Quantifying Orphaned Annotations in Hypothes.is

Mohamed Aturban, Michael L. Nelson, and Michele C. Weigle

Dept of Computer Science, Old Dominion University, Norfolk, VA, 23529
{maturban,mln,mweigle}@cs.odu.edu
http://www.cs.odu.edu/

Abstract. Web annotation has been receiving increased attention recently with the organization of the Open Annotation Collaboration and new tools for open annotation, such as Hypothes.is. In this paper, we investigate the prevalence of *orphaned annotations*, where a live Web page no longer contains the text that had previously been annotated in the Hypothes.is annotation system (containing 6281 highlighted text annotations). We found that about 27% of highlighted text annotations can no longer be attached to their live Web pages. Unfortunately, only about 3.5% of these orphaned annotations can be reattached using the holdings of current public web archives. For those annotations that are still attached, 61% are in danger of becoming orphans if the live Web page changes. This points to the need for archiving the target of annotations at the time the annotation is created.

Keywords: Web Annotation, Web Archiving, HTTP

1 Introduction

Annotating web resources helps users share, discuss, and review information and exchange thoughts. Haslhofer et al. [9] define annotation as associating extra pieces of information with existing web resources. Annotation types include commenting on a web resource, highlighting text, replying to others' annotations, specifying a segment of interest rather than referring to the whole resource, tagging, etc.

In early 2013, Hypothes.is[1], an open annotation tool, was released and is publicly accessible for users to annotate, discuss, and share information. It provides different ways to annotate a web resource: highlighting text, adding notes, and commenting on and tagging a web page. In addition to that, it also allows users to share an individual annotation URI with each other as an independent web resource. The annotation is provided in JSON format and includes the annotation author, creation date, target URI, annotation text, permissions, tags, comments, etc.

One of the well-known issues of the Web is that Web pages are not fixed resources. A year after publication, about 11% of content shared on social media

[1] http://hypothes.is

will be gone [12, 13], and 20% of scholarly articles have some form of reference rot [10]. Lost or modified web pages may result in *orphaned annotations*, which can no longer be attached to their target web pages.

Figure 1 shows a web page `http://caseyboyle.net/3860/readings/against.html` which has 144 annotations from Hypothes.is. The text with darker highlights indicates more users have selected this part of the page to annotate. The issue here is that all of these annotations are in danger of being orphaned because no copies of the target URI are available in the archives. Figure 2 shows the annotation "Who does that someone have to be?" on the highlighted text "Get someone integrate it into bitcoin/litecoin/*coin" at the target URI `http://zerocoin.org/`, created in January 2014. In January 2015, this annotation can no longer be attached to the target web page because the highlighted text no longer appears on the page, as shown in Figure 3. Although the live Web version of `http://zerocoin.org/` has changed and the annotation is orphaned, the original version that was annotated has been archived and is available at Archive.today (`http://archive.today/20131201211910/http://zerocoin.org/`). The annotation could be re-attached to this archived resource.

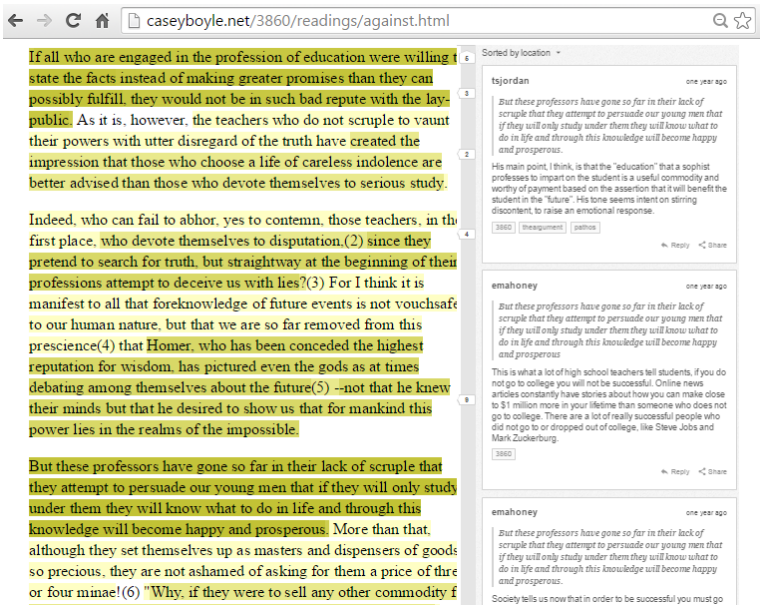


Fig. 1: Using the Hypothes.is Browser Extension to View the 144 Annotations of http://caseyboyle.net/3860/readings/against.html

In this paper, we present a detailed analysis of the extent of orphaned highlighted text annotations in the Hypothes.is annotation system as of January, 2015. We also look at the potential for web archives to be used to reattach these

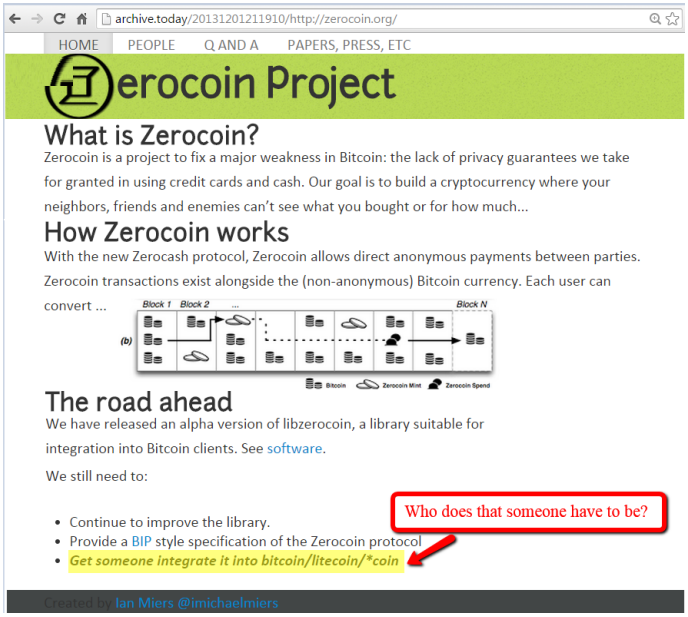


Fig. 2: http://zerocoin.org/ in January 2014

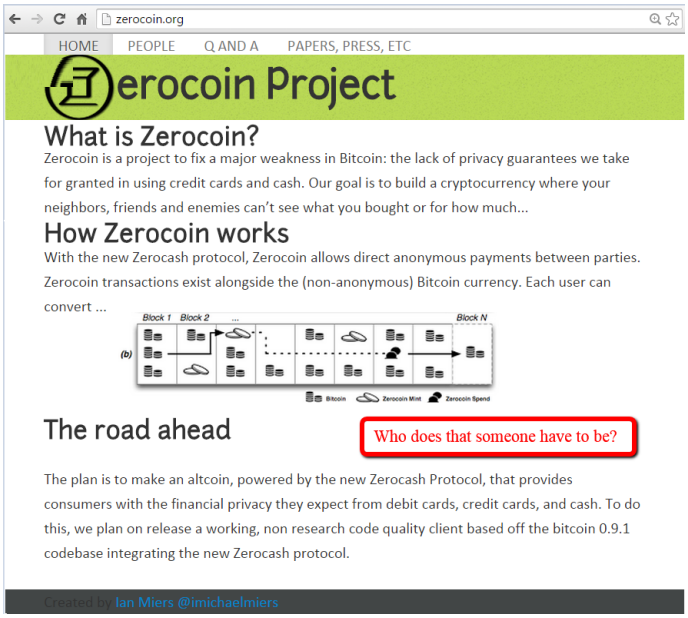


Fig. 3: http://zerocoin.org/ in January 2015

orphaned annotations to archived versions of their original targets. We find that 27% of the highlighted text annotations at Hypothes.is are orphans, and only

a few can be reattached using web archives. Further, we show that 61% of the currently attached annotations could potentially become orphans if their live Web resources change, because there are no archived versions of the annotated resources available. Our analysis points to the potential for reducing orphaned annotations by archiving web resources at the time of annotation.

2 Related Work

Annotation has long been recognized as an important and fundamental aspect of hypertext systems [11] and an integral part of digital libraries [1], but broad adoption of general annotation for the Web has been slow. Annotations have been studied for digital library performance [6, 17] and methods have been explored for aligning annotations in modified documents [4], but typically such studies are limited to annotation systems specific for a particular digital library. While orphaned annotations of general web pages have been studied in the context of Walden's Paths [7, 5], our study of Hypothes.is is a more recent evaluation of annotation and page synchronization in a widely deployed system.

Memento [18] is an HTTP protocol extension that aggregates information about the resources available in multiple Web archives. We can use Memento to obtain a list of archived versions of resources, or mementos, available in several web archives. In this paper, we use the following Memento terminology:

- URI-R - the original resource as it used to appear on the live Web. A URI-R may have 0 or more mementos (URI-Ms)
- URI-M - an archived snapshot of the URI-R at a specific date and time, which is called the Memento-Datetime, *e.g.*, $URI\text{-}M_i = URI\text{-}R@t_i$
- TimeMap - a resource that provides a list of mementos (URI-Ms) for a URI-R, ordered by their Memento-Datetimes

There has been previous work in developing annotation systems to support collaborative work among users and in integrating the Open Annotation Data Model [15] with the Memento framework. The Open Annotation Collaboration (OAC) [9] has been introduced to make annotations reusable through different systems like Hypothes.is. Before publishing OAC, annotations would not be useful if the annotated web pages were lost because annotations were not assigned URIs independent from the web pages' URIs. By considering annotations as first-class web resources with unique URIs, annotation not only would become reusable if their targets disappear, but also would support interactivity between systems. Sanderson and Van de Sompel [16] built annotation systems which support making web annotations persistent over time. They focus on integrating features in the Open Annotation Data Model with the Memento framework to help reconstructing annotations for a given memento and retrieving mementos for a given annotation. They did not focus on the case of orphaned annotations and assumed that the archived resources were available in web archives. Ainsworth et al. have estimated how much of the web is archived [2]. The result indicated that 35-90% of publicly accessible URIs have at least one archived

copy, although they did not consider annotations in their work, the result might estimate the number of orphaned annotations by factors like how frequently web pages are archived and the archiving process coverage. In other work [14, 8] researchers built annotation systems that can deliver a better user experience for specialized users and scholars. The interfaces allow users to annotate multimedia web resources as well as medieval manuscripts in a collaborative way. In this paper, we focus on orphaned annotations and investigate how web archives could be used to reattach these annotations to the original text.

3 Methodology

We performed our analysis on the publicly accessible annotations available at Hypothes.is. The interface allows users to create different types of annotations: (1) making a note by highlighting text and then adding comments and tags about the selected text, (2) creating highlights only, (3) adding comments and tags without highlighting text, and (4) replying to any existing annotations. In January 2015, we downloaded the JSON of all 7744 publicly available annotations from Hypothes.is. Figure 4 shows the JSON of the annotation from Figure 2 with relevant fields shown in bold. The "`updated`" field gives the annotation creation date, "`source`" provides the annotation target URI, "`type`":"`TextQuoteSelector`" indicates that it is a highlighted text annotation, "`exact`" contains the highlighted text, and "`text`" contains the annotation text itself. We focus only on annotations with highlighted text ("`type`":"`TextQuoteSelector`"), leaving 6281 annotations for analysis. To determine how many of those annotations are orphaned, for each annotation we performed the following steps:

- Determine the current HTTP status of the annotation target URIs ("`source`").
- Compare selected highlighted text ("`exact`") to the text of the current version of the URI.
- Discover available mementos for the target URI.
- Search for highlighted text within the discovered mementos.

In Table 1, we show the top 10 hosts with annotations at Hypothes.is. Many of these hosts, including the top three, are academic servers and appear to use the system for annotation of scholarly work. Apart from this listing, we did not attempt to make judgements about the content of the annotations or annotation target text in our analysis.

3.1 Determining the HTTP Status

In the first step, the current HTTP status of annotation target URIs can be obtained by issuing HTTP HEAD requests for all URIs. In addition, we extended this to detect "soft" 401, 403, and 404 URIs, which return a 200 OK status but actually indicate that the page is not found or is located behind authentication [3]. One technique we used to detect "soft" 4xx is to modify the original URI by adding some random characters. It is likely that the new URI does not exist.

```
{
  "updated": "2014-01-13T19:49:33.052047+00:00",
  "target": [
     {
          "source": "http://zerocoin.org/",
          "selector": [
             {
                "type": "RangeSelector",
                "startContainer": "/div[1]/div[3]/div[1]/
                                   ul[2]/li[3]/em[1]",
                "endContainer": "/div[1]/div[3]/div[1]/
                                 ul[2]/li[3]/em[1]",
                "startOffset": 0,
                "endOffset": 52
             },
             {
                "type": "TextQuoteSelector",
                "prefix": "cation of the Zerocoin protocol",
                "exact": "Get someone integrate it
                          into bitcoin/litecoin/*coin",
                "suffix": "Created by Ian Miers @imichaelm"
             },
             {
                "start": 5522,
                "end": 5574,
                "type": "TextPositionSelector"
             }
          ]
     }
  ],
  "created": "2014-01-13T19:49:33.052030+00:00",
  "text": "Who does that someone have to be?",
  "tags": [
    "bitcoin"
  ],
  "uri": "http://zerocoin.org/",
  "user": "acct:rdhyee@hypothes.is",
  "consumer": "00000000-0000-0000-0000-000000000000",
  "id": "SvI30rBYR52gpaCh_IiJgQ",
  "permissions": {
     "admin": [
        "acct:rdhyee@hypothes.is"
     ],
     "read": [
        "group:__world__",
           "acct:rdhyee@hypothes.is"
        ],
        "update": [
           "acct:rdhyee@hypothes.is"
        ],
        "delete": [
           "acct:rdhyee@hypothes.is"
        ]
     }
}
```


Fig. 4: An Annotation Described in JSON Format, Available at https://hypothes.is/api/annotations/SvI30rBYR52gpaCh_IiJgQ

After that, we download the content of the original URI and the new one. If the content of both web pages is the same, we consider that the HTTP status of the original URI is "soft" 4xx.

The returned responses will determine the next action which should be made for every URI. The resulting responses can be categorized into 3 different groups. The first group contains URIs with hostnames `localhost` or URIs which are

Number of Annotations	Host
1077	caseyboyle.net
886	rhetoric.eserver.org
246	umwblogs.org
131	hypothes.is
111	dohistory.org
101	www9.georgetown.edu
65	github.com
63	courses.ischool.berkeley.edu
55	www.nytimes.com
46	www.emule.com

Table 1: The Top Hosts With Annotated Pages

actually URNs. The second group has URIs with one of the following status codes: "soft" and actual 400, 401, 403, 404, 429 or Connection-Timeout. URIs with 200 status code belong to the third group.

The first group, localhost and URN URIs, were excluded completely from our analysis because these are pages that are not publicly accessible on the live Web. URIs in the second group, soft/actual 4xx and timed-out URIs, have been checked for mementos in the web archives. For URIs with response code 200, we have compared their associated highlighted annotation text with both the current version of the web page and the available mementos in the archives. Even though some annotations are still attached to their live web pages, we are still interested to see if they have mementos to know how likely those annotations are to become orphans if their current web pages change or become unavailable.

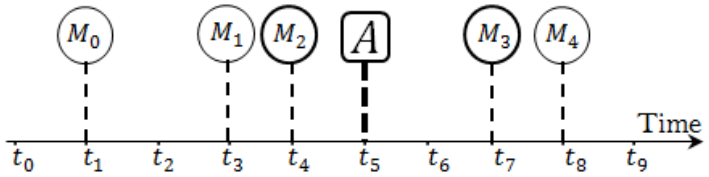
3.2 Are Annotations Attached to the Live Web?

The second step is to compare the annotated text ("`exact`") of each annotation target URI that has a 200 HTTP status code with the current version of its web page and see if they match; this can be done by downloading the web page and extracting only the text which will be compared to the highlighted annotation text. We use `curl` to access and download web pages. Then, we extract only the text after cleaning it by removing all HTML tags, extra white-space characters, and others. If the highlighted annotation text is not found in the web page, it is considered *not attached*. For example, as shown in Figure 3, the annotation text is no longer attached to the web page as the highlighted text, shown in Figure 2, has been removed from the live web page.

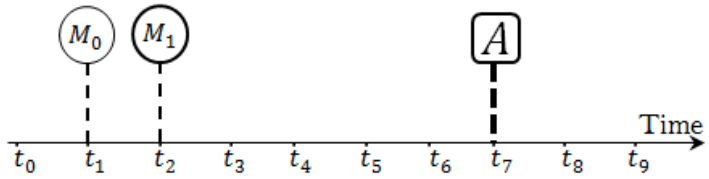
3.3 Discovering Mementos for All Valid URIs

The third step is in discovering mementos for all valid annotation target URIs. For this purpose, we used a Memento Aggregator [18], which provides a TimeMap of the available mementos for a URI-R. It would be a time-consuming task to check all available mementos for a URI-R to see whether they can be used to

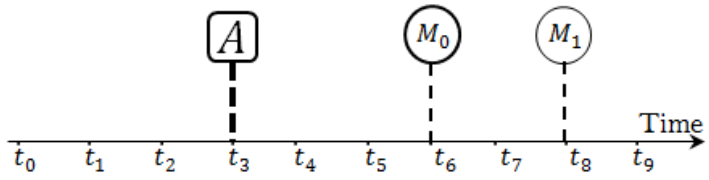
recover web pages. For example, URIs like `http://www.nytimes.com/` or `http://www.cnn.com/` have thousands of existing mementos in different archives. The strategy that we use here is effective in terms of execution time. For each URI, we only retrieve the nearest mementos to the annotation's creation date ("`updated`"). More precisely, we are capturing the closest memento(s) to the date *before* the annotation was created and the closest memento(s) to the date *after* the annotation was created.



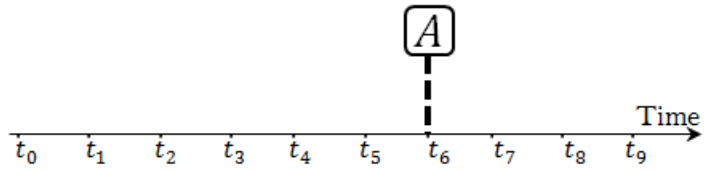
(a) Existing Mementos Before and After the Annotation Creation Date



(b) Existing Mementos Only Before the Annotation Creation Date



(c) Existing Mementos Only After the Annotation Creation Date



(d) No Mementos for the Annotation Target

Fig. 5: Annotation and Memento Creation Dates

In Figure 5(a), the annotation `A` was created at the time t_5. The closest memento to the date before t_5 was M_2 (captured at t_4) while the closest memento to the date after t_5 was M_3 (captured at t_7). So, for this annotation we picked the two closest mementos which are M_2 and M_3. Figure 5(b) is an example where

Number of Annotations	Status Code
5432	200
321	Time out
155	404
85	`localhost`
73	403
60	URN
47	401
46	410
37	Soft 401/403
17	400
8	Soft 404 and others

Table 2: HTTP Status Code for All Annotation Target URIs

mementos are only available before the annotation creation date while, in Figure 5(c), mementos are only available after the annotation creation date. It is also possible that an annotation target has no mementos at all as Figure 5(d) shows. If there are multiple closest mementos from different archives that share the same creation date (*memento-datetime*), then we consider all of these mementos for two different reasons. First, it is possible that at the time a memento is requested from an archive, there would be a technical problem or server-related issue which may affect returning the requested mementos. Second, we would like to know how different archives could contribute to provide mementos and recover annotation target text.

3.4 Are Annotations Attached to the Selected Mementos?

The final step is to see whether annotated URIs can be recovered by their mementos. The same technique introduced in Section 3.2 is used to test mementos. If the annotation target text ("`exact`") matches the text in the discovered memento, then we consider that this annotation is attached to the memento. Otherwise, we consider that the annotation cannot be attached.

4 Results

We collected 7744 annotations from Hypothes.is. Table 2 shows the results of checking the HTTP status code for the target URIs in each annotation. We find that 13.5% of the annotations have URI-Rs that are no longer available on the live Web. In our further analysis, we will focus only on the 6281 annotations that include highlighted text. After checking each annotation, we found that 4566, or 72.7%, of the highlighted text annotations are still attached to their live web pages. This means that the remaining 27.3% of the annotations are orphans.

Next for each annotation, we checked the archives for the presence of mementos of the target URI near the annotation creation date. In Table 3 we consider

Number of Annotations	Attached to Live Web Page	Attached to Memento (L)	Attached to Memento (R)
902	Yes	Yes	Yes
9	Yes	Yes	No
28	Yes	No	Yes
9	Yes	No	No
31	No	Yes	Yes
4	No	Yes	No
12	No	No	Yes
71	No	No	No

Table 3: Annotation Targets with Existing Mementos Before and After the Annotation Creation Date.

annotations that have mementos both before ("L") and after ("R") the annotation date. "No" under the L and R columns means that annotation cannot be attached to the nearest memento while "Yes" means that the annotation attaches to the nearest memento.

Table 4 shows the number of annotations that have mementos only on the L side (before annotations were created) of the annotation date, and Table 5 shows the number of annotations that have mementos only on the R side (after the annotation creation date) of the annotation date. Finally, Table 6 illustrates the number of annotations whose targets have no mementos. From these tables, we see that 1715 (27%) of the annotations can no longer be attached to their live web pages. Unfortunately, the current holdings of web archives only allow 61% of these to be re-attached. As shown in Table 6, the majority of annotations have no mementos available at all. Those that can no longer be attached to their live web version are lost, but those that are still attached can be recovered if these pages are archived before the annotated text changes.

Table 7 shows the number of annotations that can be recovered using various archives, split by whether or not they are still attached to the live web. As expected `web.archive.org` can be used to recover the most annotations, but for those annotations not attached to the live web, we find that `archive.today`, an on-demand service, can recover more orphaned annotations.

Number of Annotations	Attached to Live Web Page	Attached to Memento (L)
599	Yes	Yes
11	Yes	No
14	No	Yes
68	No	No

Table 4: Annotation Targets with Existing Mementos Only Before the Annotation Creation Date

Number of Annotations	Attached to Live Web Page	Attached to Mementos (R)
0	Yes	Yes
24	Yes	No
0	No	Yes
25	No	No

Table 5: Annotation Targets with Existing Mementos Only After the Annotation Creation Date

Number of Annotations	Attached to Live Web
2737	Yes
1289	No

Table 6: Annotation Targets with No Existing Mementos

Archive	Attached to Live Web	Not Attached to Live Web
web.archive.org	1546 (86.5%)	23 (20.3%)
archive.today	249 (13.9%)	59 (52.2%)
wayback.archive-it.org	101 (5.65%)	14 (12.3%)
wayback.vefsafn.is	74 (4.14%)	18 (15.9%)
webarchive.nationalarchives.gov.uk	0 (0%)	2 (1.76%)
Total	1786 (110.2%)	113 (102.5%)

Table 7: Annotation Targets Recovered by Different Archives

5 Conclusions

In this paper, we analyzed the attachment of highlighted text annotations in Hypothes.is. We studied the prevalence of orphaned annotations, and found that 27% of the highlighted text annotations are orphans. We used Memento to look for archived versions of the annotated pages and found that orphaned annotations can be reattached to archived versions, if those archived versions exist. We also found that for the majority of the annotations, no memento exists in the archives. This points to the need for archiving pages at the time of annotation.

References

1. Agosti, M., Ferro, N., Frommholz, I., Thiel, U.: Annotations in digital libraries and collaboratories–facets, models and usage. In: Research and Advanced Technology for Digital Libraries, pp. 244–255 (2004)
2. 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 Joint Conference on Digital Libraries (JCDL). pp. 133–136. ACM (2011)

3. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic transit gloria telae: Towards an understanding of the web's decay. In: WWW '04: Proceedings of the 13th International Conference on World Wide Web. pp. 328–337 (2004)
4. Brush, A., Bargeron, D., Gupta, A., Cadiz, J.J.: Robust annotation positioning in digital documents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 285–292. ACM (2001)
5. Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., Arora, A.: Managing change on the web. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 67–76. ACM (2001)
6. Frommholz, I., Fuhr, N.: Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In: Proceedings of the 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 55–64. ACM (2006)
7. Furuta, R., Shipman III, F.M., Marshall, C.C., Brenner, D., Hsieh, H.w.: Hypertext paths and the world-wide web: Experiences with walden's paths. In: Proceedings of the 8th ACM Conference on Hypertext. pp. 167–176. ACM (1997)
8. Haslhofer, B., Sanderson, R., Simon, R., Van de Sompel, H.: Open annotations on multimedia web resources. Multimedia Tools and Applications 70(2), 847–867 (2014)
9. Haslhofer, B., Simon, R., Sanderson, R., Van de Sompel, H.: The Open Annotation Collaboration (OAC) model. In: Proceedings of the IEEE Workshop on Multimedia on the Web (MMWeb). pp. 5–9. IEEE (2011)
10. 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), e115253 (2014)
11. Marshall, C.C.: Toward an ecology of hypertext annotation. In: Proceedings of the 9th ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space—Structure in Hypermedia Systems. pp. 40–49. ACM (1998)
12. SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: How many resources shared on social media have been lost? In: Proceedings of Theory and Practice of Digital Libraries (TPDL), pp. 125–137 (2012)
13. SalahEldeen, H.M., Nelson, M.L.: Resurrecting my revolution: Using social link neighborhood in bringing context to the disappearing web. In: Proceedings of Theory and Practice of Digital Libraries (TPDL). pp. 333–345 (2013)
14. Sanderson, R., Albritton, B., Schwemmer, R., Van de Sompel, H.: SharedCanvas: A collaborative model for medieval manuscript layout dissemination. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 175–184. ACM (2011)
15. Sanderson, R., Ciccarese, P., Van de Sompel, H.: Designing the W3C open annotation data model. In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 366–375. ACM (2013)
16. Sanderson, R., Van de Sompel, H.: Making web annotations persistent over time. In: Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 1–10. ACM (2010)
17. Soo, V.W., Lee, C.Y., Li, C.C., Chen, S.L., Chen, C.c.: Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In: Proceedings of the 2003 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 61–72. IEEE (2003)
18. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time Travel for the Web. Tech. Rep. arXiv:0911.1112 (2009)