

Chapter 14

Blog Posts and Tweets: The Next Frontier for Grey Literature

Marcus Banks, UCSF Library, USA

14.1 Introduction

My interest in grey literature began as an Associate Fellow of the US National Library of Medicine (NLM) from 2002-2004. Many colleagues at NLM and throughout the country work to improve information access for people in the public health workforce. In comparison to resources available for practicing clinical medicine, the information needs for public health are more diffuse and often require access to grey literature [1]. In response to these needs NLM has developed the Partners in Information Access for the Public Health Workforce portal, which includes some avenues to grey literature but mostly useful links and validated search strategies for PubMed [2]. In addition to NLM's Partners page, staff of the New York Academy of Medicine Library has maintained the Grey Literature Report for several years [3]. This is a portal to documents produced by reputable organizations in public health and health policy.

During my NLM fellowship years, debate about open access publishing - specifically, how to secure access to publicly funded research - was heating up. An open access publication is generally a type of white literature, and is available for free online and stored in a digital repository [4]. Debate over the proper balance between open and subscription access will continue for years, as library associations and publishers continue to hire lobbyists and issue strongly worded statements. In 2004 I was certain that pure open access would prevail, but now think a hybrid subscription-OA model is much more likely to endure. During this more optimistic phase, I argued that grey literature advocates could learn from the political strength of open access advocates, and mount a similar campaign to demonstrate the value of grey literature [5]. Grey literature is almost always free to read already, so it only needed to be found.

Today I am much less concerned about findability for grey literature. While it remains simpler to locate a journal article than a working paper, smart Google searches can easily unearth the latter (I'll provide some examples of this in the next section.) Portals to grey literature remain useful for providing context and browsability; along with the Grey Literature Report, OpenSIGLE also serves both these functions well [6]. Even so,

the core challenge of finding grey literature in the first place is much less potent than in years past. So rather than mounting a political campaign to raise the profile of grey literature, I believe that grey literature advocates should now concern themselves with strategies for preserving the ephemeral “grey data” represented in content such as blog posts and tweets [7]. After addressing improved findability, I will present the case for my position.

14.2 Improved Findability for Grey Literature

Traditional barriers to locating grey literature, in comparison to white literature, include irregular publication schedules and the lack of standard bibliographic identifiers such as volumes, issues, and page numbers. These difficulties persist, but are much less fatal in the Internet age than they were in the print-only era.

It is possible to search Google for specific file types, and/or to restrict the search to the domains of organizations that produce a significant quantity of grey literature. Examples of each type of search, with screen shots, are found below (screen shots are current as of September 5, 2009.)

Insert Figure 1: Search for specific file type: “health policy filetype:pdf”

Insert Figure 2: Search within domain of specific organization: site:rwjf.org childhood obesity filetype:pdf (Restricts to searches of the Robert Wood Johnson Foundation domain)

These examples demonstrate how easy it is to locate grey literature, *as long as searchers know the appropriate search strings*. I recognize that this syntax is obscure to a very large majority of end users. However, librarians can instruct patrons about these search strategies, and also develop search systems that invoke these kinds of strategies behind the scenes.

Findability is not the only concern with grey literature. In the first example above the documents come from a large smattering of organizations, and would be of varying use depending on an individual’s information needs. In the second search, I needed to know in advance that the Robert Wood Johnson Foundation was a useful source of information. The advantage of OpenSIGLE and the Grey Literature Report, in comparison to “raw Google,” is that they both organize materials and vet their sources.

That said, findability *per se* is no longer a concern for traditional grey literature.

14.3 “Web 2.0” Content as “Grey Data”

If we accept the premise that findability is a less pressing concern than before, how should the grey literature community focus its energies today? I submit there is a pressing need to preserve the content being generated via various Web 2.0 tools and platforms.

First, some background: the somewhat annoying moniker “Web 2.0” describes the more interactive Internet that has emerged in recent years [8]. First generation web sites tended to be static HTML pages where readers could look but not touch, and the only way to share anything was to forward web links via email. Today it is extremely easy to post articles or news clips (say, from the *New York Times* or the BBC) to one’s Facebook page. (Facebook is just one of many “social networking” services; others include LinkedIn or MySpace.) This enhances sharing between friends and colleagues, which is useful. But the most profound change lies in the ability for anyone to post “user-generated content” such as blog posts or YouTube clips. Web 2.0 tools are also beginning to influence scientific debate [9].

Blogs are now an established part of the information landscape; they are scrolling public diaries that usually allow comments. Twitter is a more recent development. It is a “micro-blogging” service that enables users to post very short messages (no longer than 140 characters) via their web browsers or mobile phones. Each message is a “tweet,” and “retweeting” interesting messages has emerged as a rapid way to broadcast information [10].

I have maintained a blog—which contains ruminations on both professional and personal matters—since January 2005 [11]. Many of my colleagues in health sciences libraries also write blogs, often with a mix of personal and professional content [12, 13, 14]. In 2004 I did not read any blogs for professional information, and today they are critical professional sources. Blogs are updated much more rapidly than traditional journals, and readily facilitate conversation (although people often do not comment on posts.) Many librarians now tweet, sometimes to make pithy observations but often to share interesting blog posts or other online content [15, 16]. Within Facebook, the status update is functionally similar to a tweet.

The various Web 2.0 tools have enhanced both my professional and personal life. They enable a more fluid and informal form of communication, and in some form they are here to stay. (This Web 2.0 stuff could all be a fad, but at least within the library realm I think it’s unlikely that we’ll return to journal articles and white papers as dominant distribution mechanisms.) If this premise is correct, then a major concern about reliance on Web 2.0 tools is that we do not yet have good mechanisms for permanently archiving content produced with these tools.

Tweets are almost by definition ephemeral, and blog posts suffer from the general “link rot” that bedevils the Web [17]. Commercial solutions for Twitter archiving are emerging [18], and at least we have the Internet Archive [19]. But we are not yet close to an equivalent to acid-free paper for content developed online. Grey literature advocates can step into this breach, particularly if we expand the definition of grey literature to include the more informal “grey data” [20].

In the meantime, there are ad hoc efforts to preserve Web 2.0 content. For example, in the summer of 2009 the US Library of Congress announced plans to preserve tweets associated with Justice Sonia Sotomayor’s successful nomination to the Supreme Court [21]. An endorsement of the value of tweets (and by extension of blog posts, as many of the tweets referred to blogs commenting on the nomination) by an institution such as LC is a powerful indicator of their importance. Given LC’s decision, to close this chapter I

will provide further arguments for why the grey literature community should accept the challenge of preserving Web 2.0 content.

14.4 Case Study: Nicole Dettmar and Clinical Reader

Before making the argument for preservation of Web 2.0 content as grey literature or “data,” let’s examine the case of my librarian colleague Nicole Dettmar. Her experience in the summer of 2009 points to the vital need for permanent archiving of tweets in order to understand how people communicate online today.

Dettmar blogs, and in July 2009 she skillfully criticized the web site Clinical Reader for falsely implying that it had earned endorsements from leading libraries and for using copyrighted images without permission [22]. The critique was both accurate and thoughtful. Clinical Reader’s initial response, via Twitter, was to “kindly request” that Dettmar remove her blog post or else face the risk of legal action [23]. As of September, Clinical Reader has apologized to Dettmar and removed the implication of non-existent endorsements.

Insert Figure 3.

In the wake of Clinical Reader’s overwrought response to her post many librarians, as well as *Guardian* columnist Ben Goldacre, leapt to her defense. Much of the conversation between Clinical Reader and its critics took place on Twitter. But it was a hard conversation to have, because representatives from Clinical Reader consistently deleted their tweets and re-emerged with new Twitter accounts. Dettmar utilized screen shots to preserve the tweets for posterity [24], and the Disruptive Library Technology Jester also tracked the action [25]. Despite these laudable efforts, it would now be difficult for even the most intrepid scholar to piece together what happened on Twitter because of the deliberate interruptions to the flow of conversation. Below is a screen shot of a Clinical Reader Twitter screen that no longer exists, because the content was deliberately deleted.

If there is anyone who should care passionately about the preservation of otherwise overlooked and forgotten discourse, it is the grey literature community.

As Dettmar modestly stated, her initial post about Clinical Reader “was not of massive importance” [26]. But the support she received was gratifying and invigorating. Another important dimension of this support was its technical fluency. Dettmar’s supporters utilized Twitter with ease, responding to her critics within Twitter and retweeting in order to bring more attention to aspects of the discussion [27]. All of this happened in real-time, at a much faster pace of discourse than existed before the Web, or even in the Web 1.0 days. Tweets are early warning devices, and the blog posts or news articles they reference provide the context for whatever controversy is brewing.

Whether or not one has much interest in the Clinical Reader controversy, the phenomenon of the “rapid stream” of comments that registered the controversy should be of interest to the grey literature community. As I write, United States residents and

political leaders are debating whether and how to reform the health care system. Here too Twitter offers a vital register of the discussion. And just as with the Clinical Reader tweets, there is no guarantee that this record will persist. Unless, that is, members of the grey literature and broader information communities resolve to preserve this record and others like it.

14.5 Connection Between Web 2.0 Content and Grey Literature/Grey Data

My initial interests in grey literature stemmed from admiration at the political savvy of open access advocates, with hopes that the GL community could learn from the open access movement and raise the profile of grey literature [28]. Now that findability for grey lit is not as large a concern, my interests have shifted to the idea of a “continuum” that will eventually collapse the distinction between grey and white literature [29].

I once assumed that this continuum only included materials that could easily be printed. My idea was that peer review could often happen online rather than behind the scenes, and I conceived of blog posts as the most radical extension of traditional forms of communication. Everything I envisioned would properly fall under the heading of “literature.”

As tweets can be no longer than 140 characters, it is a stretch to call them literature. But they are definitely useful bits of data, which collectively can aggregate into an important lens of understanding an ongoing discussion. In 2009 the lonely tweet suffers from the same findability problems as the working paper did pre-Google. Twitter facilitates searching for groups of tweets via the “hash tag” convention (example: #Obama), but the individual tweets buried within conversations can be easily lost [30].

If we conceive of the tweets as “grey data,” then the preservation imperative for the grey literature community becomes clearer. Robust archiving services for Twitter would have prevented Clinical Reader from manipulating the Twitter conversation this year.

Blog posts leave a larger footprint than tweets, although they still ephemeral and subject to the usual Internet link rot. They are also “grey” in the sense of not containing standard bibliographic identifiers. Although tweets are particularly at risk because they are so easy to proliferate, blog posts would also be worthy of the preservation attention of the grey literature community.

This begs a whole host of questions. With literally millions of bloggers of Twitterers out there, how can anyone possibly determine what to preserve? The Library of Congress’s decision regarding the Sotomayor tweets was an easy call, but this won’t always be the case. While I recognize the enormity of the challenge, the developers of the New York Academy of Medicine and OpenSIGLE portals also had to establish selection criteria (albeit on a much smaller scale). General digital preservation principles, combined with an evolving understanding of the uses of Twitter, would be necessary in developing preservation criteria for blogs and tweets. I have no words of wisdom in this regard, except to say that this is clearly an area of growth for the grey

literature community and that I would be happy to be part of any discussions in these areas.

14.6 Conclusion

My interests in grey literature have shifted considerably in the last five years, which is roughly synonymous with the emergence of “Web 2.0” tools. I almost feel like a fraud writing this chapter, because my interests are so divergent from what I presented at conferences in Nancy and New Orleans! But at the risk of seeming fraudulent, I really do believe that preservation of Web 2.0 content should be a main focus for the GL community in future years. Traditional grey literature remains important, but thankfully it is much easier to find than before. Let us now turn our attention to new and exciting challenges.

References

1. Alpi K.M. (2005), Expert searching in public health. – In: *J Med Libr Assoc*, 93 (1) Jan 2005, pp. 97-103.
<http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=545128&blobtype=pdf>
2. NLM, Partners in Information Access for the Public Health Workforce, <http://phpartners.org/>. [cited 7 Sep 2009].
3. NYAM Library, Grey Literature Report, <http://www.nyam.org/library/greyreport.shtml>. [cited 7 Sep 2009].
4. Suber P. (2003), Bethesda statement on open access publishing, <http://www.earlham.edu/~peters/fos/bethesda.htm>. [cited 7 Sep 2009].
5. Banks M. (2004), Connections between open access publishing and access to gray literature. – In: *J Med Libr Assoc*, 92 (2) April 2004, pp. 164-166.
<http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=385294&action=stream&blobtype=pdf> [cited 7 Sep 2009].
6. INIST, OpenSIGLE, <http://opensigle.inist.fr/>. [cited 7 Sep 2009].
7. Banks M. (2009), Open access, grey literature, grey data, http://mbanks.typepad.com/my_weblog/2009/07/open-access-grey-literature-grey-data.html [cited 7 Sep 2009].
8. O'Reilly T. (2005), What is Web 2.0?, <http://oreilly.com/web2/archive/what-is-web-20.html>. [cited 7 Sep 2009].
9. Economist (2008), User-generated science, http://www.economist.com/sciencetechnology/displayStory.cfm?STORY_ID=12253189 [cited 7 Sep 2009].
10. Blogging Bits (2008), How to retweet: a simple guide, <http://bloggingbits.com/the-art-and-science-of-retweeting-for-twitteraholics/>. [cited 7 Sep 2009].
11. Banks M., Marcus' World, <http://mbanks.typepad.com/> [cited 7 Sep 2009].

12. Kraft M., Krafty Librarian, <http://kraftylibrarian.com/> [cited 7 Sep 2009].
13. Plutchak T.S., T. Scott, <http://tscott.typepad.com/>. [cited 7 Sep 2009].
14. Rothman D., Exploring medical librarianship and web geekery, <http://davidrothman.net/>. [cited 7 Sep 2009].
15. Rumsey E., Eric Rumsey on Twitter, <http://twitter.com/ericrumsey>. [cited 7 Sep 2009].
16. Walden R., Rachel Walden on Twitter, http://twitter.com/Rachel_w. [cited 7 Sep 2009].
17. Wikipedia (2009), Link rot, http://en.wikipedia.org/wiki/Link_rot. [cited 7 Sep 2009].
18. QuoteURL (2009), <http://www.quoteurl.com/>. [cited 7 Sep 2009].
19. Internet Archive (2009), <http://www.archive.org/index.php>. [cited 7 Sep 2009].
20. Banks M. (2009), Open access, grey literature, grey data, http://mbanks.typepad.com/my_weblog/2009/07/open-access-grey-literature-grey-data.html. [cited 7 Sep 2009].
21. Library of Congress (2009), LOC will capture tweets on [#sotomayor](#) for its web archives on the Sotomayor nomination, <http://twitter.com/librarycongress/statuses/2222610835>. [cited 7 Sep 2009].
22. Dettmar N. (2009), Clinical Reader: Starry ethics fail, <http://eagledawg.blogspot.com/2009/07/clinical-reader-starry-ethics-fail.html> [cited 7 Sep 2009].
23. Lawson S. (2009), Clinical Reader: From zero to negative sixty with one bogus threat, http://stevelawson.name/seealso/archives/2009/07/clinical_reader_from_zero_to_negative_sixty_with_one_bogus_threat.html [cited 7 Sep 2009].
24. Dettmar N. (2009), Twitter search on “legal ramifications,” July 16, ’09 on Twitpic, <http://www.twitpic.com/al4eg> [cited 7 Sep 2009].
25. Disruptive Library Technology Jester (2009), On the pitfalls of social media: The case of Clinical Reader, <http://dltj.org/article/clinical-reader-background/> [cited 7 Sep 2009].
26. Dettmar N. (2009), Gratitude, <http://eagledawg.blogspot.com/2009/07/gratitude.html> [cited 7 Sep 2009].
27. Lozeau P.M. (2009), RT @adr Dear @ClinicalReader: Stop threatening bloggers with bogus legal claims. OK? Thanks. (i.e <http://ur1.ca/78oc>). [cited 7 Sep 2009].
28. Banks M. (2004), Connections between open access publishing and access to gray literature. – In: J Med Libr Assoc, 92 (2) April 2004, pp. 164-166. <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=385294&action=stream&blobtype=pdf> [cited 7 Sep 2009].
29. Banks M. A. (2005), Towards a continuum of scholarship: the eventual collapse of the distinction between grey and non-grey literature. – In: GL7 : Seventh International Conference on Grey Literature, Nancy (France), December 2005. <http://eprints.rclis.org/5803/> [cited 7 Sep 2009].

30. Tech for Luddites (2009), The Twitter hash tag: what is it and how do you use it?, <http://www.techforluddites.com/2009/02/the-twitter-hash-tag-what-is-it-and-how-do-you-use-it.html> [cited 7 Sep 2009].