



BIROn - Birkbeck Institutional Research Online

Eve, Martin Paul (2017) Towards the digital preservation of DOM-node-keyed scholarly web annotations. *Journal of Librarianship and Scholarly Communication* 5 (1), eP2178.1-eP2178.13. ISSN 2162-3309.

Downloaded from: <http://eprints.bbk.ac.uk/18661/>

Usage Guidelines:

Please refer to usage guidelines at <http://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | <http://jpsc-pub.org>

Volume 5, General Issue (2017)

Towards the Digital Preservation of DOM-Node-Keyed Scholarly Web Annotations

Martin Paul Eve

Eve, M. P. (2017). Towards the Digital Preservation of DOM-Node-Keyed Scholarly Web Annotations. *Journal of Librarianship and Scholarly Communication*, 5(General Issue), eP2178. <https://doi.org/10.7710/2162-3309.2178>



© 2017 Eve. This open access article is distributed under a Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

Towards the Digital Preservation of DOM-Node-Keyed Scholarly Web Annotations

Martin Paul Eve

Professor of Literature, Technology and Publishing, Birkbeck University of London

The current generation of annotation technologies uses a set of keying techniques, often based on the Document Object Model (DOM) for representing HTML content, that links an annotation to its target content. However, when the DOM structure changes, for any reason, or browser rendering engines parse the underlying source differently, annotations can be orphaned and incorrectly re-attached. This article explores the preservation strategies that are required to ensure the longevity of scholarly annotations that use such technologies. These recommendations range from the social changes needed for the perception of annotations as first-class scholarly objects through to the technological changes and infrastructures that are needed for the preservation of such objects. It concludes with a series of recommendations for changes in practice and infrastructure that work towards the digital preservation of DOM-node-keyed scholarly web annotations.

IMPLICATIONS FOR PRACTICE

1. Publishers should commit to providing versions of articles in a strict context free grammar language, such as XHTML Strict, so that annotations can always accurately be re-keyed to a DOM.
2. A set of guidelines/briefing documents for publishers about the challenges in DOM reattachment for annotation technologies should be produced and circulated.
3. Those writing annotation tools should stick to the smallest set of commonly used technologies that is possible, so that publisher barriers to understanding the digital preservation of annotations are not too high.
4. Those writing annotation tools should provide export functionality that can generate static XHTML versions of a current DOM representation with annotations pre-attached (and statically rendered) for preservation purposes. These could, for instance, be archived periodically on publisher sites and then ingested via LOCKSS or CLOCKSS.

Received: 01/13/2017 Accepted: 05/05/2017

Correspondence: Martin Paul Eve, Birkbeck University of London, Department of English and Humanities, School of Arts, 43 Gordon Square, London WC1H 0PD, martin.eve@bbk.ac.uk



© 2017 Eve. This open access article is distributed under a Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

In recent years, “the future of the academic library” or even “the future of academic librarians” has been a topic much in vogue. Some, such as Deborah Schwarz, approach this debate from a technological perspective, calling for increased specialization among librarians to counter a perceived erosion of the physical library space (Schwarz, 2016). Others, such as Emily Drabinski, point out that Schwarz and others demanding such specialization are self-serving in their arguments, running companies that offer “cheaper access to librarian labor” through market segmentation (Drabinski, 2016, p. 59). Such debates look set to continue for quite some time.

One element that seems to be at stake in such debates, though, is the way in which the library and its staff act as *enframing contexts* for information. In an era dominated by print, the material scarcity of the budget and the stack-shelf space required often acted as legitimation criteria and valorisation signals for academics and students. That is, the presence of work within the collection of a university library conferred a certain value upon that work. This does not mean that such works would always be positively appraised; many history sections of university libraries contain copies of Adolf Hitler’s writings, as just one example. The value conferred here is of a different type: it is one of significance. It is to deem the work worthy of discussion without determining the final appraisal of the work’s truth.

The age of digital abundance begins to threaten this model of value conferral. This is not necessarily a terrible event since it entails an academic reflection on the structures and proxy measures through which we decide where to spend scarce attention (that is: scarce academic labour time). It is also not quite so imminent as the major detractors feel. The potential of the internet for unfettered, infinite, and open dissemination of nonrivalrous digital objects, as Peter Suber puts it (Suber, 2012, p. 46), is constrained by the material economics of publisher business models. While we may have the technological capacity to expand publication indefinitely, we do not have the material economic resources to underwrite the labour to do so. Or, at least, we do not have the business models in place to effect the same financial distribution as is currently achieved by subscription or purchase mechanisms.

Despite this slow start, the day is nonetheless foreseeable when the scarcity of print and the paucity/unevenness of budget will no longer work as value signals in scholarly communication environments. The works that we will need to discuss, as global communities, are less likely to be denoted through digital frames that would be recognizable to academics of the twentieth century. Such value-conferring discussions have been envisaged as taking place in virtualised spaces, be it in annotation environments or online discussion fora. Indeed, prominent library blogs are now already recommending the use of online

annotation tools within scholarly environments, and library training programmes for doctoral researchers are promoting such software for general use (Troia, 2017).

While it may be some time until we can conceive of the value frames and information literacy needs that will make such discussions possible and productive, several pieces of software have emerged that allow end-users to annotate documents presented in the (eX-tensible) HyperText Markup Language ((x)HTML) or other comparable web-accessible markup languages. Among these, “Hypothes.is,” founded by Dan Whaley, has taken a prominent place, based upon the previous work of Annotator.js. The basic premise of use for such tools is that users’ annotations are “keyed” to specific “locations” on a page. Many users can annotate a single document. This is theorized to be particularly useful in the context of scholarly communications environments since enframing discussion can take place in direct proximity to the material that is under consideration. Indeed, the importance of (digital) annotation as a generalised scholarly activity has been evidenced not only through a series of grants from the Andrew W. Mellon Foundation to the Open Annotation Collaboration (OAC) and then to Hypothes.is, but also in Jisc reports and in the fact that John Unsworth referred to annotation as one of the seven “scholarly primitives” in 2000 (Jisc, 2015; Unsworth, 2000; Waters & Cullyer, 2014).

Yet such annotation technologies pose a range of challenges for libraries and for digital preservation, and they currently sit among the “enormous amounts of digital information [that is] already lost forever” (Kuny, 1998, p. 2). It is also the case that, in the specific instance of annotation technologies and their linkages, the metadata for attachment to a web page is by necessity so rich as to add considerable overhead for preservation purposes (Rosenthal, 2014). While the software that provides the annotation functionality is usually open source (Hypothes.is and Annotator.js, for instance) and this is also often the case for the data that sit underneath such software, there are a set of social, technological, and infrastructural challenges in preserving these paratextual materials in a way that can be accurately reconstructed at an undefined future moment. In this article I set out the social difficulties in understanding the open-source technologies that are used to key annotations to the Document Object Model (DOM) in Hypothes.is and Annotator.js; the technological problems of orphan annotations within a context of DOM mutation; and finally the practical infrastructure that is becoming necessary to ensure the ongoing accessibility of digital annotations on scholarly material. If the library is to continue to provide frames of value construction in its imagined digital futures, one such function must be to ensure ongoing access to otherwise digitally ephemeral resources. This is especially the case when those working within libraries are beginning to recommend such tools without fully appraising the preservation needs of such material.

UNDERSTANDING ANNOTATION TECHNOLOGIES AND THE DOCUMENT OBJECT MODEL

In order to understand the way in which the current breed of web annotation technologies function, it is first necessary to understand a little of the underlying document presentation techniques upon which they rest. XHTML or HTML documents are part of a series of technologies that power the modern World Wide Web and that derive from XML (eXtensible Markup Language) and SGML (Standard Generalized Markup Languages), first proposed by Tim Berners-Lee at CERN in the early 1990s (Berners-Lee, 1990). The basic premise is that such documents encode their information content within semantic “tags”. For instance, a paragraph in HTML is represented thus: “<p>A paragraph.</p>”, consisting of an opening “p” tag, the information content, and a closing tag. In the most basic scenarios, HTML documents are typically written on the server side, often by computer programs themselves (in order to provide dynamic content), and then served to client browsers over the internet.

Browser rendering engines, such as Mozilla’s Gecko, Google’s Blink, and KDE’s/Apple’s WebKit, take such marked-up content and render it visible for an end user within browsers such as Firefox, Chrome, or Safari. Since the nested-tag structure of HTML and related technologies makes them conducive to representation in a tree format, the typical first task of a browser engine is to create an in-memory representation of the document that *models* the underlying HTML. This is called the DOM, and it is the result of a twofold process of tokenization and tree construction. This in-memory construction (the DOM) of the HTML document is then rendered to the end-user and made available for modification by javascript engines on the page.

While, in an ideal situation, “the DOM has an almost one-to-one relation to the markup”, Tali Garsiel has noted that, since most implementations of HTML parsing do not treat the language as having a context free grammar, there is neither a standardised top-down nor bottom-up parser-lexer process for generating the DOM (Garsiel, 2011). The reasons for this, according to Garsiel, lie threefold in: “1.) The forgiving nature of the language, 2.) The fact that browsers have traditional error tolerance to support well known cases of invalid HTML, and 3.) The parsing process is reentrant. For other languages, the source doesn’t change during parsing, but in HTML, dynamic code (such as script elements containing `document.write()` calls) can add extra tokens, so the parsing process actually modifies the input” (Garsiel, 2011). In short, “[u]nable to use the regular parsing techniques, browsers create custom parsers for parsing HTML” (Garsiel, 2011). This is important since, in theory, it then becomes possible for the same HTML input to generate different internal DOM tree representations when viewed in different browser layout engines.

Hypothes.is and Annotator.js use a variety of methods to key their annotation text to a specific target location within the DOM. These annotations are stored in a database that is owned by the host institution that is running the centralized server software (and that must be independently preserved, an aspect that I do not cover here). The techniques for re-attaching annotations to the DOM are defined in a proposed Web Annotation Data Model as a W3C Candidate Recommendation, as of the 22 November 2016 (Sanderson, Ciccarese, & Young, 2016). The methods that such software uses to identify fragments within an HTML document are: fragment selectors, CSS selectors, XPath selectors, Text Quote selectors, and Range selectors. While a full description of each of these selector methods is outside the scope of this article, and they are instead each covered in more detail in the W3C document, it is important to note that at least four events can break the ability of a selector to accurately link to an annotation's target fragment:

1. A change in the parsing process of the browser engine that results in a different DOM;
2. A difference between the parsing processes of different browser engines, resulting in a different DOM between different browsers;
3. Dynamic javascript (or other script) injection modifying the underlying DOM structure; and
4. Publishers modifying an HTML document's structure.

These correspond, respectively, to a subset of the threats that David S. H. Rosenthal *et al.* have identified (Rosenthal, Robertson, Lipkis, Reich, & Morabito, 2005). That is: software failure/software obsolescence (1, 2 and 3); and operator error (4). These structures are, of course, dependent upon the highly correlated lower-level threats that the authors present within the LOCKSS threat model (hardware failure etc.). It is also true that these threats are a type of format migration issue. Although Rosenthal has argued prominently that disproportionate resources are poured into format migration against unlikely scenarios (Rosenthal, 2007), the case here is somewhat different since the very availability of these artefacts depends upon a set of dynamic renderers that attempt to match against a presumed underlying stable document.

Although all technical problems are, in reality, the results of complex social processes and can, therefore, be defined as social problems (Eve, 2014, pp. 43–44), it may nonetheless be helpful to see the above list as subdivided into technical problems (items 1 to 3) and a social problem (item 4). Moving to the latter of these divisions first (social problem), I will then return to the technological difficulties that result in orphan annotations and non-accessibility from a digital preservation perspective.

THE SOCIAL CONTEXTS FOR THE PRESERVATION OF SCHOLARLY ANNOTATIONS

Most of the problems around the preservation of annotations are problems of resourcing. While it is claimed that these artefacts are valued, few resources have thus far been invested in their long-term storage and availability. As in all cases of digital preservation, there are financial, social choices that we must make in the present about what we value for the future.

That said, and returning to my above point about document structure modification, the reasons for which publishers change the online layouts of their websites that contain scholarly material have not been formally studied, and this is an area for future work. However, I hypothesize that the core reasons lie within a combination of the business logic of brand aesthetics and the emergence of new technologies. For instance, Taylor & Francis noted of their 2016 redesign that key reasons included a “responsive design” paradigm that would display well on multiple devices of different kinds alongside “clear access indicators” (to show the subscription status, or otherwise, of work) and a renewed focus on “prominent journal branding” (Taylor & Francis, 2016). Importantly, matters of digital preservation do not feature in any of the commentary on publisher website redesign here.

While many publishers are aware of the challenges of the digital preservation of scholarly articles (to which I am deliberately limiting the discussion here for the sake of concision), they also seem unaware that the strict form/content dichotomy within which website redesigns are undertaken poses problems for the preservation of paratexts such as annotations. Although LOCKSS and CLOCKSS, for example, can ingest the HTML version of an article, and many publishers understand that the fundamental content of the article must remain the same, even small changes such as adding a correction or addendum notice to an article can corrupt the DOM integrity for the purposes of annotation keying (see, for example, the correction notice on Lawson, Gray, & Mauri, 2016). Such notices, while important, cause DOM mutation from the state stored in existing annotations and may render XPath keying methods unusable for their re-attachment.

There are two core social reasons that I posit for this. The first is that annotations are not currently viewed as first-class citizens in the scholarly publishing community. For, while many prominent scholarly publishers are members of the “Annotating All Knowledge” coalition, a set of secondary or supporting literatures (that is, annotations) that are outside of these entities’ control are likely to always be of secondary significance to the business concerns of such actors (“Annotating All Knowledge,” 2016). To see active social change in the valorisation of such paratextual discussions would require a coordinated effort by many agents, including scholarly societies that provide citation guidance on how to cite annotations as first-class scholarly objects. Since many scholars have taken a long time to recognise

(or even still do not recognise) other grey literatures, such as blogs, this social change could be a long way off.

The second core social problem, though, is that publishers do not often possess the requisite in-house technical expertise to understand the complexities of the way in which annotations are keyed to the DOM. Indeed, to understand the codebase of Hypothes.is requires a developer who is full-stack proficient and who has experience with the Pyramid framework, Annotator.js, Javascript, Coffeescript, XMLHttpRequest (XHR) channels, Node.js, AngularJS, XQuery, XPath, Python, JQuery, Elasticsearch, Postgres, and an array of other technologies. It is extremely difficult and expensive to hire developers with such a range of knowledge, particularly when annotations are also not yet valued as first-class objects.

That said, without such an understanding of the consequential implications for annotation of redesigning web pages or amending content flows by additional node insertion, we risk creating a set of information resources that cannot be accurately preserved and reconstructed. Furthermore, as I will shortly show, at least one solution to this problem—a centralized preserved first copy version of an article—creates further social difficulties for publishers who wish to control the version of record for the purposes of metric gathering. The social problems around the preservation of scholarly annotations pertain, therefore, to the stability of the document *structure*.

THE TECHNOLOGICAL DIFFICULTIES OF PRESERVING SCHOLARLY ANNOTATIONS

In addition to these social problems, there are at least the three central technological challenges (outlined above) that relate perhaps less to the storage but more to the reconstruction of preserved scholarly annotations. For the most part, these pertain to the differing *interpretation* of the document structure *by browser technologies*.

If, as Reagan Moore puts it, a “preservation environment is the software middleware that shields records from the rapid evolution of technology,” then what do we do when the very viewing environment for the records (web browsers and browser rendering engines) are part, themselves, of a rapid technological evolution (Moore, 2008, p. 65)? There have been proposals for ways in which we might preserve such software (Matthews, Shaon, Bicarregui, & Jones, 2010). However, since different browsers, at different points in their development history, can create different DOM trees from different parsing processes, even on the same documents, is it necessary to preserve browsers and operating systems within standardised virtualised containers so that the precise rendering process for annotation attachment at any moment can be reconstructed? Might it also be the case, though, that such a preservation system would cause such technical overhead as to deter the majority of readers from

seeking out the preserved annotations (von Suchodoletz & van der Hoeven, 2009)? Such structures also divert economic resources away from an already underfunded preservation environment (Rosenthal, 2015, p. 31). Certainly, there is also an instance of Gödel's incompleteness theorem if we try to preserve a formally complete axiomatic system for access (that is, to preserve information about all the formats in which we store information would take more available storage space than is available in the universe).

Since this virtualised solution seems unlikely to come to fruition within any realistic time-frame (and, currently, since they hold little realistic prospect of usability for the purposes of accessing preserved annotation content), a different set of practices seems necessary that could militate against the variability of DOM interpretation between browsers over time, despite the existence of formalised test standards that attempt to codify this. The two changes that would make the most difference here, it seems to me, are: 1. to use an XML presentation language, such as XHTML strict, which allows for context free grammar parsing without ambiguity in DOM interpretation; and 2. to strip all DOM-modifying javascript from scholarly communications web pages.

However, because scholarly communications take place over a range of disaggregated systems, organizations, and processes, it seems unlikely that the necessary coordination between entities could be gathered. To this end, I want to turn, in the final section of this article, to a description of an interim prototype solution that was developed as part of a three day collaborative workshop between Columbia University's Group for Experimental Methods in the Humanities and Birkbeck, University of London's Centre for Technology and Publishing. The goal of the workshop, held in late 2016 and organised by Alex Gil, was to discuss the challenges of preserving scholarly web annotations and, if possible, to rapidly prototype a solution that was more immune to the problems set out above (Eve & Gil, 2016).

The result of this workshop was a simple piece of software, dubbed "Cemmento," that performed the following actions:

1. When a user asks to annotate a page, instead of loading the Hypothes.is sidebar, the software instead first queries the Internet Archive to ascertain whether a copy has been stored;
2. If there is no existing copy in the Internet Archive, the software stores a copy;
3. The software redirects the user to the version stored in the Internet Archive and loads the Hypothes.is sidebar on that version.

This process was designed to shield against several of the social issues set out above. Firstly, for instance, it certainly protects against DOM changes by the publisher. By storing a centralized copy, and annotating this, we circumvent the need for publishers to understand the complex DOM keying structures inherent within most contemporary web annotation procedures. On the other hand, within the prototype that we built, using the Internet Archive, we still have no guarantee that the centralized backend store will not, itself, change the layout and/or DOM structure thereby resulting in orphaned annotations.

A set of further problems exist within our prototype: we do not mitigate the challenges of trans-browser interpretative errors with respect to in-memory DOM presentation, and we do not strip out javascript that may result in DOM mutation. Furthermore, the Internet Archive is unable to correctly ingest content that sits behind a paywall, which rules out the vast majority of current scholarly content. Indeed, in order to thoroughly guard against these types of problem we would instead need a set of infrastructural changes and provisions.

INFRASTRUCTURAL CHANGES AND PROVISIONS

If annotations, as they are implemented in current software solutions as of 2017, are to be robustly digitally preserved, a range of infrastructural changes and provisions are needed in an ideal situation:

1. Publishers should commit to providing versions of articles in a strict context-free grammar language, such as XHTML Strict. This avoids the challenges of DOM parsing causing problems for annotation reattachment within browser rendering engines.
2. A neutral storage backend with access credentials and authentication/authorisation procedures for paywalled content is required for a proxy such as Cernanto that can commit to providing unframed, straight rendering of its content and that strips out all DOM-modifying javascript from the preserved copy. This mitigates the problem of publisher content changing in the event of site redesign, although it also creates a separate problem of version proliferation and also has the preservation problem of unfaithful reproduction (for instance, some future user may wish to study the DOM-modifying javascript as an integral part of the document. Hence this preservation strategy works only for annotations).
3. Annotation of resources that are to be preserved should take place on a central, canonicalised resource that combines points #1 and #2 above. This solves the problem of orphan annotations and their re-attachment but comes with considerable social challenges since publishers are unlikely to be keen to see

annotation taking place away from their own platform. It also re-introduces an element of centralization into an otherwise de-centralized system.

These specific recommendations vary in the degree of complexity for implementation. The first is not a technically difficult challenge, particularly as many publishers generate their (X)HTML output as a result of XSL transformations upon Journal Article Tag Suite (JATS) XML. The second is also, likewise, not a technically difficult task. However, it is a complex social problem to create a financially resilient and durable, neutral organisation that can provide such infrastructure over the long term (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008). The final proposed point here is a large social challenge. It is unlikely that the many heterogeneous bodies involved in scholarly communications structures would agree on where best to place this off-site annotated copy and will also likely object towards such centralisation away from their own platforms.

Given the unlikelihood of these measures coming to fruition in the short term, the following action points could serve as an intermediate set of practicable guidelines to properly pave the way to the digital preservations of scholarly web annotations:

1. As above, publishers should commit to providing versions of articles in a strict context free grammar language, such as XHTML Strict. Indeed, the production of a set of guidelines/briefing documents for publishers around the challenges in DOM reattachment for annotation technologies would also help here.
2. Those writing annotation tools should stick to the smallest set of commonly used technologies that is possible, so that publisher barriers to understanding the digital preservation of annotations are not too high.
3. Those writing annotation tools should provide export functionality that can generate *static XHTML* versions of a current DOM representation with annotations pre-attached (and statically rendered) for preservation purposes. These could, for instance, be archived periodically on publisher sites and then simply ingested via LOCKSS or CLOCKSS.

True change in the preservable status of annotations is still a long way off. Despite being seen as scholarly primitives, the preservation of these paratexts is of a lower priority than the challenges that face us elsewhere. However, as libraries seek to position themselves as enframers, as just one strategy in a time of digital change, it becomes necessary to consider the implications of new tools for long-lasting access to such paratexts. The above recommendations could, I hope, lead us on a path towards the thinking, if not yet the practice, of the digital preservation of DOM-node-keyed scholarly web annotation.

ACKNOWLEDGEMENTS

The author wishes to thank Columbia University's Group for Experimental Methods in the Humanities and Birkbeck, University of London's Centre for Technology and Publishing for the productive workshop that underpins this article. In particular, thanks should go to Alex Gil for hosting the author in late 2016 and greatly advancing the thinking herein.

REFERENCES

- Annotating All Knowledge. (2016). Retrieved December 24, 2016, from <https://hypothes.is/annotating-all-knowledge/>
- Berners-Lee, T. (1990). *The original proposal of the WWW, HTMLized*. Retrieved December 24, 2016. Retrieved from <https://www.w3.org/History/1989/proposal.html>
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008). *Sustaining the digital investment: Issues and challenges of economically sustainable digital preservation*. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf
- Drabinski, E. (2016). Editorial response: A librarian by another name. *Journal of New Librarianship*, 1(1). <https://doi.org/10.21173/newlibs/2016/1/drabinski.1>
- Eve, M. P. (2014). *Open Access and the Humanities: Contexts, Controversies and the Future*. Cambridge: Cambridge University Press. Retrieved from <https://doi.org/10.1017/CBO9781316161012>
- Eve, M. P., & Gil, A. (2016). MartinPaulEve/cemmento. Retrieved from <https://github.com/MartinPaulEve/cemmento>
- Garsiel, T. (2011). *How browsers work: Behind the scenes of modern web browsers*. Retrieved from <https://www.html5rocks.com/en/tutorials/internals/howbrowserswork/>
- Jisc. (2015). Conference report: Jisc CNI meeting 2014: Opening up scholarly communications. 10-11 July 2014, Bristol. Retrieved from <https://www.jisc.ac.uk/sites/default/files/scholarly-communication-the-journey-towards-openness-october-2015.pdf>
- Kuny, T. (1998). The digital dark ages? Challenges in the preservation of electronic information. *International Preservation News*, 17, 8–13.
- Lawson, S., Gray, J., & Mauri, M. (2016). Opening the black box of scholarly communication funding: A public data infrastructure for financial flows in academic publishing. *Open Library of Humanities*, 2(1). <https://doi.org/10.16995/olh.72>
- Matthews, B., Shaon, A., Bicarregui, J., & Jones, C. (2010). A framework for software preservation. *International Journal of Digital Curation*, 5(1), 91–105. <https://doi.org/10.2218/ijdc.v5i1.145>

- Moore, R. (2008). Towards a theory of digital preservation. *International Journal of Digital Curation*, 3(1), 63–75. <https://doi.org/10.2218/ijdc.v3i1.42>
- Rosenthal, D. S. H. (2007, April 29). Format obsolescence: Scenarios. Retrieved from <http://blog.dshr.org/2007/04/format-obsolescence-scenarios.html>
- Rosenthal, D. S. H. (2014, March 31). The half-empty archive. Retrieved from <http://blog.dshr.org/2014/03/the-half-empty-archive.html>
- Rosenthal, D. S. H. (2015). *Emulation & virtualization as preservation strategies*. The Andrew W. Mellon Foundation. Retrieved from <https://mellon.org/Rosenthal-Emulation-2015/>
- Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine*, 11(11). <https://doi.org/10.1045/november2005-rosenthal>
- Sanderson, R., Ciccarese, P., & Young, B. (2016). *Web annotation data model*. Retrieved December 24, 2016, from <https://www.w3.org/TR/annotation-model/>
- Schwarz, D. (2016). Editorial essay: A librarian by another name. *Journal of New Librarianship*, 1(1). <https://doi.org/10.21173/newlibs/2016/1/schwarz>
- Suber, P. (2012). *Open Access*. Cambridge, MA: MIT Press. Retrieved from <http://bit.ly/oa-book>
- Taylor & Francis. (2016). Redesigned, refreshed, responsive: Your new Taylor & Francis Online. Retrieved December 24, 2016, from <http://explore.tandfonline.com/page/gen/introducing-the-new-taylor-francis-online>
- Troia, L. (2017, January 25). Hypothes.is and the dream of universal web annotation. Retrieved from <http://acrlog.org/2017/01/25/hypothes-is-and-the-dream-of-universal-web-annotation/>
- Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? Presented at the Humanities Computing: Formal methods, experimental practice, King's College London. Retrieved from <http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html>
- von Suchodoletz, D., & van der Hoeven, J. (2009). Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4(3), 146–155. <https://doi.org/10.2218/ijdc.v4i3.1188>
- Waters, D. J., & Cullyer, H. (2014, June 5). Annotating the Scholarly Web. Retrieved December 24, 2016, from <https://mellon.org/resources/news/articles/annotating-scholarly-web/>

