Chapter 1

Networking the News

In the previous chapters, I have outlined the ways in which archives, and critical readings of them, have expanded from a fixed and graspable entity to a suite of interconnected parts, constantly shifting and adapting to new information. The web, when seen as an archive of archives, is itself "an active and evolving repository of knowledge," rather than a fixed, bordered entity or set of categories.[1] This chapter hones in specifically on the structure of news stories and publishing archives, and the ways online publishers and legacy news outlets are treating their digital and digitized archives in this new era of continuous reclassification.

Although news has been called "a first rough draft of history," in newsrooms, the archive is traditionally known as "the morgue": a place where stories go to die. But new technologies and conditions have led to many recent attempts to reanimate the news archive, and there seems to be an "archive fever" developing amongst news publishers. Nicole Levy wondered if 2014 is "the year of the legacy media archive" in a story about *Time* magazine's archival "Vault."[2] She points to *The Nation*'s "back issues," *The New Yorker*'s open archive collections, and the *New York Times*' TimesMachine and @NYTArchives Twitter account as examples of old publishers endeav-

1. Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data* (Morgan Kaufmann, 2003), 2.

2. Nicole Levy, "Time.com opens its 'Vault' | Capital New York," Capital New York, November 12, 2014, accessed November 12, 2014, `http://www.capitalnewyork.com/article/media/2014/11/8556503/timecom-opens-its-vault`.

oring to use their rich histories to create something new. Back archives like Harper's and the National Geographic are held up as examples of combining rich content with historical context, improving credibility and brand recognition in the process.

The Times closely examined its own archives in their celebrated *Innovation* report of 2014, suggesting that a clever use of archives could revitalize new content by seamlessly integrating with historical context. "Our rich archive offers one of our clearest advantages over new competitors... [b]ut we rarely think to mine our archive, largely because we are so focused on news and new features," arguing that "we can be both a daily newsletter and a library."[3] The report suggests that arts and culture content, more likely to be evergreen, could be organized "more by relevance than by publication date," and the topic homepages should be more like guides than wires.[4] The report goes on to enumerate successful experiments with repackaging old content in collections, organized by categories and themes. They suggest allowing users to create their collections of stories– something that readers could also do without risk to the Times brand. By creating "no new articles, only new packaging," the Times can easily give new life to old stories.[5]

In 2014 we also saw another trend towards "explainer journalism," and an intense focus on context provision for readers. Vox.com, the poster child for the explainer movement, aims "to create a site that's as good at explaining the world as it is at reporting on it."[6] Explainer journalism aims to take a step back from the immediate news event and place it in a larger phenomenon, and it reflects a deep shift in the roles and practices of online journalists; as news is increasingly broken and scooped on social media, journalists are increasingly becoming summarizers, filterers, and context providers. News has traditionally been delivered in a stream format, full of boilerplate text that is repeated across every story related to a given theme. In the archive and explainer movements, we see a pattern among some news outlets attempting to evade

3. *Innovation* (New York Times, March 24, 2014), 28, accessed January 24, 2015, `https://www.scribd.com/doc/224332847/NYT-Innovation-Report-2014`.

4. Ibid., 29-30.

5. Ibid., 34.

6. Ezra Klein, "Vox is our next," The Verge, January 26, 2014, accessed December 23, 2014, `http://www.theverge.com/2014/1/26/5348212/ezra-klein-vox-is-our-next`.

and reconsider the news cycle's obsession with speed and feeds, instead experimenting with new forms of what a news story can be, and how it can connect to other stories within the archive and around the web.

As many publishers emphasize the potential value of archives and context for the future of digital journalism, this moment is rich for closely examining this connection. By looking at the challenges and methods in digitizing and structuring legacy media archives, we can gain a sense of how news stories are structured on a small scale, and how a collection of them creates context on a larger scale. This lets us think closely about the structure of online content, and the ways that news publishers can continuously keep their archives relevant and context at hand.

In this chapter I will start by examining the structure of news stories to grasp at the challenges specific to news organizations when archiving. I will then outline the stages of "linking the archive" for news publishers, and consider current efforts in legacy news and digital preservation. Finally, I will discuss the relationship between archives and explainer journalism, examining the convergence of the roles of journalist and developer in the newsroom and the corresponding shift in the function and practices of journalists.

1.1 The structure of stories

Newspaper and magazine publishers prove an ideal study for examining the potentials of hypertext archives. If we treat a newspaper as a proto-hypertextual document, it becomes apparent that online news might be a natural extension of reading the newspaper. Few readers go through a newspaper sequentially, paying equal attention to every article; instead the reader jumps around from page to page, skimming some sections for its raw information while reading longer pieces more deeply. A website homepage reads like a newspaper's front page, with snippets and teasers that aim to draw the reader deeper. A given page can hold several articles, and an interested reader might be distracted or intrigued by a "related article" next to the one he or she came to read. Some works are categorized into sections—arts, sports, letters to the

editor—while others might be paired with a certain advertisement or reaction article. These examples point to the inherently interlinked nature of newspapers, and the endless potential for insightful metadata; newspapers might seem to naturally lend themselves to the digital world.

The pre-hypertextual newspaper started as a response to a sort of historical information overload; the newspaper frontpage and summary lead paragraph, both solidified in 1870, were part of a broader trend towards "helping readers to economize their scarce time in scanning a paper."[7] Larger type, illustrations, and bolder headlines drew criticism for trying to grab attention, but they also directed and focused attention to the major stories of the day, allowing for nonlinear readings of a newspaper as a fragmented collection of stories, headlines, or leads. A newspaper's layout and seriality therefore scaffold a pseudo-hypertextual structure, one that can be computationally mined for insights. Some libraries and cultural heritage institutions are leading these endeavors, such as Europeana and Trove.[8]

But traditional newspapers have a major limitation: they cannot *explicitly* link to other work in a structured and idiomatic way. Scholars have long relied on the footnote and bibliography to systematically track influence and dialogue, and networks of citations can be created out of them. This forms the basis for citation analysis, or bibliometry, a practice with a long history and strong conventions that I will dive into more closely in the following chapter. Its essential principle is that the more an item is cited, the more influential and credible it is. The online version is known as "webometrics," and it applies certain new standards which online newspapers can take advantage of, both in measuring impact on the web and inside their own archives. But citation is "as old as written language itself," and it is *itself* a language, with its own idioms, syntaxes and exceptions.[9] The footnote has its limitations, only linking back to the past—but newspapers don't even get footnotes.

The journalistic affordances that the web brings can be conceptually divided into

7. Paul Starr, *The Creation of the Media: Political Origins of Modern Communications* (Basic Books, 2004), 254.

8. **europeana**; **trove**.

9. Chakrabarti, *Mining the Web*, 1.

a few core features, which Mark Deuze outlines as hypertextuality, multimediality, and interactivity.[10] Peter Dahlgren adds a fourth and fifth: for him, media logic is also *figurational* and *archival*. Discussing archivality, he asserts that "users of cyberspace for journalism are in principle no longer so bound to the present," and points to hypertextuality as enabling new, more usable archives.[11] While many projects have closely mapped the role of hypertextuality in online journalism—examining newsrooms' approaches towards hyperlinking through network analysis, surveys, interviews, and newsroom ethnography—there has been less research considering hypertextuality's influence on newsrooms' archival practices.

Hyperlinks allow for a new standard of citation, reference, and context provision for news. The link can even go beyond the footnote by linking in both directions, allowing readers to see who referenced the story; an old article in *The New York Times*, for instance, can link out to more recent related Times articles, other publishers or blogs that picked up on the story, or conversations in the Times forum or on Twitter. Linking offers great potential, not only for enlivening the reading experience, but for creating a traceable dialogue that can improve a story's discoverability in the future. A number of search algorithms, such as Google's PageRank and Jon Kleinberg's HITS system create "hyperlink-induced communities,"[12] between websites, and the same principles can be adopted an expanded *within* news websites.

A human editor who is tagging a story is equivalent to the archivist in the library, attempting to predict every possible way that a user might search for the story in the future, whether it's "Sports" or "Breaking" or "Opinion"—and editors don't have the extensive training and professional expertise that comes with being a librarian or archivist. Journalists are trained to explain, contextualize, and curate rather than structure and tag. Given the impossibility of explicitly and expertly tagging in advance for every possible present and future use, as well as the arbitrariness of

10. MArk Deuze, "The Web and its Journalisms: Considering the Consequences of Different Types of Newsmedia Online," *New Media & Society* 5, no. 2 (June 1, 2003): 203–230, accessed February 8, 2015, `http://nms.sagepub.com.libproxy.mit.edu/content/5/2/203`.

11. Peter Dahlgren, "Media Logic in Cyberspace: Repositioning Journalism and its Publics," *Javnost - The Public* 3, no. 3 (January 1, 1996): 66, accessed March 9, 2015, `http://dx.doi.org/10.1080/13183222.1996.11008632`.

12. Chakrabarti, *Mining the Web*, 12.

tagging *the story* instead of its constituent parts, we can turn to entities and links as supplements to categories and tags in the newsroom archive.

1.1.1 Units of news

In technical terms, stories are usually objects in a database that have associated text, images and tags. Stories contain multitudes, and a typical story might have a variety of metadata attached to it; authors, dates, versions, categories, images, events and collections it's a part of, tags, and so on.[13] While more metadata and structure requires more investment at the outset, smart use of such metadata prepares a story for archival reuse. Some of it will be useful for linking or embedding on the website, others for use in an API or application. Stories can include other stories as part of their metadata too, either related manually (by a hyperlink or a human editor) or automatically (via similarity algorithms that analyze the words or topics in the article, the communities it reaches, and so on).

The story has long been the basic unit of news, and so it tends to have a one-to-one relationship with the URL, the basic unit of the web. One section of the *Times' Innovation* report announces that they produce "more than 300 URLs a day," using URL as a sort of "thing" word, their default unit of work.[14] Most publishers will assign a "canonical URL" to a given story, which serves as its unique identifier, and often, practically speaking, it is the only information that a researcher or search engine can feasibly obtain about a particular document on the web.[15] You can be sure to find the most canonical version at the canonical URL, but the article lives in various forms across the web.

But if stories contain multitudes, then why is the article the basic unit of information for news? An article can pull paragraphs from one source, photos and charts from another. It is an ecosystem of media itself, and it can contain other stories in

13. Figure here with typical story objects; a sample database schema
14. *Innovation*, 27.
15. While the researcher or bot crawler could, of course, request the webpage to get the information, each request can take several seconds and some processing power, so it becomes infeasible at a larger scale.

turn. The news app Circa organizes its content around "atoms" of news: single facts, quotations, statistics, and images that can be reaggregated and remixed as needed. Systems like Vox and Circa aim to create a baseline repository to build upon rather than recreate from scratch every time.

This approach rethinks how we organize news items and structure stories. A "story" can be a collection or dialogue of items; indeed, most stories already are. A journalist can still create, but also curate, collect, and contextualize, or allow users to do the same. All of these remixes and reuses can improve the classification and discoverability of the content in turn. Thinking of a story as a collection or mash-up offers a new framework of a story as a highly linked entity, one that can start to organize itself.

As discussed in previous chapters, organizing by link and tag has often proven a more effective form of sorting things out on the web than has organizing by overarching taxonomy or ontology. It has been the lifeblood of Google, as its PageRank algorithm made it the dominant search engine over rivals like Yahoo! and HotBot.[16] Yahoo! began in 1994 as a hierarchical directory of useful websites. This is a natural first step for an online search engine, since computer users have grown accustomed to the tree-like document and file structure pioneered by Douglas Engelbart and others, and replicated by Berners-Lee's domains and paths on the web. It also builds relationships between categories into its structure—parents, children, and siblings—which readily enables features like "More like this."

But Google succeeded by crawling in the weeds rather than commanding from on high. For Google, the links sort everything out. Tim Berners-Lee proved that networks could work with many links—and in fact, if you had a lot of links, as Clay Shirky argues, "you don't need the hierarchy anymore. There is no shelf. There is no file system. The links alone are enough."[17]

Shirky and David Weinberger champion the tag as a hybrid hierarchical/networked organizational structure. On one hand, tagging relies on an individual singularly clas-

16. Clay Shirky, "Ontology is Overrated: Categories, Links, and Tags," 2005, `http://www.shirky.com/writings/ontology_overrated.html`.
17. Ibid.

sifying an object under a certain discourse. On the other hand, users are generally free to tag as many times as they want, and using whatever scheme they desire. Tags could range from "World War II" to "articles I want to read." Studies and businesses alike have proven that at web scale, even with users tagging items for personal and idiosyncratic reasons, distinct and simple patterns emerge that allow for collaborative classification.[18] Such collaborative classification systems, sometimes called "folksonomies," emerge as manifestations of the "boundary infrastructures" proposed by Bowker and Star in *Sorting Things Out*.[19]

Tags have their limitations; if another user tags an item "World War 2," the system needs to recognize that it means the same thing as "World War II," and publishers employ controlled vocabularies to avoid such ambiguities. Some research has shown that the first tags on an item are likely to influence future tags in turn, resulting in a sort of ontological groupthink.[20] Still, whether a Flickr tag, a Delicious bookmark, or a Twitter hashtag, these crowdsourced approaches to tagging function as links between content; it is not about the tag itself, but the *connection* being made to other content. Shirky even considers the possibility that syonymous terms aren't desirable; perhaps the people who are searching for "films" would be better served by just seeing results tagged as "films," and not "movies" or "cinema."[21] This suggests an even more direct reliance on language to sort things out. Still, sometimes hierarchies are crucial; a user searching for movies set in Massachusetts would also want movies tagged "Boston."

The New York Times sees tagging as core to its business, calling it the primary reason that the Times has remained the "paper of record" for decades.[22] This title was bestowed to them largely on the back of the legacy Times Index, which has offered an annual reference version of Times stories since 1913, still published in hard copy. There is no doubt that the Times' tagging practices have helped them remain

18. Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero, "Semiotic dynamics and collaborative tagging," *Proceedings of the National Academy of Sciences* 104, no. 5 (January 30, 2007): 1461–1464, accessed February 8, 2015, `http://www.pnas.org/content/104/5/1461`.

19. Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT Press, August 28, 2000).

20. Cattuto, Loreto, and Pietronero, "Semiotic dynamics and collaborative tagging."

21. Shirky, "Ontology is Overrated: Categories, Links, and Tags."

22. *Innovation*, 41.

a library, information hub, and general authority on contextual information. But the *Innovation* Report sees them falling behind, adhering too much to the needs of the hard copy Index. They also note crucial limitations with identifying, splitting, and lumping categories; it took seven years for the Times to start tagging stories "September 11." Their team of librarians is required to shepherd about 300 new articles a day into the archive, making sure to keep them discoverable under any context. Tags can concern more than just the contents or event of a story—the Report suggests tagging stories by timeliness, story tone, and larger "story threads"—but they admit that many of the more exciting tagging potentials would require them to "better organize our archives."[23]

So the linking of the archive can occur not only through explicit hyperlinks, but implicit tags and entities that reside within the stories themselves. Most news articles rely on tagging to be connected to other media; if *The Boston Globe* writes a story about New England Patriots coach Bill Belichick, an editor might tag the story "football" in order to place it in dialogue with other football stories, landing on the Globe's football topic page, and so on. But this belies a more nuanced dialogue between the words, images, and hyperlinks used within the story itself; a short story that has "Bill Belichick" and "Cincinnati Bengals" in the text is likely to be referencing a recent game or a trade, while a longer story that brings up his family members or his hometown is likely to be a biographical piece about his life and upbringing. By combining natural language processing and entity linking tools, a story can be automatically and dynamically tagged according to the myriad possible contexts that a user might search for, whether she is looking for stories about football, the Patriots, or Bill Belichick himself.[24]

23. *Innovation*, 41-42.
24. explainer on NLP and linking?

1.1.2 From topics to tags to links

While tagging practices can be enhanced in a variety of ways, news developer, researcher, and blogger Stijn Debrouwere thinks that even "tags don't cut it."[25] As an expert in news analytics and a co-creator of link-tracking platform NewsLynx, he knows well the limitations of the web and newsrooms' content management systems. His blog series "Information architecture for news websites" dives into the headaches that result when journalists think of stories as blobs of text in feeds and streams, rather than structured systems of facts that carry value in their own right.[26]

He cites a famous blog post by 2006 by Adrian Holovaty, co-creator of the Django web framework and its now-retired "benevolent dictator for life." Holovaty's essay revolves around one idea: "newspapers need to stop the story-centric worldview."[27] Each story, he notes, contains a vast amount of structure that is being thrown away with every click of the "publish" button. He leads through several examples, such as:

- An obituary is about a *person*, involves *dates* and *funeral homes*.
- A *birth* has parents, a child (or children) and a date.
- A *college graduate* has a *home state*, a *home town*, a *degree*, a *major* and *graduation year*.
- A drink special has a *day of the week* and is offered at a *bar*.
- A *political advertisement* has a *candidate*, a *state*, a *political party*, multiple *issues*, *characters*, *cues*, *music* and more.

Holovaty links to context everywhere above, using hyperlinks to literally highlight the information that's otherwise locked away behind stories. Of course we don't need all of this context all the time, but we may really need *some* of the context *sometime*,

25. Stijn Debrouwere, "Tags don't cut it," stdout.be, April 6, 2010, accessed September 15, 2014, `http://stdout.be/2010/04/07/tags-dont-cut-it/`.

26. Stijn Debrouwere, "Information architecture for news websites," stdout.be, April 5, 2010, accessed March 8, 2015, `http://stdout.be/2010/04/06/information-architecture-for-news-websites/`.

27. Adrian Holovaty, "A fundamental way newspaper sites need to change," Holovaty.com, September 6, 2006, accessed March 8, 2015, `http://www.holovaty.com/writing/fundamental-change/`.

and it's easier to structure it now than to unlock it later. The better structured this information, Holovaty argues, the more serendipity can foster new features and applications. Proper story scaffolding can lead to more happy accidents of "wouldn't it be cool if…" later. Want to map the births and deaths of a famous family, or the happy hours in a neighborhood? You might already have that information buried in your stories.

Debrouwere expands on Holovaty, summarizing his frustration with tags: "each story could function as part of a web of knowledge around a certain topic, but it doesn't." Tags are our only window into content at the level of a story's metadata (which, too often, is all we have). For all their weblike strengths, they are still often inconsistent, outdated, and stale. "The whole purpose of tags is to relate one piece of content to another," and given the dozens of ways that one can type "George Bush," they can't even do that.

Debrouwere concludes that we need "a way of indicating how content relates to other content on our website and on other websites that is more powerful and more expressive than tags." He suggests using vocabularies: set people, places, organizations, events and themes. Knowledge bases like DBpedia, OpenCalais, OpenNLP, AlchemyAPI, or the Getty Vocabularies allow for deep context at low cost, basing its tagging on "entities, not labels." He also advocates for indexing relationships rather than contents, which borrows from Semantic Web principles to add detail to a link. "A tag on an article says 'this article has something to do with this concept or thing.' But what exactly?" Rather than tagging an article "Rupert Murdoch," a tag has more value if it can say "criticizes Rupert Murdoch." For Debrouwere, "we don't need the arbitrary distinction between a label and the thing it labels on a website. Let's unlock the full potential of our relationships by making them relationships between things."

Such a scheme could benefit an end user in many ways. Topic pages, such as New York Times' over 5000 pages ranging from "A.C. Milan" to "Zimbabwe," could be smarter, reflecting the most popular articles or most related topics. Entity- and link-oriented schemes can create cascades of relationships, synonyms, and homonyms.

Journalists as well as readers would gain improved access to their organization's history, improving the research, context, and tagging of future stories. Debrouwere is suggesting a return of structure to the open web; he envisions a tagging system where a tag can double as a card or widget, linked in turn to other cards and widgets in a network of knowledge. This could be extended to events and phenomena as well as proper names and entities; some emerging systems can recognize phrases like "Barack Obama announced his candidacy for president" and ground it as as a unique, unambiguous newsworthy event.[28] While this research has a long way to go, it is a promising start towards extending the "ankle-deep semantics" that Chakrabarti advocates.[29]

This return of structure does not have to be a step backwards; instead of manually and unilaterally structuring from on high, we can focus on the structure built into the stories already. This solution doesn't replace stories, or require editors to exhaustively tag every component of a piece; it just automatically supplements a story with new metadata that gives its inherent information an afterlife. Every story—whether a blog post, a map, a listicle, or an interview—has its own structures and patterns.

One example comes from the Boston Globe's March 2015 coverage of the Boston Marathon bombing trial. Data editor Laura Amico knew well that trials are far from linear stories; since they are told by lawyers, they are a series of conflicting arguments. Trials are unique in providing two stories: the chronological narrative of the trial repieces the narrative of the original event. Amico and her team knew there was more than one way to tell this story, so they decided to focus on the facts; the witnesses, exhibits, and arguments are all entered into a spreadsheet, which can generate snippets of facts and entities—designed as cards—for later review.[30] Each of these cards can embed and contain other cards, or be combined to form a story. Such a framework recognizes the interlinked nature of stories and takes advantage of

28. Joel Nothman, "Grounding event references in news" (August 31, 2013), accessed February 8, 2015, `http://ses.library.usyd.edu.au:80/handle/2123/10609`.

29. Chakrabarti, *Mining the Web*, 289.

30. Benjamin Mullin, "How The Boston Globe is covering the Boston Marathon bombing trial," March 8, 2015, accessed March 9, 2015, `http://www.poynter.org/news/mediawire/325301/how-the-boston-globe-is-covering-the-boston-marathon-bombing-trial/`.

the structure of an event (in this case, a criminal trial).

One of the most obvious, low-hanging and underexplored structures is the hyperlink. A year after publishing his "Information architecture" series, Debrouwere followed up by questioning many of his own allegiances; tags still don't cut it, but maybe taxonomies don't either. He realized: "The best content recommendations on news websites are inside of the body copy: inline links. With recommendations, you never know what you're getting. It's mystery meat. With links, a writer tells you why she's pointing at something the moment she's pointing at it." It is better to draw on the connections from these links than to rely on automated recommendation engines to organize content. Journalists are better at explaining and contextualizing than they are at tagging and structuring, which are a librarian's craft. Debrouwere knows that newsroom developers are building for journalists, and he ends by asserting that he wants to build "prosthetics, not machines."[31]

Another under-mined source of insight lies in the plethora of "human-generated lists," as Google calls them, around the web.[32] Whether collecting articles, photos, books, songs, or tweets, people obsessively collect and curate, and some are known experts at doing so. These range from Amazon wish lists to mixed-media stories on Storify. Thinking of lists as links between contents, weighted by expertise, leads to interesting potentials. The title of the list, or its other metadata, could tell us more about the context behind the link; a list of local Mexican restaurants is linked by country and location, while a list of my favorite hip-hop albums of 2014 is linked by year, quality, and musical genre. The Times' *Innovation* report suggests allowing users to create lists, since it could allow for deep interactivity without risk to their brand; such a system could use readers' collective wisdom by asking users to specify the context behind their lists.

These web-native and polyhierarchical approaches to classification reflect the grow-

31. **debrouwere_taxonomies_2011**.

32. Discovering and scoring relationships extracted from human generated lists, U.S. Classification 707/765, 707/803; International Classification G06F17/30, G06F7/06; Cooperative Classification G06F17/30053, G06F17/30657, G06F17/30867, G06F17/30663; European Classification G06F17/30E4P (US8108417 B2, filed January 31, 2012), accessed March 9, 2015, `http://www.google.com/patents/US8108417`.

ing need for newsrooms to find weblike ways to organize their content; automatic and dynamic tagging, linked entites, image recognition, and tapping into the knowledge of experts and crowds are some of the solutions. But despite the news story's rigid structure and the limitations of indexing, there is no reason to believe that the article is going away as the core unit of news. Link-based platforms and services like RSS and social media feeds still rely on stable and consistent resources at given URLs. Innovations in story structure could even be at odds with the very notion of revitalizing the archive. In aiming to blend the present with the past, archive-oriented publishers are bringing past conventions back into the present. Still, some new forays into interactive, multimedia, and app-driven journalism enhance or bypass the URL and hyperlink—I will touch on these at greater length in the conclusion. My aim is not to suggest that we restructure the news story; only that we rethink how they work under the hood. Stories are not uniform resources, and they should not be uniformly tagged and categorized.

1.2 Stages of digital history

In 1997, John Pavlik suggested that there were three stages of development in online versions of newspapers. Keeping in mind this early date, Pavlik observed that newspapers' first stage online was to copy the print edition to a given website (a stage known and maligned as "shovelware"), followed by supplementing the copy with interactive features (like hyperlinks or comments), then in the third and final stage, writing copy specifically for the online version of the story.[33] Twenty years later, it is safe to say that publishers have all reached the final stage. I aim to suggest a three-stage development of my own in the state of the digital legacy news archive, referring to each stage respectively as *digitizing*, *atomizing*, and *linking*.

The first stage for any legacy publisher is to ***digitize*** the archive. This tends to consist of scanning the pages of old publications, running OCR (optical character

33. John V. Pavlik, "The Future of Online Journalism," *Columbia Journalism Review* 36, no. 2 (August 7, 1997): 30–36, accessed February 8, 2015, `http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=9709105522&site=eds-live`.

recognition) on each page, and exposing the results to a search interface for researchers and, perhaps, interested readers. It is a crucial first step for enlivening the archive, but a physical record can often limit the digital equivalent's potentials. Digital versions of physical articles often do not leverage links, images, and mixed media to the same effect. While a digital-native version of a print article might directly cite more sources or feature an intriguing interactive, these elements remain second-class citizens to the print article, which digital versions must remain faithful to. The Times' *Innovation* Report argues that by modeling their website and apps on their print structure, the Times "ask[s] too much of readers."[34] So it is crucial to remember at this stage that the digital archive has different potential from its physical counterpart. As many theorists and historians remind us, too, a paper's physical appearance and content are closely linked together, so simply "digitizing" and newspaper changes it massively, reshaping a great deal of context.[35] Richard Abel breaks down the promises and challenges in generating archival "big data" in research on 1910 US cinema. Using digital newspapers led to "a wealth of unexpected documents," but he notes the unreliability of completeness and searchability, and the collapse of community.[36]

Given the print newspaper's proto-hypertextual status, it presents a unique metadata challenge for archivists. Paul Gooding, a researcher at University College London, sees digitized newspapers as ripe for analysis due to their irregular size and their seriality.[37] In order to learn more about how people use digitized newspaper archives, Gooding analyzed user web logs from Welsh Newspapers Online, a newspaper portal maintained by the National Library of Wales, hoping to gain insight from users' behavior. He found that most researchers were not closely reading the newspapers

34. *Innovation*, 26.

35. James Mussell, "Elemental Forms," *Media History* 20, no. 1 (January 2014): 388-389; Marlene Manoff, "Archive and Database as Metaphor: Theorizing the Historical Record," *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398, accessed December 2, 2013, `http://muse.jhu.edu.libproxy.mit.edu/login?auth=0&type=summary&url=/journals/portal_libraries_and_the_academy/v010/10.4.manoff.html`.

36. Richard Abel, "The Pleasures and Perils of Big Data in Digitized Newspapers," *Film History* 25, no. 1 (January 2013): 1–10, accessed March 9, 2015, `http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=88234573&site=ehost-live`.

37. Paul Gooding, "Exploring Usage of Digitised Newspaper Archives through Web Log Analy…" (DH2014, Lausanne, Switzerland, July 9, 2014), accessed March 9, 2015, `http://www.slideshare.net/pmgooding/dh2014-pres`.

page by page, but instead searching and browsing at a high level before diving into particular pages. He sees this behavior as an accelerated version of the way people browse through physical archives—when faced with boxes of archived newspapers, most researchers do not flip through pages, but instead skip through reams of them before delving in. So while digital newspapers do not replace the physical archive, they do mostly mimic the physical experience of diving into an archive. Still, something is lost when the physical copy becomes digital; the grain of history—the old rip, annotation, or coffee stain—is reduced to information.

It might go without saying, but it is also crucial to back up the archive. More and more evidence shows that digital content could have a shorter shelf-life than tapes, film, or other analog media.[38] It is important to diversify the format of the archived material as well. The Missouri School of Journalism's *Columbia Missourian*, lost 15 years of stories and seven years of images in a single server crash.[39] For Missouri's Reynolds Journalism Institute, "News archives serve as a form of institutional knowledge allowing newer staff members to understand and convey the historical context of their stories to the readership. It is difficult to calculate the full value of news archives given the countless hours of reporting and editing they distill, not to mention the treasure they represent in terms of their community's cultural heritage."[40] Obsolete backup formats only compound the problem. Some experts are pessimistic, suggesting that the sheer cost of maintaining the records is not worth the benefit.

In response to the server crash, the Reynolds Institute led a survey of 476 news websites, finding that found that 88–93 percent of them highly value their archives, but about a quarter of them had lost significant portions of their archive due to

38. Ian Sample, Science Editor, and in San Jose, "Google boss warns of 'forgotten century' with email and photos at risk," the Guardian, February 13, 2015, accessed March 9, 2015, `http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf`.

39. Lene Sillesen, "Minus proper archives, news outlets risk losing years of backstories forever," Columbia Journalism Review, July 21, 2014, accessed July 24, 2014, `http://www.cjr.org/behind_the_news/minus_proper_archives_many_new.php`.

40. Edward McCain, "Saving the news: When your server crashes, you could lose decades of digital news content ¯ forever | RJI," Reynolds Journalism Institute, July 16, 2014, accessed March 9, 2015, `http://www.rjionline.org/blog/saving-news-when-your-server-crashes-you-could-lose-decades-digital-news-content-forever`.

technical failure. Without these full archives, as Tom Warhover says, "You can't offer up a comprehensive product to sell—your archives—if they aren't complete. You can't be sure you've really vetted a candidate for school board or city council. You can't find those historical pieces from events that now are historic and worth reporting again on anniversaries."[41] Those publishers who have digitized their archives have already made the pledge to preserve and organize, and should be careful to preserve and protect their archive from obsolescence.

News archives form a symbiotic relationship with data-driven news apps, whether they exist for a specific event (think World Cup coverage or election results), or exist as standalone and evergreen frameworks (such as the HomicideWatch, Syria Deeply, or Timelines app). Such apps bypass the URL, but in the process they have their own challenges when saving and archiving. Scott Klein, luminary of news apps at the *Times*, brings up Adrian Holovaty's ChicagoCrime.org, which Holovaty described as "one of the original map mashups, combining crime data from the Chicago Police Department with Google Maps."[42] It is now defunct and unreachable in its original form; while the data survives, we have lost the presentation, and more importantly, the work, process, and context behind it.

There's no doubt that software constantly races against obsolescence, and some apps must be retired when their event passes or their function is done. But the lost process and context is lost knowledge. In March 2014, a group of NICAR conference attendees gathered at the Newseum to brainstorm the challenges and potentials of preserving news apps, suggesting more collaboration with libraries, museums, and cultural heritage institutions.[43] Some such institutions are offering novel ways of preserving and maintaining digital work for the future. At the Cooper–Hewitt Museum, a team led by Seb Chan has been preserving the previously for-profit Planetary app as "a living object." For the Cooper–Hewitt, preserving an app is more like running a zoo than a museum, where "open sourcing the code is akin to a panda breeding

41. Ibid.

42. Adrian Holovaty, "In memory of chicagocrime.org," Holovaty.com, January 31, 2008, accessed March 9, 2015, `http://www.holovaty.com/writing/chicagocrime.org-tribute/`.

43. "OpenNews/hackdays/archive," MozillaWiki, accessed March 9, 2015, `https://wiki.mozilla.org/OpenNews/hackdays/archive`.

program."[44] They'll preserve the original, but also shepherd the open-source continuation of app development, thereby protecting its offspring. While the Cooper–Hewitt is currently guarding Silicon Valley technology, overlaps and partnerships between newspapers and cultural heritage institutions could lead to similar experiments.

Some publishers have thrown up their hands altogether, relying on third-party services to organize and provide access to their own archives.[45] Reporters might suddenly see old stories disappeared, locked away behind services like LexisNexis. Such services provide fast and effective text search at low cost; but at what cost to an organization's brand, legal rights, and sense of history? A digital story is not just text, and increasingly, an archive includes images, videos, charts, maps, interactives, facts, statistics, quotations, comments, and annotations. Newer forms of classification can take a more holistic view of media, allowing a researcher to browse through text, image, sound, and video alike, and bypass the language limitations of search. This will become increasingly important as media evolves in a "post-text" web.[46] Although newsroom librarians are increasingly disappearing and publishers are seeing a proliferation of new media enter their archive, the next generation of media companies cannot rely alone on text search to access their past.

The second stage is to ***atomize*** the archive, or to break these scanned pages into their consituent parts. But what metadata is worth saving: the text, the subtext, the pictures? The photo or pullquote on the side? Is the image in the center of the page associated with the article on the left, the right, or both?

Newspapers are rich archival documents, because they store both ephemera and history. Journalists sometimes divide these types of news into "stock" and "flow"; the constant stream of information built for *right now*, versus the durable, evergreen stuff, built to stand the test of time.[47] Newspapers also have advertisements, classifieds,

44. Seb Chan, "Planetary: collecting and preserving code as a living object," Cooper Hewitt Smithsonian Design Museum, August 26, 2013, accessed March 9, 2015, `http://www.cooperhewitt.org/2013/08/26/planetary-collecting-and-preserving-code-as-a-living-object/`.

45. Jim Romenesko, "U.S. News deletes archived web content published before 2007," Romenesko, February 18, 2014, `http://jimromenesko.com/2014/02/18/u-s-news-deletes-content-published-before-2007/`.

46. Felix Salmon, "Why I'm joining Fusion," Medium, April 23, 2014, accessed March 9, 2015, `https://medium.com/@felixsalmon/why-im-joining-fusion-4dbb1d82eb52`.

47. Robin Sloan, "Stock and Flow," Snarkmarket, January 18, 2010, accessed April 22, 2014, `http:`

stock quotes, and weather diagrams. Many researchers rely on such ephemera—James Mussell calls it "a key instrument of cultural memory"—so from the archivist's perspective, everything needs to be stored.[48] But historians might treat or navigate through ephemera differently, and each set of documents could have its own metadata or interface as a result.

A newspaper is a very complex design object with specific archival affordances; their irregular size, seriality, and great care in page placement make them ripe for unique forms of automated analysis. For some researchers, placement will be important (was an article's headline on the first page? Above or below the fold? Was there an image, or a counterpoint article next to it?). Others could be examining the newspaper itself over time, rather than the contents within (for instance, did a paper's writing style or ad placement change over the course of a decade?) Still others may be hoping to deep-dive into a particular story across various journals. In each case, we can glean information from where and when it was published on the page.

The project of atomizing the archive should take advantage of the signals built into newspapers in the first place, accumulating metadata from its size, shape, and context. An atomized archive should also provide a solid interface for viewing the original in its context. When legacy news publishers refer to a "linked" record in a digital archive, they are refering to this ability to see the original source page. Some publishers do not even have linked records for their entire archive, which makes context difficult to grasp for interested researchers. Even legacy publishers with atomized archives often have clunky ways of accessing them. Small usability issues add up, and lead to lower use of archives. Atomizing the archive is not helpful without a good interface.

It is telling that many of the digitization projects, begun decades ago, focused exclusively on salvaging the text. This ignores substantial information in the archive, of course, and speaks to the shortsightedness of many projects aimed at digitizing the past. Images, advertisements, maps, formatting, and related metadata were all

//snarkmarket.com/2010/4890.

48. James Mussell, "The Passing of Print," *Media History* 18, no. 1 (February 2012): 77–92, accessed March 9, 2015, http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=70332004&site=ehost-live.

lost, and many of them are being re-scanned by publishers, at great expense, in order to properly atomize the archive; to capture the details that they ignored years ago. Nicole Maurantonio criticizes old newspapers for ignoring the visual in favor of text, "propelling scholars down a misguided path."[49] Keith Greenwood finds that newspapers diligently archived their photographs for daily newspaper use, but did not tag items with public historical value in mind, rendering many of them useless as historical records.[50]

Historical images are one of the greatest potential sources of engagement and revenue for news archives, and it would be relatively easy for some news archives to sell old photographs with historic value.[51] Some metadata projects in the publishing world are aiming specifically at images, like the New York Times' *Madison* project, which hopes to crowdsource insight about 1950s *Times* advertisements.[52] Outside the publishing sphere, Kalev Leetaru took an image-centric approach to the Internet Archive. The Internet Archive's OCR software threw out images, and Leetaru's would save whatever it threw out as an image file. He has since put 2.6 million of these Internet Archive images onto Flickr for open use. "They have been focusing on the books as a collection of words," he told the BBC. "This inverts that."[53] Newspaper and journal images provide a richer glimpse of history, and one that might prove more engaging to digital readers than dated text. You get a sense of the visual language and associations of the time; as any visual critic or cultural studies scholar can tell you, photos and advertisements provide a revealing window into the patterns and contingencies of culture and history.[54]

The final stage is to ***link*** the archive, which when considered on a massive scale,

49. Nicole Maurantonio, "Archiving the Visual," *Media History* 20, no. 1 (January 2014): 90.

50. Keith Greenwood, "Digital Photo Archives Lose Value As Record of Community History," *Newspaper Research Journal* 32, no. 3 (2011): 82–96, accessed March 9, 2015, `http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=65540347&site=ehost-live`.

51. someone's gotta be doing this already, find them

52. Madison can be found at http://madison.nytimes.com/.

53. Leo Kelion, "Millions of historical images posted to Flickr," BBC News, August 29, 2014, accessed March 9, 2015, `http://www.bbc.com/news/technology-28976849`.

54. Susan Sontag, *On Photography* (New York: Farrar, Straus / Giroux, 1977); Roland Barthes, "Rhetoric of the Image," in *Image-Music-Text* (Hill / Wang, 1978), 32–51.

is a quixotic endeavor along the lines of the Radiated Library or Project Xanadu. We can't predict all possible links between every possible piece of content. But linking the archive requires learning from the explicit and implicit references that already reside in the stories. This combines manual and automatic means, supplementing dedicated search in the surfacing of archival content, and using a "push" rather than "pull" method for finding archival materials. A user doesn't always know exactly what he or she wants, and a linked archive can work with a user to surface it. If the archive doesn't have the resource a user needs, could it at least point the user in the right direction? Could it interface with other knowledge bases to retrieve the answer?

The linked archive borrows from, but is distinct from the notion "link journalism" or "networked journalism." As a term popularized by Jeff Jarvis to refer to the growing citizen journalism movement, networked journalism has also led to Jarvis's succinct motto of "Cover what you do best, link to the rest.".[55] Building on this notion, Charlie Beckett posits that linking between sources leads to editorial diversity, connectivity and interactivity, and relevance.[56] A networked archive turns the conversation inward; as Mark Deuze and others note, links that point inward are vastly different from those that point out, but they can adhere to the same principles of diversity, connectivity, and relevance.

It is unhelpful to have a massive, borderless archive, but linked archives can expand their borders strategically through clever use of APIs. As Anne Helmond's "Boundaries of a website" and the Open Knowledge Foundation's "The News Reads Us" project remind us, publishing websites rarely operate alone; they rely on third-party platforms and services for analytics, sharing, commenting, and recommendations.[57] One could similarly integrate with APIs that offer archival resources from

55. Jeff Jarvis, "New rule: Cover what you do best. Link to the rest," BuzzMachine, February 22, 2007, accessed March 9, 2015, `http://buzzmachine.com/2007/02/22/new-rule-cover-what-you-do-best-link-to-the-rest/`; Jeff Jarvis, "Networked journalism," BuzzMachine, July 5, 2006, accessed February 7, 2015, `http://buzzmachine.com/2006/07/05/networked-journalism/`.

56. Charlie Beckett, "Editorial Diversity: Quality Networked Journalism," Polis, March 15, 2010, accessed March 9, 2015, `http://blogs.lse.ac.uk/polis/2010/03/15/editorial-diversity-quality-networked-journalism/`.

57. Anne Helmond, "Exploring the Boundaries of a Website. Using the Internet Archive to Study Historical Web Ecologies" (MIT8, Cambridge, MA, 2013), accessed December 15, 2013, `http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-`

around the web. If a user is coming to a publisher's search bar instead of Google, it's because they want more context than a mere list or index of items. They want less containment and more connection. A user should be able to see response articles, comments, tweets, timelines, images and videos, from around the web (as long as these are visually separate from the main content to avoid confusion). Otherwise, users will continue to go to Google and Wikipedia for information.

A publisher's archival search interface could include results from Wikipedia, Creative Commons images from Flickr, or resources from digital libraries like Europaeana and the Digital Public Library of America—not to mention partnering with other organizations to merge archive rights, or at least indices. The linked archive is therefore intricately indexed on a small scale, but also effectively connected on a large scale, seamlessly interfacing with other archives and collections around the web.

using-the-internet-archive-to-study-historical-web-ecologies/; Stefan Wehrmeyer, Annabel Church, and Friedrich Lindenberg, "The News Reads Us," Open Knowledge Foundation, accessed March 9, 2015, https://github.com/okfde/the-news-reads-us.

Bibliography

Abel, Richard. "The Pleasures and Perils of Big Data in Digitized Newspapers." *Film History* 25, no. 1 (January 2013): 1–10. Accessed March 9, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=88234573&site=ehost-live.

Barthes, Roland. "Rhetoric of the Image." In *Image-Music-Text*, 32–51. Hill / Wang, 1978.

Beckett, Charlie. "Editorial Diversity: Quality Networked Journalism." Polis. March 15, 2010. Accessed March 9, 2015. http://blogs.lse.ac.uk/polis/2010/03/15/editorial-diversity-quality-networked-journalism/.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press, August 28, 2000.

Cattuto, Ciro, Vittorio Loreto, and Luciano Pietronero. "Semiotic dynamics and collaborative tagging." *Proceedings of the National Academy of Sciences* 104, no. 5 (January 30, 2007): 1461–1464. Accessed February 8, 2015. http://www.pnas.org/content/104/5/1461.

Chakrabarti, Soumen. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, 2003.

Chan, Seb. "Planetary: collecting and preserving code as a living object." Cooper Hewitt Smithsonian Design Museum. August 26, 2013. Accessed March 9, 2015. http://www.cooperhewitt.org/2013/08/26/planetary-collecting-and-preserving-code-as-a-living-object/.

Dahlgren, Peter. "Media Logic in Cyberspace: Repositioning Journalism and its Publics." *Javnost - The Public* 3, no. 3 (January 1, 1996): 59–72. Accessed March 9, 2015. http://dx.doi.org/10.1080/13183222.1996.11008632.

Debrouwere, Stijn. "Information architecture for news websites." Stdout.be. April 5, 2010. Accessed March 8, 2015. http://stdout.be/2010/04/06/information-architecture-for-news-websites/.

———. "Tags don't cut it." Stdout.be. April 6, 2010. Accessed September 15, 2014. http://stdout.be/2010/04/07/tags-dont-cut-it/.

Deuze, MArk. "The Web and its Journalisms: Considering the Consequences of Different Types of Newsmedia Online." *New Media & Society* 5, no. 2 (June 1, 2003): 203–230. Accessed February 8, 2015. http://nms.sagepub.com.libproxy.mit.edu/content/5/2/203.

Discovering and scoring relationships extracted from human generated lists. U.S. Classification 707/765, 707/803; International Classification G06F17/30, G06F7/06; Cooperative Classification G06F17/30053, G06F17/30657, G06F17/30867, G06F17/30663; European Classification G06F17/30E4P US8108417 B2, filed January 31, 2012. Accessed March 9, 2015. http://www.google.com/patents/US8108417.

Gooding, Paul. "Exploring Usage of Digitised Newspaper Archives through Web Log Analy. . .." DH2014, Lausanne, Switzerland, July 9, 2014. Accessed March 9, 2015. http://www.slideshare.net/pmgooding/dh2014-pres.

Greenwood, Keith. "Digital Photo Archives Lose Value As Record of Community History." *Newspaper Research Journal* 32, no. 3 (2011): 82–96. Accessed March 9, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=65540347&site=ehost-live.

Helmond, Anne. "Exploring the Boundaries of a Website. Using the Internet Archive to Study Historical Web Ecologies." MIT8, Cambridge, MA, 2013. Accessed December 15, 2013. http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-using-the-internet-archive-to-study-historical-web-ecologies/.

Holovaty, Adrian. "A fundamental way newspaper sites need to change." Holovaty.com. September 6, 2006. Accessed March 8, 2015. http://www.holovaty.com/writing/fundamental-change/.

———. "In memory of chicagocrime.org." Holovaty.com. January 31, 2008. Accessed March 9, 2015. http://www.holovaty.com/writing/chicagocrime.org-tribute/.

Innovation. New York Times, March 24, 2014. Accessed January 24, 2015. https://www.scribd.com/doc/224332847/NYT-Innovation-Report-2014.

Jarvis, Jeff. "Networked journalism." BUZZMACHINE. July 5, 2006. Accessed February 7, 2015. http://buzzmachine.com/2006/07/05/networked-journalism/.

———. "New rule: Cover what you do best. Link to the rest." BUZZMACHINE. February 22, 2007. Accessed March 9, 2015. http://buzzmachine.com/2007/02/22/new-rule-cover-what-you-do-best-link-to-the-rest/.

Kelion, Leo. "Millions of historical images posted to Flickr." BBC News. August 29, 2014. Accessed March 9, 2015. http://www.bbc.com/news/technology-28976849.

Klein, Ezra. "Vox is our next." The Verge. January 26, 2014. Accessed December 23, 2014. http://www.theverge.com/2014/1/26/5348212/ezra-klein-vox-is-our-next.

Levy, Nicole. "Time.com opens its 'Vault' | Capital New York." Capital New York. November 12, 2014. Accessed November 12, 2014. http://www.capitalnewyork.com/article/media/2014/11/8556503/timecom-opens-its-vault.

Manoff, Marlene. "Archive and Database as Metaphor: Theorizing the Historical Record." *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398. Accessed December 2, 2013. http://muse.jhu.edu.libproxy.mit.edu/login?auth=0&type=summary&url=/journals/portal_libraries_and_the_academy/v010/10.4.manoff.html.

Maurantonio, Nicole. "Archiving the Visual." *Media History* 20, no. 1 (January 2014): 88–102.

McCain, Edward. "Saving the news: When your server crashes, you could lose decades of digital news content ¯ forever | RJI." Reynolds Journalism Institute. July 16, 2014. Accessed March 9, 2015. http://www.rjionline.org/blog/saving-news-when-your-server-crashes-you-could-lose-decades-digital-news-content-forever.

Mullin, Benjamin. "How The Boston Globe is covering the Boston Marathon bombing trial." March 8, 2015. Accessed March 9, 2015. http://www.poynter.org/news/mediawire/325301/how-the-boston-globe-is-covering-the-boston-marathon-bombing-trial/.

Mussell, James. "Elemental Forms." *Media History* 20, no. 1 (January 2014): 4–20.

———. "The Passing of Print." *Media History* 18, no. 1 (February 2012): 77–92. Accessed March 9, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=70332004&site=ehost-live.

Nothman, Joel. "Grounding event references in news" (August 31, 2013). Accessed February 8, 2015. http://ses.library.usyd.edu.au:80/handle/2123/10609.

"OpenNews/hackdays/archive." MOZILLAWIKI. Accessed March 9, 2015. https://wiki.mozilla.org/OpenNews/hackdays/archive.

Pavlik, John V. "The Future of Online Journalism." *Columbia Journalism Review* 36, no. 2 (August 7, 1997): 30–36. Accessed February 8, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=9709105522&site=eds-live.

Romenesko, Jim. "U.S. News deletes archived web content published before 2007." Romenesko. February 18, 2014. http://jimromenesko.com/2014/02/18/u-s-news-deletes-content-published-before-2007/.

Salmon, Felix. "Why I'm joining Fusion." Medium. April 23, 2014. Accessed March 9, 2015. https://medium.com/@felixsalmon/why-im-joining-fusion-4dbb1d82eb52.

Sample, Ian, Science Editor, and in San Jose. "Google boss warns of 'forgotten century' with email and photos at risk." The Guardian. February 13, 2015. Accessed March 9, 2015. http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf.

Shirky, Clay. "Ontology is Overrated: Categories, Links, and Tags." 2005. http://www.shirky.com/writings/ontology_overrated.html.

Sillesen, Lene. "Minus proper archives, news outlets risk losing years of backstories forever." Columbia Journalism Review. July 21, 2014. Accessed July 24, 2014. http://www.cjr.org/behind_the_news/minus_proper_archives_many_new.php.

Sloan, Robin. "Stock and Flow." Snarkmarket. January 18, 2010. Accessed April 22, 2014. http://snarkmarket.com/2010/4890.

Sontag, Susan. *On Photography*. New York: Farrar, Straus / Giroux, 1977.

Starr, Paul. *The Creation of the Media: Political Origins of Modern Communications*. Basic Books, 2004.

Wehrmeyer, Stefan, Annabel Church, and Friedrich Lindenberg. "The News Reads Us." Open Knowledge Foundation. Accessed March 9, 2015. https://github.com/okfde/the-news-reads-us.