shows.

Zhong et al. [20] studied why and how people curate using data sets of three in January 2013 for Pinterest, and over the month of December 2012 for Last.fm. They found that curation tends to focus on items that may not be high ranked in popularity and search rankings, which slightly contradicts our finding in Section 4.3. They also found that curation tends to be a personal activity more than being social.

Storify has been used in many studies by journalists [16] and also to explore how curation works in the classroom [9, 8]. Cohen et al. believe that Storify can be used to encourage students to become empowered storytellers and researchers [3]. Laire et al. [8] used Storify to study the effect of social media on teaching practices and writing activities.

Stanoevska-Slabeva et al. [16] 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. [6] proposed a method for predicting the popularity of social curation content based on a data set from Storify. 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 the curator features perform well for detecting the popularity. In this paper, we also investigate if there is relation between the number of views and the features of the story, such as the number of elements.

3 Constructing the Data Set

We created the data set by querying the Storify Search API[1] with the most frequent 1000 English keywords issued to Yahoo[2]. We retrieved 400 results for each keyword, resulting in a total of 145,682 stories downloaded in JavaScript

[1] `http://dev.storify.com/api`
[2] `http://webscope.sandbox.yahoo.com/catalog.php?datatype=l`

Object Notation (JSON). 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 only zero or one elements or zero views, resulting in only 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

Table 1 contains the distribution of story views, web and text elements, and number of subscribers for the story authors. We show the distribution percentiles instead of averages because the distribution of the data is long-tailed. The timespan is the time interval (in hours) in which users edit their stories and is calculated by taking the difference between the story creation-date and 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" rather than mean.

Table 1. Distribution of the features of the stories in the data set. Timespan is measured in hours.

Features	Views	Web elements	Text elements	Subscribers	Timespan
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	2,216	559	1,726,143	36,111

4.1 What Kind of Resources are in Stories?

Using the Storify-defined categories reflected in the UI (Figure 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?

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.

Table 2 contains the top 25 domains of the resources ordered by their frequency. The list of top 25 hosts represents 92.3% of all the resources. The table also contains the global rank of the domains according to Alexa as of March 2015. 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 (Section 4.1).

Although the top 25 list of domains appearing in the stories is dominated by globally popular web sites (e.g., Twitter, Instagram, Youtube, Facebook), the long-tailed distribution results in the presence of many globaly lesser known sites. In Section 4.3 we investigate the correlation between Alexa global rank and rank within Storify.

Table 2. The top 25 hosts that are used in human-generated stories. The percentage is the frequency of the host out of 1,150,399.

Host	Frequency	Percentage	Alexa Global Rank as of 2015-03	Category
twitter.com	943,859	82.05%	8	Social media
instagram.com	45,188	3.93%	25	Photos
youtube.com	22,076	1.92%	3	Videos
facebook.com	13,930	1.21%	2	Social media
flickr.com	7,317	0.64%	126	Photos
patch.com	5,783	0.50%	2,096	News
plus.google.com	3,413	0.30%	1	Social media
tumblr.com	3,066	0.27%	31	Blogs
blogspot.com	1,857	0.16%	18	Blogs
imgur.com	1,756	0.15%	36	Photos
coolpile.com	1,706	0.15%	149,281	Entertainment
wordpress.com	1,615	0.14%	33	Blogs
giphy.com	1,055	0.09%	1,604	Photos
bbc.com	966	0.08%	156	News
lastampa.it	927	0.08%	2,440	News
pinterest.com	892	0.08%	32	Photos
softandapps.info	861	0.07%	160,980	News
photobucket.com	768	0.07%	341	Photos
nytimes.com	744	0.06%	97	News
soundcloud.com	736	0.06%	167	Audio
wikipedia.org	736	0.06%	7	Encyclopedia
repubblica.it	682	0.06%	439	News
theguardian.com	588	0.05%	157	News
huffingtonpost.com	572	0.05%	93	News
punto-informatico.it	570	0.05%	42,955	News

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. This captures the behavior of users including tweets in the stories because the tweets are surrogates for embedded content. We sampled 5% from Twitter resources (47,512 URIs). Of sampled tweets in the stories, 32% (15,217) have embedded resources, of which there are 14,616 unique URIs. Of the 15,217, 46% are photos from twitter.com (hosted at `twimg.com`). Table 4 contains the most frequent 10 domains for the embedded resources, which

represent 61.6% of the all the URIs embedded in tweets. Note that some Storify stories (0.49%) point to other stories in Storify.

Table 3. The most frequent 10 hosts 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

4.3 Correlation of Global and Storify Popularity

We calculate the correlation between the frequency of the domains and their Alexa global traffic ranking. Table 4 shows Kendall's Tau τ correlation coefficient for the most frequent n domains. Statistically significant ($p < 0.05$) correlations are bolded. The highest correlation is 0.45 for list of 15 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 [20], which found that the most frequent sites on Pinterest had low Alexa Global Ranking. That possibly returns to the different nature 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, events. The most used subject areas that are being used by Pinterest users are food and drinks, décor and design, and Apparel and Accessories [5]. Most of the pins on Pinterest come from blogs, or uploaded by users. In Storify, the people tend to use social media and web resources to create their narratives about events, or something of interest.

Table 4. The correlation between the most frequent n domains in the stories and their global Alexa Ranking.

n	10	15	25	50	100
Kendall' τ	0.1555	**0.4476**	**0.3372**	**0.3194**	**0.2485**

4.4 What is the Average Timespan for Stories?

Table 5 shows the percentage of the stories in each time interval. The table also shows the corresponding features of the stories which are divided by their

timespan. We normalized the number of views by the age of the story (the time of existence of the story on the live web). The first two intervals represent the stories that were created and modified, 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 Storify users who update their stories over a long period of time, with the longest timespan in our data set covering more than four years and with more than 13,000 views. Curiously, it had only 33 web elements and 51 total elements. Although the story with the longest timespan did not have the largest number of elements, from Table 5 we can see that based on the median number of elements in each interval there is nearly linear relation between the time length of the story and the number of elements.

Table 5. The percentage of the stories based on the editing interval along with the median of web elements, text elements, and views. The percentage is out of 15,568 stories.

Intervals	Percentage	Median web elements	Median text elements	Median views
0-60 seconds	14.0%	15	0	23
1-60 minutes	26.7%	19	0	53
1-24 hours	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

5 Decay of Elements in Stories

Resources on the web are known to disappear quickly [18, 7, 14]. 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 checked the live web and public web archives for 265,181 URIs (202,452 URIs from story web elements + 47,512 randomly sampled tweet URIs + 15,217 URIs of embedded resources in those tweets), in which there are 253,978 unique URIs. We examined the results of the five most frequent domains in the stories (twitter.com, instagram.com, youtube.com, facebook.com, flickr.com).

5.1 Existence on the Live Web

From all the web resources, 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 [2]. The left two columns of Table 6 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 websites.

Table 6. The existence of the resources on the live web (on the left) and in the archives (on the right). Available represents the requests which ultimately return "HTTP 200", while missing represents the requests that return HTTP 4xx, HTTP 5xx, HTTP 3xx to others except 200, timeouts, and soft 404s. Total is the total unique URIs from each domain.

	Existence on live web			**Found in archives**		
Resources	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%	3,140
Facebook	95.2%	4.8%	12,793	0.6%	0.49%	80
Flickr	95.6%	4.4%	6,859	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%	1,257

5.2 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 an average value of 10.3% per story. Figure 2 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. This finding is very similar to the findings by SalahEldeen and Nelson [13], in which they found that resources linked to from social media resources disappeared at rate of 11% the first year and 7% for each following year.

5.3 Existence in the Archives

We checked the 253,978 pages for existence in general web archives in March 2015. The existence in the web archives was tested by querying Memento proxies and aggregator [17].

The right-most columns of Table 6 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. From the table we notice that the social media is not well-archived like the regular web [1]. Facebook

uses `robots.txt` to block web archiving by the Internet Archive, but the other sites do not have this restriction.

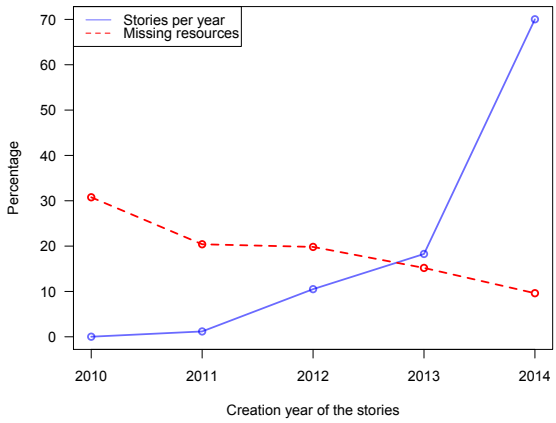


Fig. 2. The distribution of the stories per year and the decay rate of the resources in these stories through time.

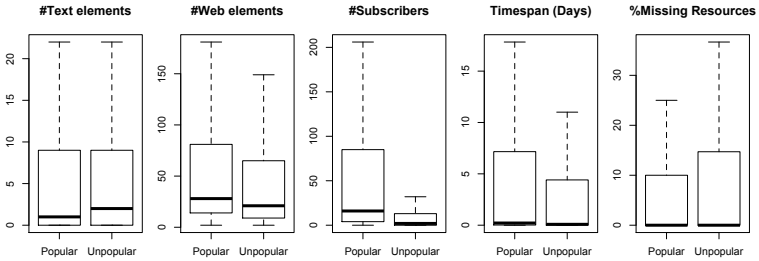
6 What Does a Popular Story Look Like?

In this section, we establish structural features for what differentiates popular stories from normal stories for building a baseline for the stories 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 they were available on the web. We took the top 25% of stories (3,642 stories) that have the most views and consider those as the popular stories. The 75th percentile of the views that we separated the data based on is 377 views/year.

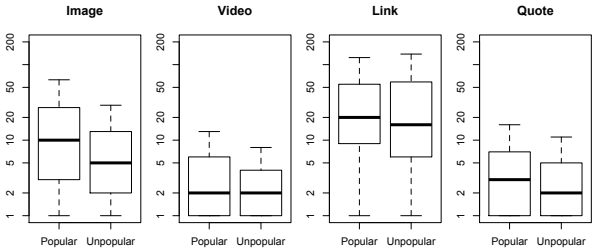
6.1 The Features of the Stories

We considered the distributions of several features of the stories: number of web elements, the number of text elements, and the editing timespan. We also check if there is a relation between the popular stories and the relative number of subscribers. Furthermore, we test if popular stories are different from the unpopular stories using Kruskal-Wallis test, 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 most of the features: number of web elements, text elements, timespan, and subscribers. Figure 3(a) shows that popular stories tend to have more web elements (medians of 28 vs. 21) and a longer timespan (5 hours vs. 2 hours) than the unpopular stories. The number of elements in



(a) The distributions for the features of the stories.



(b) The distributions for the elements of the stories.

Fig. 3. Characteristics of popular and unpopular stories.

the popular stories is between 2 to 1950 web elements with $median = 28$ and text elements from 0 to 559 with $median = 1$. The popular stories tend to have longer editing time intervals than the unpopular stories. For the popular stories, 38% have an editing timespan of at least one day, while 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.

6.2 The Type of Elements

Figure 3(b) shows the distributions for the popular and the unpopular stories for each element type. The figure shows that the distribution of the images in popular stories is higher than the distribution for the images in the unpopular stories. The median number of images in popular stories is 10, while it is 5 in the unpopular stories. For the videos the median is 2 for both popular and unpopular. Although the unpopular stories tend to use links more than the popular stories, the median of the links in popular stories (20 links) is higher than the unpopular stories (16 links). We also test if popular stories are different from the unpopular stories using the Kruskal-Wallis test, based on the elements and found $p \leq 0.05$ for all tests.

6.3 Do Popular Stories have a Lower Decay Rate?

We checked the existence of the missing resources of the popular stories and the unpopular stories to investigate if there is a correlation 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 3(a) 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 tends to stay longer than the resources of the unpopular. The 75th percentile of decay rate per popular story is 10% of the resources, while it is 15% in the unpopular stories.

7 Conclusions and Future Work

In this paper, we presented the structural characteristics of the 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 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 relation 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.

Future work will include investigating if these structural characteristics of stories hold for other social media storytelling services, such as `paper.li`, `scoop.it`, and `pinterest.com`. This study also will inform our future work of automatically creating stories to summarize collections of archived web pages. Using the structural characteristics of human-generated stories, such as number of elements, timespan, and a distribution of domains and types of resources, will provide us with a template with which to evaluate our automatically created stories.

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, ACM Press (2011) 133–136
2. 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 (2004) 328–337
3. Cohen, J., Mihailidis, P.: Storify and News Curation: Teaching and Learning about Digital Storytelling. In: Second Annual Social Media Technology Conference & Workshop. Volume 1. (2012) 27–31

4. 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)
5. Hall, C., Zarro, M.: Social curation on the website pinterest.com. American Society for Information Science and Technology **49**(1) (2012) 1–9
6. Kieu, B.T., Ichise, R., Pham, S.B.: Predicting the popularity of social curation. In: Knowledge and Systems Engineering. Springer (2015) 413–424
7. Klein, M., Nelson, M.: Find, new, copy, web, page-tagging for the (re-) discovery of web pages. Research and Advanced Technology for Digital Libraries (2011) 27–39
8. Laire, D., Casteleyn, J., Mottart, A.: Social Media's Learning Outcomes within Writing Instruction in the EFL Classroom: Exploring, Implementing and Analyzing Storify. Procedia-Social and Behavioral Sciences **69** (2012) 442–448
9. Mihailidis, P., Cohen, J.N.: Exploring Curation as a Core Competency in Digital and Media Literacy Education. Journal of Interactive Media in Education (2013) 19
10. 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 (2014) 386–395
11. 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 (2012) 437–438
12. Palomo, B., Palomo, B.: New information narratives: the case of storify. Hipertext.net **12** (2014)
13. SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: How many resources shared on social media have been lost? In: Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries. TPDL'12, Springer-Verlag (2012) 125–137
14. 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 (2013) 1075–1082
15. Seitzinger, J.: Curate me! exploring online identity through social curation in networked learning. In: Proceedings of the 9th International Conference on Networked Learning. (2014) 7–9
16. 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)
17. Van de Sompel, H., Nelson, M.L., Sanderson, R.: RFC 7089 - HTTP framework for time-based access to resource states – Memento. `http://tools.ietf.org/html/rfc7089` (2013)
18. Weiss, R.: On the Web, Research Work Proves Ephemeral. `http://stevereads.com/cache/ephemeral_web_pages.html` (2003)
19. 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, ACM (2014) 305–314
20. 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 (2013)