The Missing Links:
An Archaeology of Digital Journalism

by

Liam Phalen Andrew

B.A., Yale University (2008)

Submitted to the Department of Comparative Media Studies/Writing
in partial fulfillment of the requirements for the degree of

Master of Science in Comparative Media Studies

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . .
Department of Comparative Media Studies/Writing
May 8, 2015

Certified by . . . .
William Uricchio
Professor of Comparative Media Studies
Thesis Supervisor

Accepted by . . . .
T.L. Taylor
Director of Graduate Studies, Comparative Media Studies

The Missing Links:

An Archaeology of Digital Journalism

by

Liam Phalen Andrew

Submitted to the Department of Comparative Media Studies/Writing
on May 8, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Comparative Media Studies

Abstract

As the pace of publishing and the volume of content rapidly increase on the web, citizen journalism and data journalism have threatened the traditional, 20th century role of journalism and institutional newsmaking. Journalists and publishers are beginning to adapt to this new news landscape, but the role of legacy media institutions, and even digital-native outlets and news aggregators, is still in flux and under debate. Many legacy institutions are drawing newfound value from their archival stories, but their potential remains limited by technical challenges and institutional inertia.

In this thesis I propose a framework for considering the news institution of the digital era as a *linked archive*: equal parts news provider and information portal, the linked archive places historical context on the same footing as new content, and emphasizes the journalist's role as news explainer and verifier. Informed by a theoretical, historical, and technical understanding of the web's structural affordances and limitations, and especially by the untapped networking power of the hyperlink, I suggest how publishers can offer an archive-oriented model of sustainable and scalable journalism. I draw from concepts and lessons learned in library and computer science, such as link analysis, network theory, and polyhierarchy, to offer an archivally-focused journalistic model that can save time for reporters and improve the research and reading process for journalists and audiences alike. This allows for a treatment of news items as part of a dynamic conversation rather than a static box or endless feed, revitalizing the news archive and putting the past in fuller and richer dialogue with the present.

Thesis Supervisor: William Uricchio
Title: Professor of Comparative Media Studies

Acknowledgments

In the spirit of attribution, this thesis would not exist without the influence and support of many advisors, colleagues, and friends at MIT and beyond.

Thanks first and foremost to my committee members—William Uricchio, Kurt Fendt, and Ethan Zuckerman—each of whom brought a unique and invaluable mix of expertise, curiosity, encouragement, and criticism. I'm also grateful to the many faculty members and advisors who have taught me and given feedback, especially T.L. Taylor, Heather Hendershot, Sherry Turkle, David Weinberger, Jim Paradis, Rahul Bhargava, and Matt Carroll.

My CMS colleagues helped me hone my topic and my writing, both in class and over countless pitchers at the Muddy Charles Pub. Many thanks in particular to the 2015 cohort: Chelsea Barabas, Heather Craig, Suruchi Dumpawar, Sean Flynn, Desi Gonzalez, Jesse Sell, Erik Stayton, Ainsley Sutherland, and Wang Yu.

I've been lucky to learn from several research groups, all of whom provided many ideas and inspirations, especially the HyperStudio crew (Jamie Folsom, Gabriella Horvath, and Rachel Schnepper). Thanks as well to the Nieman Lab, Knight Lab, and my fellow Google Fellows for an enlightening summer, particularly Josh Benton, Caroline O'Donovan, Justin Ellis, and Yossi Lichterman. I'm also grateful to Civic Media, the Berkman Center, and the Open Documentary Lab for many enlightening discussions, particularly with Sands Fish, Alexis Hope, and Lara Baladi.

Thanks to the newsroom librarians, editors, and developers who spoke with me, notably Lily Rothman, Josh Rothman, and Lisa Tuite. Thanks also to the Lasertag team (Meredith Broussard, Lindsey Cook, James Flanagan, Scott Shigeoka, and Matt Waite) for tackling archive archaeology with me in Doha. And I owe the seeds of this thesis to the original Wiser team (Sandeep Ayyappan, Cara Matteson, Tom Weingarten, and Andy Whalen).

Thanks to my family, especially James and Joey for their hospitality, and of course Ninja, Maple, and Percy (RIP). And thanks most of all to Liz, who has endured this at least as much as I have, with undying support, humor, and spicy food.

Contents

List of Figures

0.1 Preface

For the past two years, I've been doing research online about the problems with doing research online. This has made my head spin more than once. I keep running into the very problems I want to address. I've discovered so many tantalizing online resources, only to encounter the dreaded `404 NOT FOUND`. My note repositories and reference lists have ballooned to unmanageable sizes. I've shared frustrations and frequent discussions with my colleagues about the best tools and techniques for organizing resources and ideas. And I've spent countless hours trying to collect my own thoughts, and make connections between everything I read, process, and understand. Computers are very good at helping us store and remember information, but they are less adept at making connections between the bits of data that they remember. I want my notes and citations to reflect and expand on my own memories and ideas; too often they obfuscate and distract from them instead.

I spent the summer in between these two years at Harvard's Nieman Journalism Lab, and here I turned my attention towards the digital journalism sphere. Beleaguered graduate students and academics are not the only ones facing the perils of online research; the modern journalist is increasingly playing the role of online researcher and information filterer, too. As amplifiers of important stories and topics for public discussion, journalists form a crucial bridge between the individual and the collective; their information problems affect their readers' knowledge and understanding in turn. This also highlights how online actors—publishers, information technology companies, and cultural heritage institutions alike—are fighting for readers' attention and preserving knowledge for the future.

For centuries, librarians and archivists collected and sorted out our history and access to the past. Now the majority of our personal history is collected and sorted by newer, more automated, and less tested methods for determining what something is about and whether it's worth saving. And the librarians and archivists aren't the main ones doing the sorting. Some of these developments offer great potential and excitement; the web has enabled an explosion of data and access to it. The past is bigger and closer to us than ever before. This is an exciting prospect, and much of today's utopian rhetoric around big data and information access draws from such potential; but the landscape is new and the actors involved often operate in secrecy. This extends beyond a journalist or academic conducting efficient research; these individual-scale problems combine to affect our collective understanding of history.

This thesis is, perhaps ironically, a mostly linear document about hypertext. This is not to say that each chapter must be read completely and sequentially; some chapters might be more pertinent to publishers, journalists, and newsmakers, while others may resonate with information theorists and digital humanists. But I endeavor to make the whole greater than the sum of its parts, and my goal is to highlight the overlaps between these worlds, which are each enabled and threatened by the same technological and economic forces. So this linear thesis hopes to bring these camps together in a single piece rather than encouraging fragmentary readings down familiar pathways. When Ted Nelson republished his famous *Literary Machines*, he lamented: "It should have been completely rewritten, of course, but various people have grown to love its old organization, so I have grafted this edition's changes onto the old structure." Perhaps this thesis should also be completely rewritten, and structured as hypertext; but for now, I will merely aim to graft its connections and diversions onto the usual format, and cohere into a linear structure that I will leave for future efforts to take apart.

Chapter 1

Introduction

"News is the first rough draft of history," as the saying goes. Or is that the saying? Long attributed to *Washington Post* president and publisher Philip L. Graham, it turns out that many might have said it before him, in many variations. His widow Katharine points to a speech he made in London 1963, but journalist Jack Shafer finds Graham using the phrase in 1948 and 1953, too.[1] Even earlier, Barry Popik discovers its use in a 1943 *New Republic* book review, not said by Graham at all but by Alan Barth. The *Yale Book of Quotations* credits Douglass Cater, who wrote in 1959 "The reporter [is] one who each twenty-four hours dictates a first draft of history."[2] The variations don't stop here. Did he say "first draft" or "first rough draft"? Was it "journalism," "the news," or just "newspapers"? A quick search on Google finds thousands of results under all six of these variations ("newspapers are the first draft of history" gets 36,000 hits, while "news" and "journalism" each break 10,000).

Attributing a famous phrase has never been easy, but on the web it's an even tougher challenge. On one hand, attribution is facilitated through the foundational element of the web: the hyperlink. It has never been easier to point to your source. But on the other, the lack of standards and ease of copying make tracing any information to its definitive source nearly impossible. It requires following detailed digital traces, which are each owned and controlled by various actors and can be wiped out

[1] Jack Shafer, "Who Said It First?," *Slate* (August 30, 2010), accessed April 19, 2015, http://www.slate.com/articles/news_and_politics/press_box/2010/08/who_said_it_first.html.

[2] Fred R. Shapiro, *The Yale Book of Quotations* (Yale University Press, 2006), 139.

at a moment's notice.

But the seemingly fleeting, ephemeral nature of online media is at odds with content's perennial availability on the web. Entertainment journalist Andy Greenwald joked in a 2015 tweet, "Remember: Newspapers are the first draft of history. But Twitter is a heavy stone tablet that you have to wear around your neck forever."[3] When compared to newspaper articles, journalists' tweets could be seen as both even rougher drafts and as more permanently accessible artifacts. Newer platforms like Snapchat purport to erase these constant digital traces, but these promises are unverifiable, and there are ways to overcome them; online content is usually here to stay, even as it is being produced on tighter, hastier deadlines. By default, we store everything, and whether or not anything has a "right to be forgotten" is one of the major legal and ethical issues of our time.[4]

In following the genesis and the echoes of a famous or iconic quote, article, photograph, or recording, a researcher becomes a digital sleuth trawling through databases and following links in hopes of reaching some sort of origin or record of influence. This is not only how online researchers do their work (journalists and scholars alike)—it is also how search engine algorithms gain an understanding of influence and impact. The writers of these algorithms influence the results for the human researchers in turn, and the lack of transparency surrounding many of their methods is a substantial cause of debate. While journalism serves as a crucial case study in the problems of measuring attribution, impact, and influence online, the phenomenon runs much deeper. Debates over linking, forgetting, and storing have most recently been fought in the context of privacy and surveillance; this aspect is crucial, but it is only part of the story. These debates also consider what will and will not constitute our past,

[3] Andy Greenwald, "Remember: Newspapers are the first draft of history. But Twitter is a heavy stone tablet you have to wear around your neck forever.," @andygreenwald, March 31, 2015, accessed April 19, 2015, https://twitter.com/andygreenwald/status/582913646788886528.

[4] See, e.g., Charles Arthur, "Explaining the 'right to be forgotten' - the newest cultural shibboleth," *The Guardian* (May 14, 2014), accessed April 19, 2015, http://www.theguardian.com/technology/2014/may/14/explainer-right-to-be-forgotten-the-newest-cultural-shibboleth; Danny Hakim, "Right to Be Forgotten? Not That Easy," *The New York Times* (May 29, 2014), accessed April 19, 2015, http://www.nytimes.com/2014/05/30/business/international/on-the-internet-the-right-to-forget-vs-the-right-to-know.html.

whether it is the memory of a single person or the history of a political or cultural revolution. To access our past on the web, we can do little more than trace the links, which are themselves the subject of much policing and debate.

1.1 The Stakes

In the spirit of examining the relationship between news and history, I will introduce a few examples from the recent news, all of which occurred in the year I wrote the bulk of this thesis. These events are merely the latest in a string of similar smaller events and controversies that will quickly be forgotten, or at best brought in as passing evidence for larger phenomena. It's possible they will even be deleted from the web outright, saved only as phantom traces in the Internet Archive or Google. But here they will be baked into this scan of this printout of this copy of a Masters thesis, where a physical copy might reside in the MIT libraries for decades.

On April 13, 2015, an argument broke out on Twitter. This would not typically be news, except that it involved two titans of a new wave of digital, data- and technology-oriented journalists: Nate Silver of FiveThirtyEight, and Ezra Klein of Vox. Silver cast the first stone, saying "Yo, @voxdotcom: Y'all should probably stop stealing people's charts without proper attribution. You do this all the time, to 538 & others."[5] When some respondents demanded proof of stealing, FiveThirtyEight writer Carl Bialik linked to two stories about the 2016 presidential election, each featuring a FiveThirtyEight chart. However, Vox added "attribution" – a link – in under 10 minutes, which confused readers who came late to the story and saw a link back to the source. More debate ensued about whether Vox had updated the timestamp on the article to reflect the attribution that was made.

This debate and the ensuing confusion bring up a series of issues that stem from the structural limitations of the web: there is no default storage of attribution, and no built-in way to dive back in time and see what Vox's page looked like 10 minutes

[5] Nate Silver, "Yo, @voxdotcom: Y'all should probably stop stealing people's charts without proper attribution. You do this all the time, to 538 & others.," @NateSilver538, April 13, 2015, accessed April 16, 2015, https://twitter.com/NateSilver538/status/587646527855849472.

before. The web has these features grafted onto it, rather than built in; we rely on web browsers, Google, and the Internet Archive to save, store, and index for us, and there's no telling where the information originated. Even journalists are copying and repurposing; in response to the Twitter debacle, Vox's Ezra Klein wrote a guide to "How Vox aggregates," where he suggests that aggregation and linking are at the core of modern digital news. Still, he insists that an aggregator needs to *link* to the source, and not just reference it. Klein defends Vox's use of aggregation in the article, "but the post didn't include a link. This was carelessness, not malice, but it's a violation of Vox's internal standards. Our policy requires attribution, and any time we fail that policy is inexcusable."[6]

Aggregation is nothing new; 19^{th} century newspapers and magazines freely lifted articles from one another, on the back of substantial federal subsidies that allowed newspapers to share "exchange copies" before such work was considered intellectual property.[7] Editors would literally cut, paste, and repurpose other papers' reporting for new formats and audiences; the Viral Texts project, based at Northeastern University, endeavors to map the spread of important texts in these newspapers from the era.[8] Klein points out that *Time* magazine began as a source that would go through "every magazine and newspaper of note in the world" and take notes about it for their readers. Even at the time this was sometimes considered questionable journalism, but unlike the web, traditional newspapers did not have a direct window to the source text. As Klein emphasizes, the hyperlink serves as a proxy for attributing a source, which is a core component of doing good journalism. Because of this new ability to directly attribute, it can, arguably, make aggregation a more journalistically respectable practice. Links are also the basis of Google's PageRank algorithm, and play a substantial role in search engine optimization. Looking deeper, battles over when to link and when not to pervade modern journalism. One of the primary roles

[6] Ezra Klein, "How Vox aggregates," Vox, April 13, 2015, accessed April 14, 2015, `http://www.vox.com/2015/4/13/8405999/how-vox-aggregates`.

[7] Paul Starr, *The Creation of the Media: Political Origins of Modern Communications* (Basic Books, 2004), 90.

[8] D.A. Smith, R. Cordell, and E.M. Dillon, "Infectious texts: Modeling text reuse in nineteenth-century newspapers," in *2013 IEEE International Conference on Big Data* (October 2013), 86–94.

of editors is to help journalists stick to Jeff Jarvis' core tenet of online journalism: "do what you do best, link to the rest."[9] Other editors might realign the anchor text of a link, or reword a headline for Twitter, so that it becomes more or less enticing to click on.[10] The seemingly innocuous hyperlink thus turns into a powerful instrument in every news outlet's struggle for attention and economic value.

1.1.1 The new face of plagiarism

In summer 2014, several similarly fleeting news items occurred, which each placed a mark of their own on this larger debate. It began with writer Benny Johnson of BuzzFeed accusing the Independent Journal Review of plagiarizing his work—once again, on Twitter.[11] Two Twitter users known only as @blippoblappo and @crushingbort saw an irony in a BuzzFeed writer accusing another publisher of stealing; BuzzFeed has long been accused of aggregating, remixing, and appropriating other outlets' work without attribution.[12] Perhaps because of this, the pair of detectives searched the web for examples of Johnson's own lifting.

They were likely not aware of how deep the copying went; the pair found three instances of unattributed sentences, with sources ranging from the Guardian to Yahoo! Answers.[13] When BuzzFeed editor Ben Smith replied to the plagiarism allegations by

[9] Jeff Jarvis, "New rule: Cover what you do best. Link to the rest," BuzzMachine, February 22, 2007, accessed April 20, 2015, `http://buzzmachine.com/2007/02/22/new-rule-cover-what-you-do-best-link-to-the-rest/`.

[10] Derek Thompson, "Upworthy: I Thought This Website Was Crazy, but What Happened Next Changed Everything," *The Atlantic* (November 14, 2013), accessed April 19, 2015, `http://www.theatlantic.com/business/archive/2013/11/upworthy-i-thought-this-website-was-crazy-but-what-happened-next-changed-everything/281472/`.

[11] Benny Johnson, "Repeat after me: Copying and pasting someone's work is called "plagiarism" http://www.ijreview.com/2014/07/159684-president-george-h-w-bushs-guide-picking-perfect-pair-socks/," @bennyjohnson, July 23, 2014, accessed April 19, 2015, `https://twitter.com/bennyjohnson/status/491953975785426944`.

[12] See, e.g., Farhad Manjoo, "How To Make a Viral Hit in Four Easy Steps," *Slate* (June 26, 2012), accessed April 19, 2015, `http://www.slate.com/articles/technology/technology/2012/06/_21_pictures_that_will_restore_your_faith_in_humanity_how_buzzfeed_makes_viral_hits_in_four_easy_steps_.single.html`; Adrian Chen, "Remix Everything: BuzzFeed and the Plagiarism Problem," Gawker, July 28, 2012, accessed April 19, 2015, `http://gawker.com/5922038/remix-everything-buzzfeed-and-the-plagiarism-problem`.

[13] @blippoblappo and @crushingbort, "3 Reasons Benny Johnson Shouldn't Call Out Plagiarism: He's A Plagiarist, He's A Plagiarist, and He's A Plagiarist," Our Bad Media, July 24, 2014, accessed April 19, 2015, `https://ourbadmedia.wordpress.com/2014/07/24/benny-johnson-probably-`

calling Johnson "one of the web's deeply original writers," @blippoblappo and @crushingbort responded with six more offenses, with a similarly broad range of sources, including About.com, Wikipedia, and the New York Times.[14] This new set forced BuzzFeed to investigate, and a day later they fired Johnson and apologized to their readers; they had found a full 41 plagiarized phrases among 500 Johnson pieces.[15] The rate and ease at which these offenses seem to have been found is startling. If two researchers found so much copying in one day, and BuzzFeed's internal investigation had turned up dozens more, how could they—how could *anyone* not have not discovered these during Johnson's two years as a BuzzFeed writer? The offenses were hiding in plain sight.

The Washington Post's Erik Wemple suggested that some of these transgressions could have come from the specific demands of BuzzFeed; Johnson's "multi-topical viral beat" might have left him with not enough time to fully process the material, and not enough patience to link to every single source.[16] Ben Smith points out that BuzzFeed is certainly not the first major publisher to deal with plagiarism in its ranks; this is of course true, but there is something new at play here. It can also be found in a February 2015 plagiarism case involving Mic.com's Jared Keller, who was outed for using unattributed text from *The Atlantic* and others. Here the line between plagiarism and shoddy journalism is blurred: sometimes, "Keller lifted text from sources that he credits and links to in the post itself, but without setting that text in quotation marks or block quotes, and without noting anywhere on the page that the text was not his own."[17] Even crediting doesn't go all the way; where and how

shouldnt-call-people-out-for-plagiarism/.

[14] @blippoblappo and @crushingbort, "More Plagiarism From "One Of The Web's Deeply Original Writers"," Our Bad Media, July 25, 2014, accessed April 19, 2015, https://ourbadmedia.wordpress.com/2014/07/25/more-plagiarism-from-one-of-the-webs-deeply-original-writers/.

[15] Ben Smith, "Editor's Note: An Apology To Our Readers," BuzzFeed, July 25, 2014, accessed April 19, 2015, http://www.buzzfeed.com/bensmith/editors-note-an-apology-to-our-readers.

[16] Erik Wemple, "The ravages of BuzzFeed's Benny Johnson," *The Washington Post* (July 27, 2014), accessed April 19, 2015, http://www.washingtonpost.com/blogs/erik-wemple/wp/2014/07/27/the-ravages-of-buzzfeeds-benny-johnson/.

[17] J.K. Trotter, "Plagiarist of the Day: Mic News Director Jared Keller," Gawker, February 11, 2015, accessed April 19, 2015, http://tktk.gawker.com/plagiarist-of-the-day-mic-news-director-jared-keller-1684959192.

you link can be the line between good journalism, shoddy journalism, and outright plagiarism.

The problem that places like Vox, BuzzFeed, and Mic are facing is still fairly new; they are trying to ethically aggregate and reappropriate from other online sources, but the protocol for doing so is still unclear. While there's no doubt that Johnson and Keller stepped across this ethical line, where exactly is the line? Ben Smith's first reaction to Johnson's transgressions suggested that three offenses was not enough; he also implied that plagiarism on older articles or trite listicles would be more acceptable than on more "serious," investigative pieces. The aggregational method does not only raise new legal and ethical questions, but it also changes the mentality and practice of journalism and research. This is the case for both what can actually be found online, and what we perceive to be findable online. It is surprising that Johnson did not see himself as vulnerable; despite his obvious offenses, he accused others of plagiarism as well. Keller even offered links to the sources that he was lifting from, perhaps assuming that no one would click on them anyway.

These incidents reflect a new paradigm of attribution and authorship that blurs the distinction between original and copy, between individual and collective, and between highbrow and lowbrow. Johnson pilfered language from everywhere between Yahoo! Answers to the New York Times, with little distinction between the two. His most frequent transgressions, however, did seem to come from anonymous sources. As Wemple put it, he "viewed [Wikipedia] as an open-source document," and grabbed phrases from government reports.[18] His liberal use of Yahoo! Answers and About.com also leads to more speculation; did he somehow feel that it was more ethical (or just less risky) to take from anonymous sources than other professional writers? Who should get the original credit, and how should they be compensated? Moreover, why did Johnson feel the need to pass them off as original? Johnson's safest option would have been to simply *link to* the sources, but this would be tedious. It would also interrupt the story if the reader decided to click on a link, possibly never to return to Johnson's article again. And of course, it would lay bare Johnson's curation of often

[18] Wemple, "The ravages of BuzzFeed's Benny Johnson."

dubious sources, not only to readers, but to machines.

BuzzFeed understands well this double power of the link. Tellingly, their apology post about the Benny Johnson incident likewise did not include links to the tainted articles. When internet users pushed back on this omission, BuzzFeed updated the post with plaintext URLs, without adding hyperlinks to them.[19] Why would they do this? While it might slightly increase the friction for an interested user to get to the article, it is more likely that it was to keep web crawlers and search engines from knowing about the connection. On the web, you are what you link to, and this post didn't want to link to—or be linked to—dozens of plagiarized articles. BuzzFeed has also deleted thousands of its older posts, which they claim did not adhere to their newer, more rigorous journalistic standards; it is telling that rather than offering contextual explanation, they deleted these posts outright. While CEO Jonah Peretti offers that they were deleted because they were "not worth improving or saving because the content [wasn't] good," J.K. Trotter of Gawker also found Johnson-like moments of slippery near-plagiarism in the deleted posts.[20]

1.1.2 The summer of archives

Meanwhile, legacy news outlets have been aiming to revive rather than delete their past. 2014 was known as the "summer of archives" at the Digital Public Library of America, and the New Yorker took this saying to heart as they opened up their digital archive completely, while experimenting with a new paywall. Prior to this summer, legacy media's default move was to hide their archival content behind strict paywalls; archives were considered one of the first incentives for digital publishers to entice readers into novel subscription models. The New York Times' 2014 *Innovation* report similarly featured archives as a trove of untapped potential, and outlets like Time, the Nation, and National Geographic focused on their archives with sleek redesigns

[19] Smith, "Editor's Note."

[20] J.K. Trotter, "Over 4,000 BuzzFeed Posts Have Completely Disappeared," Gawker, August 12, 2014, accessed April 19, 2015, `http://gawker.com/over-4-000-buzzfeed-posts-have-completely-disappeared-1619473070`; J.K. Trotter, "Don't Ask BuzzFeed Why It Deleted Thousands of Posts," Gawker, August 14, 2014, accessed April 19, 2015, `http://gawker.com/don-t-ask-buzzfeed-why-it-deleted-thousands-of-posts-1621830810`.

and dedicated editorial and development staff.

The archive summer was accompanied by a historic event farther from the journalism industry; Theodor Nelson's Project Xanadu was finally released on the web. First conceived over 50 years prior, Project Xanadu was famously the first hypertext project, a distant ancestor of the web, under active development for decades.[21] Belinda Barnet says Xanadu was supposed to be "like the web, but much better;" in hindsight, Xanadu lives as a vision of an alternate hypertext system, one in which many of the pitfalls of the web — the problems of attribution, measurement, and research — are laid bare to be scrutinized and reimagined.[22] The 2014 version was made for the web, which seems like a sort of admission of defeat; it also still lacks many of the core features that comprise Nelson's Xanadu dream. But Xanadu maintains a set of devoted acolytes and followers, and the project's persistence and rebirth demonstrates a drive for understanding and incorporating some of its features.

Among Xanadu's admirers are the founders of NewsLynx, a research project and platform under development at Columbia University's Tow Center for Digital Journalism. In August 2014, NewsLynx wrote about the perils of online linking and tracking; specifically, they lamented the web's ability to only link in one direction, and praised Nelson's Xanadu for its foresight in recognizing this problem. Two-way linking would allow a user to see every document that links *to* a page as well as *from* it; without it, we are left with a "hole at the center of the web" which has allowed Google to "step in and play librarian."[23] NewsLynx's goal was to develop a platform that would help small and nonprofit news outlets better measure the impact of their work. They hoped to give each story a sort of biography; how often was a story read, or shared, and where? Who were the catalysts for distributing the work around the web? How long is a typical story's lifespan? They found these problems to be incredibly difficult at a smaller, research-oriented or nonprofit scale. Measurement methods have to be

[21] See section 3.3.1 for a detailed discussion of Xanadu.

[22] Belinda Barnet, *Memory Machines: The Evolution of Hypertext* (London: Anthem Press, July 15, 2013), "The Magical Place of Literary Memory: Xanadu".

[23] Brian Abelson, Stijn Debrouwere, and Michael Keller, "Hyper-compensation: Ted Nelson and the impact of journalism," Tow Center for Digital Journalism, August 6, 2014, accessed April 19, 2015, http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/.

grafted onto the web at great expense, rather than built into its core structure or proposed as a radical alternative.

Much like the steady stream of news on the web, these incidents from the summer of 2014 do not singularly cohere into an overarching narrative; but they do all point to a shifting understanding of original, quality journalism, and a fundamental difference in the ways that journalists do research, deliver information to readers, and understand the value and impact of their work. Where traditional plagiarism might involve faking a conversation with a source, the new face of plagiarism is simply lifting information from the web and failing to link to it. As Klein points out, sometimes this is oversight and other times it is malicious; in either case, it points to the foundational role that linking appears to play in the political economy of today's online news, and its flattening effect in producing attributable journalism.

The subtle but pervasive politics and style of linking on one hand, and the changing focus of journalism towards context and aggregation on the other, are two sides of the same coin; they are two symptoms of a larger shift in the journalistic landscape towards documents, citations, and explanation. At the same time, they reflect divergent impulses; to link *out* on one hand and *in* on the other. Most researchers tend to assume that linking out and attributing to external sources is an inherently good impulse, while inlinking to one's own archive is "nepotistic" rather than altruistic.[24] In reality these lines are much more blurred; outlinking at will is unhelpful if it drives the reader down too many confused paths, or only leads the reader to the same type of information or source under the veneer of sharing. Meanwhile, linking inwards can be crucial for continuing an ongoing story and maintaining contextual control over an ongoing event. The trick might be to embrace the best of both; one's own archive is a place to draw out and contain a compelling narrative, but the archive itself can link out and integrate with other sources, reactions, and annotations, embracing the spirit of the web.

[24] Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data* (Morgan Kaufmann, 2003), 213.

1.2 Outline of chapters

This thesis aims to highlight the tension on the web between acting as a source of free information and rich remix, and its competition with rising business interests and link economies. This thesis therefore takes journalism as a crucial case study, but it aims to reach beyond journalism for inspirations and conclusions. The chapters of this thesis follow a sort of funnel structure, increasingly honing in on journalism and its use of hyperlinks as an enabler of digital archives. The initial chapters will be necessarily broad in scope, pertaining to the multifaceted meanings and interpretations of archives as they are considered by librarians, information theorists, journalists, and critical theorists alike. Each chapter will then increasingly point towards its applications and implications in news and journalism.

Chapter two, "The Size and Shape of Archives," adopts a theoretical framework to examine the role of the World Wide Web as a container of content. Considering an online object as "content" frames it in the networked language of the web, and subsumes it into a series of semiotic frameworks that serve to *contain* the content at each stage. From the URL (Uniform Resource Locator), to the hyperlink, to the feeds and indexes that aggregate content, and finally the databases that store them, one finds content forging new connections and diluting its original core at each turn. The web's inherent ephemerality also plays into the sizes and structures of the web's many archives.

Chapter three, "An Intertwingled History of Linking," turns to the history of information structures and knowledge systems, especially looking for traces of the mechanics and functions of today's hyperlink in early classification schemes. It darts back and forth, considering at length the origin of hypertext, but also peppering the history with treatments of the encyclopedias, notecard systems, and taxonomies that predated the digital era. Rather than organizing chronologically, I consider this history from the lens of three themes and challenges that recur when considering hypertextual knowledge systems: spatialization, intersubjectivity, and encyclopedism.

The fourth chapter, "Networking the News," hones in on the changes in contemporary journalism that result from the web's structure and the many uses of hyperlinks. It treats the changing practices of journalists as a reflection of many of the containing forces of the web, and the simultaneous rise of "explainer journalism" and rich paywalled digital archives as similar symptoms of these forces. I then turn to the information architecture of digital news stories and the common methods for classifying them, ultimately advocating for a link-oriented classification scheme to supplement current tagging and taxonomic processes in the newsroom.

Chapter five, "Tracing the Links," examines how journalists are currently considering and using hyperlinks in their work. First I review the existing qualitative interviews and quantitative analyses that demonstrate journalists' understanding and practice of hyperlinking. I then turn to my own inquiries into quantitative link analysis, which hone in on the internal linking that publishers are currently doing within their archive. This approach begins to examine how a link-oriented classification scheme might work, and the cultural and journalistic changes that need to occur in order to enable such a scheme.

Finally, I will conclude by peering beyond the link, and examining contemporary applications and frameworks that offer new forms of hypertextuality and interactivity. These sometimes expand the function of the link, and other times completely bypass it, rendering traditional methods of tracking attribution, influence and impact obsolete. I will conclude by considering these new forms of journalistic output through two lenses: the frameworks and standards organizations that hope to develop a cohesive language around these new initiatives in order to better save them, and the artists and activists that complicate the existing picture of digital history and memory.

Chapter 2

The Size and Shape of Archives

This chapter will take an archaeological approach to digital content, first by analyzing the size and shape of content's containers. Containers include its many representations online—as links, quotes, embeds, and excerpts across search results and social media—and especially the World Wide Web as the Internet's dominant, overarching container. I will begin by defining and contextualizing the modern archive, as well as the word "archive" as its possible meanings and definitions have proliferated in the digital age. I will then outline the *layers of containment* that the web exerts on web content, which serve to homogenize content while simultaneously connecting it to a broader ecosystem of media. Finally, I will look at current efforts to store and preserve online content, which each bring their own approaches and limitations to understanding media's history. This chapter will be broad and theoretical in scope, but it serves to frame the issues that journalists and newsrooms face in preserving larger contextual history as well as a news outlet's daily content. I aim to highlight the extraordinary shift in the structure of archives in the digital age. New technologies and institutional pressures have necessitated new methods of storing and saving; archivists have long aimed to fastidiously organize and classify content, but now they can connect content in new ways, complementing the archive's traditional categories and taxonomies with an underlying network.

The digital affords new abilities for *linking* or *networking* the archive, allowing it to dynamically expand, contract, reform, and change shape. In the linked archive, we can forge new connections and create more nuanced context for the information stored inside. Most of today's digital archives and knowledge systems take advantage of some of these new linking features, but they also still inherit many of the limitations of their physical predecessors. Libraries, archives, publishers and others should strive to link and network their archival material.

A linked archive is a collection that: a) treats its contents as an ecosystem of discourses rather than a brittle item to put in boxes; b) actively forms, re-forms, and presents information in more nuanced ways than traditional search; c) gracefully takes in new content and information for future reuse; and d) interfaces with any number of other archives to expand, contract, or reframe its borders. A well-linked archive places context on the same level as content, acknowledging the constantly expanding and shifting shape of research, inquiry and history, and putting the past in full dialogue with the present.

2.1 Defining the archive

The word "archive" brings to mind a stuffy room full of closely guarded old books and manuscripts. In the traditional archive or library, books can only be in one place at one time, and always next to the same exact books on the same exact shelf. The atomic unit of information tends to be the book, manuscript, or box of papers, even though each of these contains multitudes of media (text, images, maps, diagrams) and the bibliographies and indexes that offer a window into a book's constituent parts remain limited by space and language. And if your archive dive takes you beyond the scope of the current archive, you'll have to travel to a different one.

But archives come in many forms. More recently, an archive is likely to be digitized, stored on networked servers in databases. Here the archive's stacks and files are virtual, and can be ordered and reordered at will. Books and documents are further atomized and calculable as data. If a search goes beyond the digital archive's scope, it may even be able to reach for information outside of it. "Archive" now even turns up as a common verb in digital information management; instead of deleting

Google emails, we archive them, which in a sense *de*-organizes it and hides it away. All the same, the message is clear: the email is not gone but archived, saved forever by Google's automatic librarians.

The notion of the archive has changed in both structure and function in the digital age. As both an entity and an action, "archive" has perpetually expanding and shifting meanings. Here I will endeavor to define the archive as I will use the term, first by placing it in a lineage of other remediated digital words, then in the context of its use in digital humanities, media archaeology, and software studies.

2.1.1 "Thing" words

"Archive" is one of many words that has become increasingly generic and abstract in scope with the introduction of digital media. We often need such generic, all-encompassing words—words that describe a broad swath of things in a very general manner ("things" being one such word). While reality can be sliced and diced in any number of ways, we sometimes need to talk about the undivided whole. A word like "thing" encompasses many words (and actual things) inside it, which can be envisioned as a hierarchy or set of concentric circles around an entity; for example, ordered by levels of abstraction, my tabby cat could be called a tabby, a cat, a mammal, a vertebrate, an organism, or a thing (roughly following Linnaeus' biological taxonomy). This hierarchical structure of language both reflects and shapes the ways in which we have historically classified and organized knowledge, ever since Plato began searching for the "natural joints" in reality, and through some of the most canonical examples: Linnaeus' taxonomy and Dewey's Decimal System.

Today's methods of classifying—and possibly, organizing knowledge in general—have radically changed, and we increasingly need such generic words to describe the digital, ephemeral world around us. The information age has brought us objects, data, documents, information, and content. Its processes include products, services, applications and platforms. But what is a document, or data? How does our use of these words carry contextual weight?

Terms like these are far removed from the realities they describe, and often just as far removed from their original meanings. Remediated words balance an inheritance and a distance from their original (premediated) contexts, and much work has explored the long histories of these terms. Daniel Rosenberg charted the use of the term "data" through shifting contexts since the 18th century, noting that it was initially used to describe an indisputable fact or "given" in an argument (from Latin *dare*).[1] Annette Markham likewise questions the use of the word "data" in its modern context, suggesting that, "through its ambiguity, the term can foster a self-perpetuating sensibility that 'data' is incontrovertible, something to question the meaning or veracity of, but not the existence of."[2] Johanna Drucker suggests implementing its counterpart "capta," which highlights the inherently plucked and pre-envisioned nature of all information.[3]

Other contemporary words have been similarly historicized and questioned. John Seely Brown and Paul Duguid trace the history of the word "information" in *The Social Life of Information* and forthcoming research, highlighting its long history as an "unanalyzed term."[4] Likewise, Tarleton Gillespie draws attention to the word "platform" in the context of the software industry, focusing on the implications of the term's historical meanings.[5] "Platform" was both an object (e.g. a soapbox) and a concept (e.g. a political platform), and Gillespie sees the word's use in software as borrowing from and conflating these traditional meanings. In each of these cases, the appropriation of abstract words informs and reshapes our own notions of these words and the objects and realities that they represent.

One such remediated word, foundational to the web, is the "document." It was previously understood as a physical, printed record—usually an original. A signed

[1] Daniel Rosenberg, "Data Before the Fact," in *"Raw Data" is an Oxymoron*, ed. Lisa Gitelman (Cambridge, MA: MIT Press, 2013), 15-40.

[2] Annette N. Markham, "Undermining 'data': A critical examination of a core term in scientific inquiry," *First Monday* 18, no. 10 (September 21, 2013), http://firstmonday.org/ojs/index.php/fm/article/view/4868.

[3] Johanna Drucker, "Humanities Approaches to Graphical Display," *Digital Humanities Quarterly* 5, no. 1 (2011), http://www.digitalhumanities.org/dhq/vol/5/%201/000091/000091.html.

[4] John Seely Brown and Paul Duguid, *The Social Life of Information* (Harvard Business Press, 2002).

[5] Tarleton Gillespie, "The Politics of 'Platforms'," *New Media & Society* 12, no. 3 (May 1, 2010): 347–364.

mortgage might be a document, but a photocopy was not; the word "document" went hand in hand with the idea of an original. When digital word processing tools coopted "document" as a digital artifact, this made an age-old word new and strange. In many ways, it also forged the foundation of the web, as Tim Berners-Lee used the architecture of the document and file system as the web's basis.[6] Taken for granted today, this decision was not at all a given, and in fact stirred controversy among a subset of detractors, who pointed precisely to the web's lack of an "original" document copy as its primary shortcoming.[7]

2.1.2 From archive to database

The word "archive" follows this tradition, but it has exploded even beyond its new digital meaning. Michel Foucault uses the term to refer to "systems of statements" that consist of the "history of ideas," the entirety of sayable things and their referents.[8] Foucault's epistemological archive subsumes both the stuffy room and the digital database into itself. While an archive of books, manuscripts or newspapers is not *the* archive in a Foucauldian sense, the word constantly carries this weight in research and literature about digital history.

Jacques Derrida tracks the origin of the word in his essay "Archive Fever," noting that it comes from the Greek *arkhe*, meaning at once "commencement" and "commandment."[9] The commencement is the original moment that every act of archiving attempts to capture and return to, while the commandment represents that archive's authority to singularly classify an object and determine its contextual future. Derrida treats Freud's archives as his case study, highlighting the archival moments in the act of archiving itself, and the recursive nature of storage. Here Derrida is working with

[6] Tim Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (HarperBusiness, 2000).

[7] See, e.g., Jaron Lanier, *Who Owns the Future?* (New York: Simon & Schuster, 2013), Chapter 18; Theodor H. Nelson, "Ted Nelson's Computer Paradigm, Expressed as One-Liners," 1999, accessed April 19, 2015, `http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html`.

[8] Michel Foucault, *Archaeology of Knowledge* (London: Tavistock, 1972), 128-129, 137.

[9] Jacques Derrida, "Archive Fever: A Freudian Impression," trans. Eric Prenowitz, *Diacritics* 25, no. 2 (July 1, 1995): 9.

but complicating Foucault's definition; his archives are more literal, but he still uses the singular "archive" to refer to history and its origins.

The *Oxford English Dictionary* defines *archive* as "A place in which public records or other important historic documents are kept," and as "A historical record or document so preserved," while making note of its figurative uses; even this definition implies archives containing archives. The critical and theoretical approaches to archives rely on a reframing of the original term, layering the word with additional meanings and highlighting the philosophical weight associated with collections and stores. So is the archive literal, digital, or figurative? What size and shape does it take? Does it represent an individual's memory, or collective history?

The term shifts based on the *shape* and *scope* of its referent. An archive can be personal, institutional/collective, or universal. Despite the vast difference between, say, a student's bookshelf and the entirety of the World Wide Web, each of these aggregations of information can be figuratively and colloquially considered an archive. Archives morph, connect with, and contain one another. Since the archive evokes all of these scopes and practices, the word expands and contracts in meaning.

An archive always has a border, a point at which the collection stops. It stops on both sides: the *micro* level (what is the smallest unit of information that it indexes—a book, an image, a single letter?) and the *macro* level (what information or metadata does this archive not include?). That an archive has a limit is inevitable, and useful; a limitless archive would be impossible and unhelpful, akin to Borges' exact one-to-one map of the world.[10] But ideally, an archive can expand and contract, as needed, on both scales, satisfying both the casual browser and the dedicated researcher. If a researcher asks a question too specific for any one document, the archive could break down the document into its constituent parts; if a user is browsing beyond an archive's boundaries, it might talk to other archives that have the answer. The ideal archive is elastic, polymorphous, and adaptable.

Aside from the borders of archives, there are also borders *in* archives. Traditional, physical archives are divided into sections, stacks and rows, each with dedicated

[10] Jorge Luis Borges, *Collected Fictions*, trans. Andrew Hurley (New York: Penguin, 1999), 325.

classification schemes that keep books in their right place. Librarians and experts draw and maintain these borders, while others need to speak their language to find their way. Today's digital archives are not so neatly or hierarchically drawn. Derrida uses the border metaphor to describe the recent diffusion of archives: "the limits, the borders, and the distinctions have been shaken by an earthquake from which no classificational concept and no implementation of the archive can be sheltered."[11] Claire Waterton, citing Michael Taussig, likewise suggests that the border zone is "currently expanding, proliferating, becoming permeated by itself."[12] Reflecting the postmodern skepticism towards standard categories and hierarchies, the linked archive reshapes itself into any categorization scheme that a user or collective might define.

These complications make any singular definition of *archive* impossible. Generally speaking, I will use the term to refer to any collection or repository of items that offers interfaces for those items' organization and discovery, with the aim of helping people find information, structure ideas, and do research. This includes the systems surrounding collection itself—organizational, structural, and sociocultural. To put it in Lev Manovich's terms, "data structures and algorithms are two halves of the ontology of the world according to a computer."[13] I am interested in an archive's data structures (specifically with regard to its items' indexing, metadata, and organizational schemes), as well as its algorithms (the ways to organize, aggregate, repurpose, and present these items to the user).

For my purposes, the "archive" is similar to the concept of the "database" as considered by Manovich and others. The distinctions between these two terms have been debated extensively, and some scholars have treated traditional, pre-digital archives as databases.[14] I intend to reverse this anachronism, and treat databases as archives.

[11] Derrida, "Archive Fever," 11.

[12] Claire Waterton, "Experimenting with the Archive: STS-ers As Analysts and Co-constructors of Databases and Other Archival Forms," *Science, Technology & Human Values* 35, no. 5 (September 1, 2010): 649.

[13] Lev Manovich, "Database as Symbolic Form," *Convergence* 5, no. 2 (June 1, 1999): 84.

[14] See, e.g., Marlene Manoff, "Archive and Database as Metaphor: Theorizing the Historical Record," *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398; Jonathan Freedman et al., "Responses to Ed Folsom's "Database as Genre: The Epic Transformation of Archives"," *PMLA* 122, no. 5 (October 2007): 1580–1612; Belinda Barnet, "Pack-rat or Amnesiac? Memory, the archive and the birth of the Internet," *Continuum: Journal of Media & Cultural Studies* 15, no. 2 (July 2001):

I do this in part to hone my focus onto the collections and systems that provide access to personal, institutional, and historical records for research and inquiry. The archive, unlike the database, pledges perpetual storage, future access, and availability; while a database can contain continually shifting information, a traditional archive aims to preserve original, singular documents. As Marlene Manoff says, "The notion of the archive is useful in theorizing the digital precisely because it carries within it both the ideal of preserving collective memory and the reality of its impossibility."[15] Following Jerome McGann's insights, I see the database as a technical instrument used for the structuring and enabling of archives; it is not the archive itself.[16]

Like McGann and Manoff, I also use the word to emphasize a lineage. Today's information management tools continue to inherit many ideas and techniques from traditional archives and note-taking systems—a fact that "database" doesn't emphasize. These systems are always evolving and built atop one another; traces of old technologies are present in current systems. In this sense, many of the applications we use today are systems for organizing and managing personal, institutional, and public archives: search and social media platforms (Google, Twitter), note-taking and citation tools (Evernote, Zotero), content management systems (WordPress, Drupal), ideation and productivity software (Trello, Basecamp), media repositories, codebases, and so on. Archives are also deeply embedded within and linked to one another via APIs, further complicating the picture.[17]

The rise of knowledge work has brought more and larger archives, and new computational capabilities have brought a new *kind* of archive with new affordances. We use these archives for both professional and personal ends; whether we read social media and blog posts, create and collaborate on workplace documents, or use data-driven methods to track our health and habits, we are interacting with archives. Jussi Parikka suggests that "we are all miniarchivists ourselves," calling the informa-

217–231.

[15] Manoff, "Archive and Database as Metaphor," 396.

[16] Freedman et al., "Responses to Ed Folsom's "Database as Genre"," 1588.

[17] APIs (Application Programming Interfaces) allow online services to connect to one another for various features and services; one basic example might be a Facebook "like" button on a news article.

tion society an "information management society."[18] Belinda Barnet considers it a "pack-rat" mentality, and Derrida succinctly titles the phenomenon "archive fever." Viktor Schoenberger writes that by default, the web saves rather than forgets; this is a crucial shift that has occurred on both a personal and a collective scale.[19]

2.2 The social life of content

Along with words like archive, document, data, and information, I am interested in the word "content" to describe creative works or texts residing on the web. It is a word that is bound to encounter derision, whether from "content creators" (never self-defined as such), information theorists or media scholars. In a 2014 talk at MIT, Henry Jenkins referenced the word's derivation from the Latin *contentum*, meaning "a thing contained."[20] Doc Searls frequently criticizes the term for its ties to marketing, implying a one-way web where content is a catchall term for anything that can be packaged, sold, and consumed online.[21]

Another, perhaps friendlier way to describe content is as a "link." When a friend emails you an article, he is less likely to say "here's a piece of content" than "here's a link," implying sharing and networking from the outset. Where content implies a container (containment), a link implies a connection (expansion), promising to break free from the contained information. Looking at the link favorably, if a publisher adds a hyperlink to an article, it purports to show not only erudition (the publisher has read and vetted the content within), but also altruism (the publisher is helping the content creator and the user reach one another). But here, the link surrounds

[18] Jussi Parikka, "Archival Media Theory: An Introduction to Wolfgang Ernst's Media Archaeology," in *Digital Memory and the Archive*, by Wolfgang Ernst (Minneapolis: University of Minnesota Press, 2012), 2.

[19] Viktor Schoenberger, *Useful Void: The Art of Forgetting in the Age of Ubiquitous Computing* Working Paper RWP07-022 (Cambridge, MA: John F. Kennedy School of Government, Harvard University, April 2007), accessed April 21, 2015, http://ksgnotes1.harvard.edu/Research/wpaper.nsf/rwp/RWP07-022.

[20] Andrew Whitacre, "Podcast: Henry Jenkins Returns," CMS/W, March 10, 2014, accessed April 20, 2015, http://cmsw.mit.edu/henry-jenkins-returns/.

[21] Doc Searls, "Earth to Mozilla: Come Back Home," Doc Searls Weblog, April 12, 2014, accessed May 5, 2014, https://blogs.law.harvard.edu/doc/2014/04/12/earth-to-mozilla-come-back-to-us/.

the content. In effect, it is the original container, adding the first layer of context to the content, but diluting its core in the process. In studying the origins of the link's structure and the web's infrastructural qualities, we find many ways in which the web's very structure, as well as the creators, indexers, and archivists that work with content, acts as a containing and homogenizing force. The web's simultaneous operations of containment and connection make online media more legible as a networked, aggregated mass rather than a set of distinct and multifarious texts, more often treated with macro-level analyses rather than smaller-scale textual readings.

But there is value in honing in on the smaller aspects of online content. A single link accumulates layers of context and connection at each stage of its life on the web. One could view the result as if in concentric circles, as a series of pointers and wrappers around the original source. Therefore, the original text (whether itself text, image, video, or a combination thereof) finds itself embedded under several layers of representation. The first such wrapper is the URL (Uniform Resource Locator), which serves to represent multimedia in a homogenous piece of text that renders everything "uniform." From there, several layers of representation are placed on top of it, starting with the hyperlink (an HTML element that forges connections between documents). An HTML document is a hybrid object; links contain links, and content is consecutively embedded in secondary sources, social media platforms, search results and archives. At each stage, the content acquires new metadata created by both individuals and machines, that indelibly affects our understanding of the original source. These varying layers of context and containment reflect new modes of information organization and storage, and ultimately affect the ways in which we organize and represent multimedia works.

These stages mark a sort of biography of online content, following Igor Kopytoff's "biography of things," which focuses on the transitional events that mark a thing's history. Kopytoff complicates the idea of any single, indelible act of categorization on an object—instead, an object is "classified and reclassified into culturally constituted categories."[22] This especially lends itself to digital content as well due to its

[22] Igor Kopytoff, "The Cultural Biography of Things: Commoditization as Process," in *The Social*

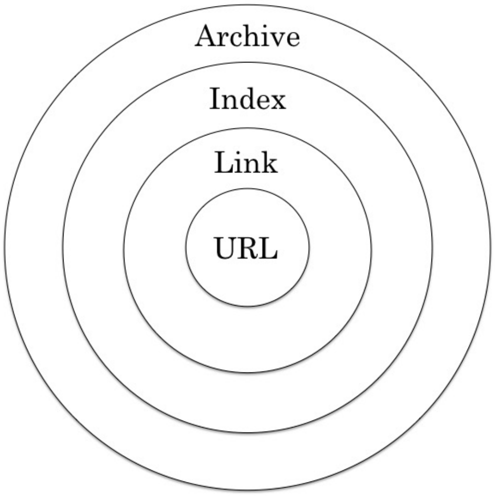


Figure 2-1: The layers of containment.

ephemerable and duplicable nature; for instance, a Flickr image might turn up in far-reaching corners of the web, activated from archives via various searches. The photo in figure 2-2 lives in eight different Flickr groups and one album, such as "USN Blue Angels," "Airplanes," and "Curbed SF."[23] One user might find it when searching for photos of the Blue Angels (part of the photo's content) or the Golden Gate Bridge (both content and location), while another could simply be searching for photos of fog, or spectacular photos (part of the photo's effect). Still another user could be looking for photos taken on a Canon EOS 7D, or photos with more than 2,000 views. Such categories can be dynamically conceived and created, and digital objects often carry the potential for a nearly limitless number of them.

The fact that digital objects carry this contextual metadata leads to a detailed history; an old photo of your ancestors won't tell you the exact date, time, and coordinates of its inception, or the camera it was taken on. All the same, that old photo carries a physical trace of reality, a *grain* of history that an archaeologist or forensic investigator might be able to decipher. Meanwhile, a digital object's origins can be forged or erased with relative ease; I could take a screenshot of the Flickr photo

Life of Things: Commodities in Cultural Perspective, ed. Arjun Appadurai (Cambridge University Press, 1986), 68.

[23] Bhautik Joshi, "Fleet Week SF 2014: swoop," Flickr, October 11, 2014, accessed April 24, 2015, `https://www.flickr.com/photos/captin_nod/15324633898/`.

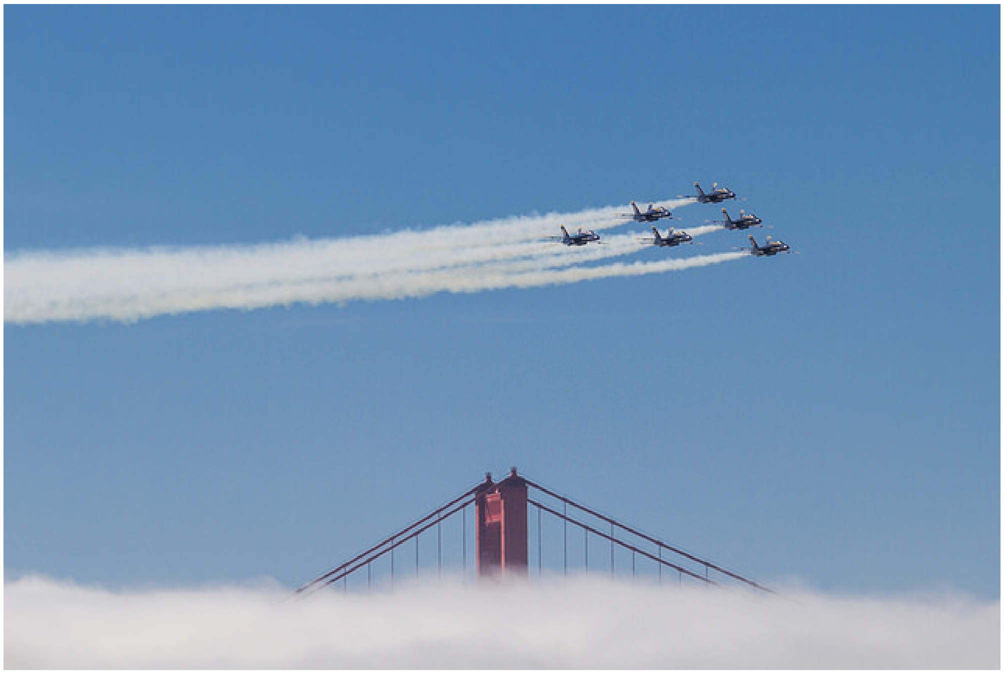


Figure 2-2: A sample photograph on Flickr. Photo by Bhautik Joshi and used under a Creative Commons license.

and give it a fresh new history, as a completely “different” object, claiming that it was taken years earlier, or with a different camera. The ephemerality of digital content and malleability of its metadata leads to many ontological and practical dilemmas; the former is explored by the field of media archaeology, while the latter forms the basis of media search and metadata verification services like Storyful and TinEye. Digital content’s history is both abundant and ephemeral, both given and plucked.

A biographical approach to digital content nods both to media archaeology and the contingencies of classification proposed by Kopytoff: “what we usually refer to as ‘structure’ lies between the heterogeneity of too much splitting and the homogeneity of too much lumping.”[24] Digital content lends itself well to these notions of shifting identity. The stock photograph is a rich example of content that is continuously recontextualized; the Blue Angels photo could turn up in a news article about the angels, or about the Golden Gate Bridge, or about the Canon EOS 7D. Its metadata forms its history; at various points of its “life” it has been tagged, grouped, searched for, resized, or recolored; some of this history has traveled with it, written by both

[24] Kopytoff, “The Cultural Biography of Things,” 70.

humans and machines, while some of it was deleted or never captured, now irretrievable. Such a rich social history with myriad possible uses cannot be predicted or summed up by simple, static categories and tags.

Geoffrey Bowker and Susan Leigh Star emphasize the perils of "splitting and lumping" in their book *Sorting Things Out: Classification and its Consequences.* Tracing the history of classification as it is used formally (in standards) and informally (in the words, framings and mental models we are perpetually forming), they argue that each act of classification affects the classification system itself, and future classifications in turn. At its most abstract level, classification is the application of language to reality; whether you are calling a kitten "cute" or a person "male," you are framing the subject at hand and privileging certain discourses and interpretations over others. Taken at scale, these acts shape our epistemology and understanding of the world. Bowker and Star see infrastructures and standards as intimately interlinked; each one inherits the values and inertias of the systems around it. They point to the 200 standards imposed and enacted when a person sends an email; these standards interact and depend on one another in important ways.[25] *Sorting Things Out* highlights many of the problems and limits with traditional classification, and suggests frameworks for rendering it more dynamic and responsive.

As such, any attempt to trace an object like a stock photo is also doubling as an analysis of the whole system in place. Content is never just content, and to describe it is also to describe its containers. This notion of embeddedness also points to actor-network theory (ANT) and its treatment of objects and the social interactions around them. Beyond content and people, there is another type of actor in this network too; the search and sorting algorithms run by Google, Facebook or other aggregators and platforms.

[25] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT Press, August 28, 2000), 7.

2.2.1 The URL

As Tim Berners-Lee tells it, the Uniform Resource Locator, or URL, "is the most fundamental innovation of the web."[26] Sometimes called the Uniform Resource Identifier or URI, it allows any address to link directly to any other address simply by pointing to it. But the URL itself is a remediation of old standards and practices. It mimics the file folders on our home computers (an intentional decision, so it could be understood and adopted quickly), implying a hierarchical, document-based structure. Interpreted hierarchically, the URL can be seen as an address, pointing us to increasingly specific locations until we arrive at the document in question. The virtual space of the web here seems to mimic physical space in the world, suggesting that one can find a document down a certain "path" under a certain "domain."[27] By turning all rich multimedia into "uniform resources," the URL is a homogenizing force, encoding all content as a textual address and turning it into a reference rather than an experience or narrative.

URLs are not created equal, however, and savvy web users can read a great deal of information in this set of text. A ".org" top-level domain (TLD), for instance, might imply a nonprofit or philanthropic institution where a ".com" connotes a business. A long, seemingly obfuscated URL might contain spyware or viruses. A URL that ends with ".html" or ".jpg" will probably be a picture, but one that ends with "/users?friendrequest=true" is more likely to be telling a social media site to request friendship with another user. Indeed, a URL could yield no content and simply trigger a piece of code, allowing any arbitrary action. Moreover, even documents are subject to change, and the web has no built-in way to track content's erasures and additions. In other words, the "Uniform Resource Locator" is not necessarily uniform, nor is it necessarily a resource. Even this vague, homogenizing definition does not hold up.

Eszter Hargittai points to the need for technical expertise in order to properly understand a URL and the implications behind it.[28] It is easier for a user with less

[26] Berners-Lee, *Weaving the Web*, 39.

[27] See section 3.1 for a historical lens on spatialization in archives.

[28] Eszter Hargittai, "The Role of Expertise in Navigating Links of Influence," in *The Hyperlinked Society* (Ann Arbor, MI: University of Michigan Press, May 23, 2008), 85–103.

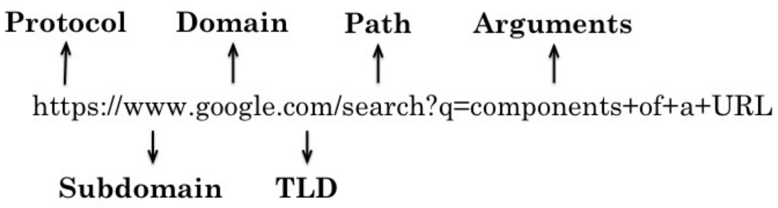


Figure 2-3: The components of a URL.

experience with the Internet to be duped by a phony URL that installs spyware or viruses; it is also more difficult for such users to find the content that they need when navigating through links. Further complicating the picture, neither the URL nor the link provide any information concerning motivation or access. In a URL, sensitive documents or paywalled media sites appear the same as free and open information. With the exception of the underlying protocol ("https" versus "http"), a URL's overall security or level of access cannot be known without visiting it. Some online publications have constructed paywalls that can easily be "broken" through a simple reordering of the URL, or by visiting it from a Google search or a new browser.[29] All of these mechanics and limitations favor information access through technical knowledge, and serve as a barrier to understanding and retrieving information for those who are in the dark.

The URL's fuzzy standards and conventions complicate any overarching attempts to understand or map the web as a whole through its URLs. Many large-scale network analyses of links (discussed at greater length in section 5.1) rely on URLs to gain insight about the content within, because it is often too technically taxing to gather richer metadata. This is despite the fact that the URL is an arbitrary reference. For instance, an organization might register a site under a variety of domains, or it might register a ".org" website despite being a for-profit operation. We're all more likely to trust a ".com" or ".org" than a ".biz"—and there's little doubt that Google takes

[29] See, e.g., Joshua Benton, "That was quick: Four lines of code is all it takes for The New York Times' paywall to come tumbling down," Nieman Journalism Lab, March 21, 2011, accessed April 20, 2015, http://www.niemanlab.org/2011/03/that-was-quick-four-lines-of-code-is-all-it-takes-for-the-new-york-times-paywall-to-come-tumbling-down-2/; Jack Smith IV, "Here's How To Get Around the Paywalls of the New York Times, Wall Street Journal and More," Observer, January 8, 2015, accessed April 20, 2015, http://observer.com/2015/01/heres-how-to-get-around-the-paywalls-of-the-new-york-times-wall-street-journal-and-more/.

domain names into account when ranking search results.[30]

Suffixes like ".es", ".uk" or ".tv" belong to Spain, the United Kingdom and Tuvalu, respectively. Some country-code TLDs restrict domain ownership to citizens of that country, while others are available to all—for a price. These varying practices make it difficult to uniformly map international communication networks by analyzing the link flows between TLDs; for instance, between Brazil and Germany, or between developing countries and Western hubs.[31] This is in part because it is relatively easy to do such analyses at a large scale, where all you need is a list of URLs and pointers. But the URL is not so easily mapped. The ".tv" domain, for instance, was sold by the small Polynesian nation Tuvalu to a Verisign company; Tuvalu maintains a 20 percent stake and a $1-million quarterly payout.[32] Hundreds of new generic (non-country affiliated) TLDs are entering the marketplace in 2014-2015, and some, like ".sucks" have been blamed for extortion, charging companies as much as $25,000 per year to protect their brand.[33] Underneath these seemingly simple and technical URLs lie complex commercial and financial transactions.

URL "shorteners" such as those employed by the New York-based company bit.ly (whose TLD would otherwise suggest that it comes from Lybia) likewise add additional layers between user and content, and further obfuscate the final destination. With a URL shortener, a small and innocuous domain (such as "bit.ly/a423e56") can take a user to any corner of the web, whether at the highest level (think "google.com") or its most specific (like "pbs.twimg.com/media/Bm6QZAGCQAADEOk.png"). Short-

[30] See Chris Liversidge, "What's The Real Value Of Local TLDs For SEO?," Search Engine Land, December 4, 2012, accessed April 20, 2015, `http://searchengineland.com/whats-the-real-value-of-local-tlds-for-seo-140519`.

[31] See, e.g., Chung Joo Chung, George A. Barnett, and Han Woo Park, "Inferring international dotcom Web communities by link and content analysis," *Quality & Quantity* 48, no. 2 (April 3, 2013): 1117–1133; Suely Fragoso, "Understanding links: Web Science and hyperlink studies at macro, meso and micro-levels," *New Review of Hypermedia and Multimedia* 17, no. 2 (2011): 163–198; Itai Himelboim, "The International Network Structure of News Media: An Analysis of Hyperlinks Usage in News Web sites," *Journal of Broadcasting & Electronic Media* 54, no. 3 (August 17, 2010): 373–390.

[32] Damien Cave, "I want my own .tv," Salon, July 24, 2000, accessed April 20, 2015, `http://www.salon.com/2000/07/24/dot_tv/`.

[33] Yuki Noguchi, "A New Internet Domain: Extortion Or Free Speech?," NPR, April 7, 2015, accessed April 20, 2015, `http://www.npr.org/blogs/alltechconsidered/2015/04/07/397886748/a-new-internet-domain-extortion-or-free-speech`.

ened URLs no longer pretend to mimic the spatial world or even most personal computer filesystems; we have replicated and obfuscated the URL to the extent that any sort of uniformity or direction is impossible. Anne Helmond calls this phenomenon the "algorithmization" of the link, a retracement of its role from navigational device to analytical instrument.[34]

2.2.2 The link

A link is more than just a URL: it wraps the URL in an HTML element that allows it to be quickly accessed from another page, containing additional mechanics and context. Without links, the web would just be a series of disconnected nodes; with links, the web gains edges, and becomes a network.

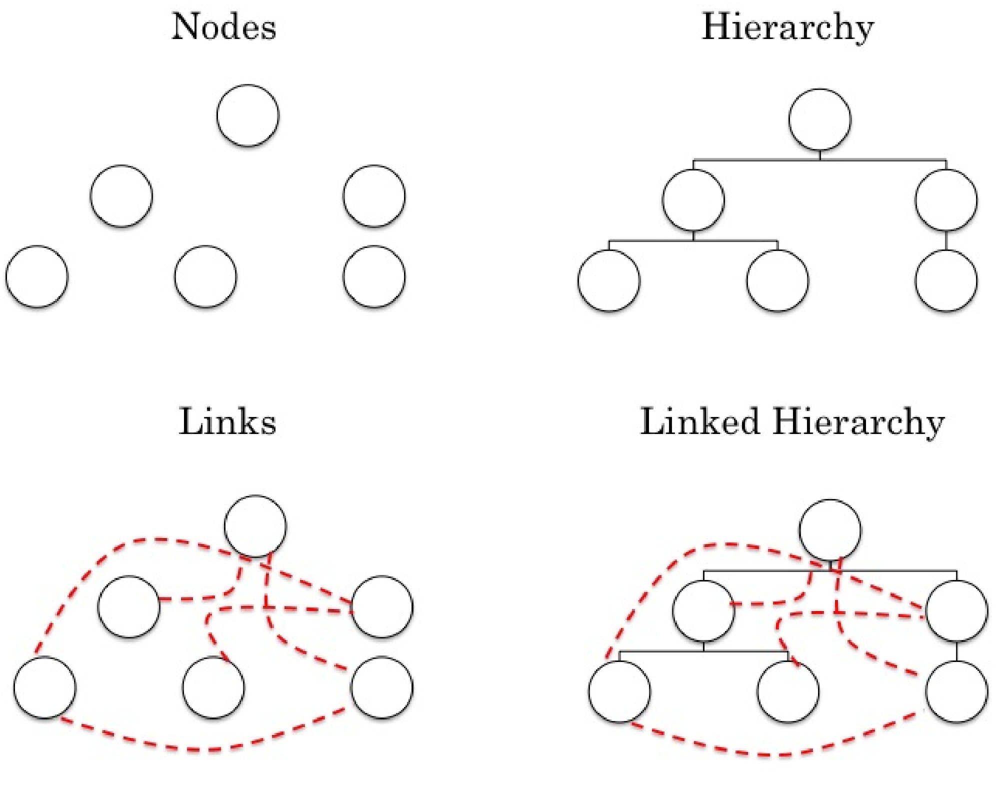


Figure 2-4: Schematic of methods for connecting content, from nodes to networks.

[34] Anne Helmond, "The Algorithmization of the Hyperlink," *Computational Culture* 3 (November 2013).

Bowker and Star suggest that links have the power to classify without any human agency or intervention, and this phenomenon forms the basis of this section: "Every link in hypertext creates a category. That is, it reflects some judgment about two or more objects: they are the same, or alike, or functionally linked, or linked as part of an unfolding series."[35] The agency shift here is important; the *link* is creating a category, rather than a human actor relying on language. Bowker and Star are not the only ones to cede agency to the link, and many disputes and debates occur over links; even in 2002, Jill Walker asserted that "links have value and *they give power*."[36] In many ways, the link is the battlefield for the political economy of the web, serving as a sort of digital currency and object of value exchange.

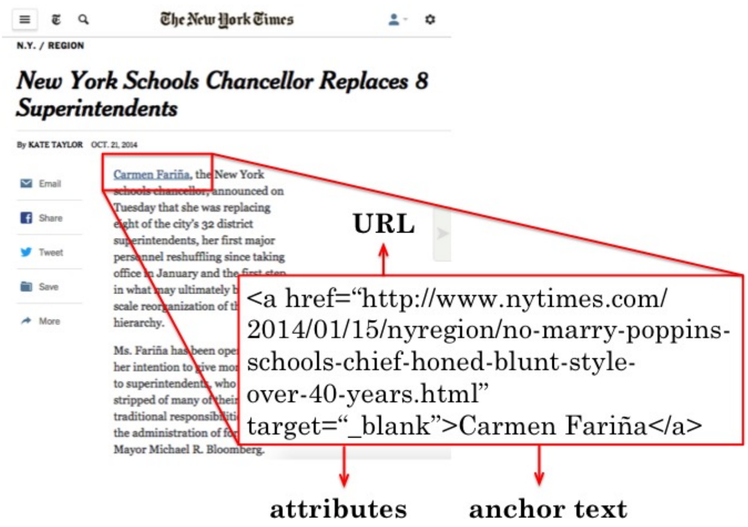
All the same, the link is a seemingly innocuous object. We usually consider it taking the form of a blue, underlined piece of text on a webpage (under the hood it is known as an anchor tag—the string "<a href>. . . </a>" and everything in between—in an HTML document). Hovering over the link reveals the true URL behind the blue text. Clicking on the link turns the object into a mechanic, leading a user down a rabbit hole of subsequent destinations and redirects (all employing some dozens of standards) before landing on the target destination—back to the URL. The URL is only one attribute of the link, along with others that determine, for instance, whether to open the link in a new tab or window—so in a literal sense, the link contains the URL.

The link is forever associated with (and perhaps plagued by) the footnote. Theodor Nelson's hypertext manifesto *Computer Lib/Dream Machines* praises the screen for permitting "footnotes on footnotes on footnotes,"[37] and Berners-Lee's web takes the traditional citation as inspiration. Nelson belies himself by consistently contrasting hyperlinks with footnotes; in some senses, one cannot escape being a remediation of the other. But the link's readable text—its manifestation in a browser, known as the anchor text—adds another layer of semiotic containment and enrichment to the

[35] Bowker and Star, *Sorting Things Out*, 7.

[36] Jill Walker, "Links and Power: The Political Economy of Linking on the Web," in *ACM Hypertext Conference* (Baltimore, MD, June 2002).

[37] Theodor H. Nelson, *Computer Lib / Dream Machines* (Self-published, 1974), DM19.



N.Y. / REGION

New York Schools Chancellor Replaces 8 Superintendents

By KATE TAYLOR OCT. 21, 2014

Carmen Fariña, the New York schools chancellor, announced on Tuesday that she was replacing eight of the city's 32 district superintendents, her first major personnel reshuffling since taking office in January and the first step in what may ultimately b
scale reorganization of th
hierarchy.

Ms. Fariña has been ope
her intention to give mor
to superintendents, who
stripped of many of their
traditional responsibiliti
the administration of for
Mayor Michael R. Bloomberg.

```
<a href="http://www.nytimes.com/
2014/01/15/nyregion/no-marry-poppins-
schools-chief-honed-blunt-style-
over-40-years.html"
target="_blank">Carmen Fariña</a>
```

Figure 2-5: The components of a link.

original content. The "jumpable interconnections" that Nelson envisions are built into the fabric of the writing rather than set aside like a footnote.

Like any sign, the anchor text has no innate relationship to its target. The many flexible uses of the hyperlink emerge at different semiotic layers; when a link says "click here" as opposed simply linking the text, it may be forming an indexical rather than symbolic relationship to the target. When a link's text is identical to its address, like "http://www.google.com," it purports to be more transparent, but there is nothing stopping someone from putting a completely different address into the anchor text. This disconnect between anchor text and target facilitates online behaviors both nefarious and playful, ranging from email phishing scams to "rickrolling."[38]

Many studies have attempted to glean insight from the link by assuming, like Bowker and Star, that links create categories. On one hand, it seems extremely liberating to sidestep the ontological dilemma of what that category *is*, and simply treat it as a raw signal. I see this ability as the basis for much of the revolutionary rhetoric of the web and the power of networks. This also forms the basis of link-oriented archival practices that I will discuss further in section 5.1. On the other

[38] "Rickrolling" is an online meme wherein an internet user purports to send another person a relevant link, but it redirects instead to Rick Astley's 1987 hit song "Never Gonna Give You Up."

hand, the lack of relation between text and target seems to point to problems with this approach: a sign is not the same thing as a signal. While some search engines and classifiers analyze textual aspects of a link (such as its anchor text, or the surrounding text in the same paragraph or post), few large-scale studies take the text into account or treat linking within a semiotic framework.

The link and its anchor text are increasingly used as a creative textual device amongst journalists, who are continuing to discover its uses. This may be best exemplified by bloggers and aggregators, who often summarize a larger topic or debate while seamlessly incorporating hyperlinks for attribution, context, and humor. Hyperlink usage may be changing too because of the increase in online literacy; the hyperlink is part of the language of the web, which users understand more and more. The default mechanic of the hyperlink has changed from a full-page refresh to opening a new tab, facilitating a new form of linking "behind" rather than "in front of" the source text. These creative and technical advances are informed by one another, and a longitudinal study of anchor text usage would help to determine the dynamic evolution of linking practices.

Instead, most studies simply take an aggregate view of link sharing, treating each connection as equal regardless of context. With rare exceptions (such as the "nofollow" attribute, which tells Google's crawlers not to follow the link), anyone who shares an article inevitably, and perhaps inadvertently, raises the article's profile and algorithmic rank. Algorithms might therefore prefer controversial links rather than universally liked, substantial, or thought-provoking ones. This could create incentives for publishers to use unnecessarily inflammatory or partisan language, with the assumption that despite how users feel about the content, they will certainly click on it, and possibly share it. Mark Coddington places a cultural studies frame around this phenomenon, where even negative linking implicitly defines the source "as part of the text's preferred reading."[39] However, some tech-savvy users have adopted techniques like "subtweeting" and screenshots of text to avoid easy detection; scholars like

[39] Mark Coddington, "Building Frames Link by Link: The Linking Practices of Blogs and News Sites," *International Journal of Communication* 6 (July 16, 2012): 20.

Zeynep Tufekci and Helen Nissenbaum have written about such intentional obfuscation, especially in the context of political upheaval and surveillance.[40]

The many cultural and contextual nuances behind a link or reference are too complex for a computer to discern. This limitation is apparent to Berners-Lee, who has in recent years championed the Semantic Web as a way to make the web more structured and machine-readable. The Semantic Web allows for links themselves to be annotated and queried, so that, for example, we could search for "users who disagreed with this article" and not just "users who linked to this article."This carries great promise not only for a machine-readable web but a new order of linkage and network formation. The W3C (the standards organization for the web) maintains appropriately revolutionary rhetoric around the Semantic Web, and has tried out scores of marketing terms in its efforts. It alternately envisions a "web of data" (rather than documents), a "Giant Global Graph," and "Web 3.0," a particularly telling attempt to couch the Semantic Web as the inevitable next step of forward progress. However, while linked data has been useful in smaller-scale initiatives, the Semantic Web movement has progressed very slowly. It also brings its own problems; while a web of documents is one level removed from the data itself (and therefore more difficult for machines to read), at least it keeps the source context intact. The Semantic Web also forces users to choose particular sets of ontologies, hierarchies and categorization schemes.

Another alternative to the web's form of linkage comes from Ted Nelson, a longtime critic of the web's architecture, whom I will discuss at greater length in section 3.3.1. As one of the original hypertext visionaries, his scheme, called Project Xanadu, floundered for decades, and has never truly been built in the way that he envisioned. When critics suggested that Xanadu was the first failed web, Nelson bristled: "HTML is precisely what we were trying to PREVENT—ever-breaking links, links going out-

[40] Zeynep Tufekci, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," *arXiv:1403.7400 [physics]* (March 28, 2014); Ethan Zuckerman, "Helen Nissenbaum on Ad Nauseum, resistance through obfuscation, and weapons of the weak," My Heart's in Accra, October 6, 2014, accessed April 20, 2015, http://www.ethanzuckerman.com/blog/2014/10/06/helen-nissenbaum-on-ad-nauseum-resistance-through-obfuscation-and-weapons-of-the-weak/.

ward only, quotes you can't follow to their origins, no version management, no rights management."[41] Xanadu's most important feature, absent from the web, is the two-way link; when one document referenced another, the target document referred back to the original in turn. The hyperlink on the web, for all its flexibility, does not escape the trappings of the footnote in this single, very important way. Like footnotes, links always move backward, and given the lack of a canonical URL on the web (another of its limitations, which the URL-shortening phenomenon compounds), finding all the citations for a single document is next to impossible. The NewsLynx project has documented its challenges in doing just that, while Jaron Lanier believes this simple omission has profoundly affected culture and economics, which forms a cornerstone of his 2013 book *Who Owns the Future?*[42]

But in the absence of replacing or reconfiguring the web's current structure, the one-way, semantically meaningless link remains the web's primary organizational scheme, and the "click" remains the proxy for attention and engagement. Clicking on a link is not only a navigational mechanic; it is a signal of intent and interest, which influences algorithmic decisions and other readers in turn. It is also often a financial transaction between unseen actors; each link clicked and page viewed is a new "impression," causing money to change hands between content distributors and advertisers. This has in turn changed the aforementioned semiotics of the link, and the meaning of its anchor text.

For instance, there has been much controversy surrounding the news headline in the hyperlinked age. Consider a story about the foundations that support tuberculosis research. Where a traditional headline might read "The Global Fight Against Tuberculosis," a more recent one is more apt to say, "It Kills 3 People a Minute, but That's Not Stopping This Group of Heroes."[43] The headline is colloquially known as

[41] Nelson, "Ted Nelson's Computer Paradigm, Expressed as One-Liners."

[42] Lanier, *Who Owns the Future?*, Ch. 18; Brian Abelson, Stijn Debrouwere, and Michael Keller, "Hyper-compensation: Ted Nelson and the impact of journalism," Tow Center for Digital Journalism, August 6, 2014, accessed April 19, 2015, `http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/`.

[43] Alana Karsch, "It Kills 3 People A Minute, But That's Not Stopping This Group Of Heroes," Upworthy, May 5, 2014, accessed April 20, 2015, `http://www.upworthy.com/it-kills-3-people-a-minute-but-thats-not-stopping-this-group-of-heroes-2`.

"click bait," playing to a user's innate curiosity (Atlantic writer Derek Thompson calls it the "curiosity gap")[44] without telling them the substance of the article or the actors in play (tuberculosis, the victims affected, the Global Fund, the Gates Foundation, and others). These actors and the issues they are tackling are reduced to pronouns. Here even the content becomes glossed, and a click is likely to signify curiosity about what the content *is*, rather than any genuine interest in the content itself. Machines are not likely to recognize these nuances, which results in false identification of public interest and discourse. Upworthy's organizational structure is telling; the company creates no original content, but instead employs people to trawl the web, find content, and repackage it with a new headline. Upworthy has built a valuable business not by creating new content, but new containers.

2.2.3 The feed, the index

Links rarely exist in isolation, and one form that the link often takes is as part of a list or sequence. Whether it is a digest (on email), a feed (on Facebook, Twitter, or RSS), a set of search results, or a list of "related articles," users are almost always confronted with several choices for what to click on. In this section, I look at the ways in which links get aggregated, indexed, and fed to users. Indexes and feeds can allow for a higher-level view of a major topic, author, or other organizing factor, but at the expense of hiding the richness of the content within.

The aggregators, indexers, and summarizers of the web are its search engines and social media platforms, run by some of the most powerful and profitable tech companies in the world. While the content creator usually has to win the attention of the distributor, the distributor in turn must always play the aggregator's game. This is evidenced by Upworthy itself, who in December 2013 found its content potentially demoted in Facebook's algorithm with no meaningful explanation, shrinking its im-

[44] Derek Thompson, "Upworthy: I Thought This Website Was Crazy, but What Happened Next Changed Everything," *The Atlantic* (November 14, 2013), accessed April 19, 2015, `http://www.theatlantic.com/business/archive/2013/11/upworthy-i-thought-this-website-was-crazy-but-what-happened-next-changed-everything/281472/`.

mense traffic to half of its previous size.[45] Another major content distributor, the lyrics annotation website Rap Genius (now a general annotation and reference site called Genius), found its pages move in December 2013 from the top hit on Google to its seventh page, due to changes in Google's algorithm.[46] These content aggregators can move around large swaths of content (millions upon millions of interlinked pieces) via slight changes in their codebases, with no obligation to inform anyone of the reasons, or even that it is occurring.

But Google did explain its reasoning for the Rap Genius demotion, and the dispute was telling. Rap Genius had launched a "Blog Affiliate" program, which clandestinely offered to tweet out any blog post in return for links back to the Rap Genius site. In other words, Rap Genius was engaging in SEO (Search Engine Optimization) spam, attempting to falsely boost its search rankings by asking bloggers to post unrelated links back to their site. This is one high-profile example of what many smaller players do every day in order to keep their businesses alive: game Google's algorithm in order to bolster their search rankings. SEO is, in effect, an entire industry built on gaming links.

This works because Google's PageRank algorithm is primarily derived from who is linking to whom. Their link-based classification scheme is part of what made them the dominant information provider that they are today. But as soon as links gained algorithmic value, advertisers and spammers began to exploit them, inserting links not for their usefulness or relation to the text, but to improve their pages' search rankings. Moreover, many website hacks and attacks occur merely in order to insert hidden links on the targeted sites. In the process, Google has had to remain one step ahead of the advertisers, with the link as the battlefield, influencing the web and changing its structure in turn. But this battle has mostly been played out by machines, which are responsible for a substantial amount of the links created—as well

[45] Nicholas Carlson, "Upworthy Traffic Gets Crushed," Business Insider, February 10, 2014, accessed April 20, 2015, `http://www.businessinsider.com/facebook-changed-how-the-news-feed-works--and-huge-website-upworthy-suddenly-shrank-in-half-2014-2`.

[46] Josh Constine, "Google Destroys Rap Genius' Search Rankings As Punishment for SEO Spam, but Resolution in Progress," TechCrunch, December 25, 2013, accessed April 20, 2015, `http://techcrunch.com/2013/12/25/google-rap-genius/`.

as the links browsed and followed—on the web. Besides a generic, easily replaceable piece of metadata in a web request, it is extremely difficult to tell whether a website request is coming from a human or a machine.

In Google's published PageRank paper, Sergey Brin and Larry Page provide a curious "intuitive justification" for their algorithm that seems to conflate human and machine:

> PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a Web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank.[47]

This is a very strange user indeed, assumed to be easily "bored," distracted, and clicking on links at random. Moreover, this was an assumed user in 1998, and the "model of user behavior" must undoubtedly be changing as the web's capabilities and browsing habits change (and indeed, Google's signals have changed in response, though links remain crucial).

While links are shared for a variety of reasons—some of them more nefarious than others—the blogging and tweeting culture of "Web 2.0" holds to the principle of link sharing for mutual interest and benefit. This was first noticeable on the "blogroll," a list of other blogs that a blogger might recommend, usually presented as links in the blog's sidebar. Social media and blogging sites in particular are places of transactional link exchange, with implicit conventions and expectations beneath each link or like. These link exchanges solidify existing networks of bloggers and content creators, perhaps galvanizing the network but at the risk of collapsing into "filter bubbles." Many studies of links have traced political homophily, public debate, blogs and global flows of information; if we take these at face value and treat hyperlink usage as a proxy for importance, impact, and communication, then link-sharing can turn conversation

[47] Sergey Brin and Lawrence Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," in *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7 (Amsterdam: Elsevier Science Publishers B. V., 1998), 110.

inward, allowing searchers to see only blogs that have overtly linked to one another (blogs which, presumably, have similar views and opinions).[48] While the web may allow for a more heterogeneous group of voices to surface than in traditional media, one must still take part in link sharing, leading bloggers into already-established and tightly wound networks. This leads to what Philip Napoli refers to as the "massification" of the hyperlink: in the editorial and algorithmic decisions that determine where links are placed and directed, there is a distinctive replication of old mass media patterns.[49]

While content creators, distributors, and aggregators are locked in this battle over links, what happens to the actual user who visits a site, application or search engine? The user is presumably after content, and unless they were provided with a direct URL, they can only access it through this series of layered containers. Moreover, the content may be replicated and embedded in different contexts and myriad places around the web. The end result, when a user goes to Google to search, is often repetition. The same piece of content appears everywhere, such as a canonical image for a popular news story, meme, or theme.

2.2.4 The archive

Online content's final resting place is in the database (or archive), but the context and metadata around a digital object is still subject to perpetual change; any new link, like, click, or download updates its history. Indeed, the four contextual layers that I have offered here, while a theoretically useful framework, belies a more complex lifecycle; sometimes the content actually reaches a database (and is thus archived) before it even has a URL (and is thus accessible).

The database is a different form of container than the others, as it is in fact not truly of the web; it merely talks to it and works with it. Web users increasingly

[48] See Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia, "Data Portraits: Connecting People of Opposing Views," *arXiv:1311.4658 [cs]* (November 19, 2013).

[49] Philip Napoli, "Hyperlinking and the Forces of "Massification"," in *The Hyperlinked Society: Questioning Connections in the Digital Age*, ed. Lokman Tsui and Joseph Turow (Ann Arbor, MI: University of Michigan Press, May 23, 2008).

treat the web as a database, but there is no single database; instead there are very many, housed on servers around the world. Each of them faces a similar challenge: how to flatten and store the infinite possible contexts, networks, and signals that the web has created around each piece of content, into a format that allows a user to find it efficiently using any number of contexts. Perhaps a user is looking for everything stored in a specific time frame, a certain format, a dedicated category, or any combination thereof; in each case, the archive serves the role of storing and retrieving the information needed.

The ideal archive would anticipate any possible need from any possible user, whether they request content today or far into the future. Any signal that is left out is lost potential knowledge. So an archivist, most often associated with storing the past, also plays a crucial role in predicting and affecting the future. Derrida calls the archive "a *pledge*, and like every pledge, a token of the future.",[50] but there is no reasonable way to store every possible route through a database that a user might take; this would require infinite storage and processing power.

Seen in this way, the database is perhaps the only truly containing force on the web; the prior stages are in fact expanding contexts and meanings for each piece of content, and it is only in retrospect (through the archive) that it becomes contained. However, we cannot see the content *except* through the archive. And with the assumption that a border must be drawn through the expansive, innately borderless web, the question is where and how to draw it. Lisa Gitelman laments the way in which the archive reduces "ideas into character strings," or in the case of rich multimedia, encoded, flattened and unsearchable bits.[51] Character strings and encoded bits are devoid of context and semantic meaning. They certainly do little justice to the richness of the original content, which points to a proliferation of associated narratives.

My aim is not to suggest any overarching solution to the limitations of the archive; it is this very impulse that has often set back the work of retaining knowledge and his-

[50] Derrida, "Archive Fever," 18.

[51] Lisa Gitelman, "Response to "Algorithms, Performativity and Governability"," in *Governing Algorithms* (New York, NY, May 5, 2013).

tory. Bowker and Star point to the myriad efforts of "universal classification," dating back to the Tower of Babel, all of which have essentially failed. In order to fully recognize and remember this, they suggest the framework of "boundary infrastructures" to acknowledge and work with the limitations of traditional classification. Boundary infrastructures make use of boundary objects: "those objects that both inhabit several communities of practice and satisfy the informational requirements of each of them."[52] In practice, these objects (and the infrastructures that work with them) will maintain slightly different meanings in each community, but they are common enough to be recognizable to multiples. Boundary infrastructures emerge as more of a framework than a solution, but they rightly discourage the drive for an overarching schema for every object and community. By recognizing that no system will ever be perfect, it instead highlights the need for a loosely linked multiplicity of them. Such a framework can help when considering the structure of any archival endeavor.

Likewise, the web itself should not be universally schematized, and its content will never be singly and correctly categorized. In a sense, the proliferation of databases and motives for classification that the web provides allows for more "ways in" to the content than if the web were stored at a single endpoint. The Semantic Web is an interesting hybrid of universal centralization and networked distribution; it aims to bridge traditional taxonomy and contemporary chaos through its use of user-generated ontologies. In order for machines to understand a network, everything must be definitively categorized, but the categorization scheme itself is subject to change. Certain standards arise, but each individual or community is free to create its own form of linked data. This has allowed certain communities, most notably the medical industry, to make great use of it; if a 50-year-old smoker is complaining about shortness of breath and a fever, a doctor can ask a linked database for all diagnoses of similar patients. Linked data has also been a factor in the news industry, with many media publishers connecting to OpenCalais or *The New York Times*' Semantic API for added context. But linked data on the web has proven difficult, and the slow adoption of the Semantic Web may have to do with its reliance on ontologies. Even if

[52] Bowker and Star, *Sorting Things Out*, 297.

multiple ontologies can coexist, they are still trying to compromise the web's inherent disorder.

2.2.5 Erasure and afterlife

Content has an afterlife when it is reactivated from the archive at a client's request. Some content stays dormant indefinitely: nearly one-third of all reports on the World Bank's website have never once been downloaded.[53] While this may seem dire, it is not to say that the knowledge contained in the World Bank's documents has been utterly forgotten. The document could be distributed at another URL, or by email, or presented at a conference—the inability to track it is part of the problem. But the information is not helpful in the current format. If the World Bank's PDFs are lying dormant, they might consider using HTML, adding links and references from around the web, repackaging and remixing the data, or inviting user comments. All of these variations and annotations could help to unlock and link their data and insights.

But some content may be worthless, misleading, slanderous, detrimental, embarrassing, or outright false. Whether it's an outdated scientific fact, a politician's off-color comment, or a teen's suggestive selfie, not everything should be stored and remembered forever, as-is, without context. Samuel Arbesman's *The Half-life of Facts* emphasizes the drift in knowledge over time, reminding us that information requires constant care and upkeep to remain useful (not to mention maintaining privacy, legality, and security).[54] But how should such context be presented, and who should decide or control what is saved and stored? And how can one control a digital object's history and context after it has been replicated and remixed around the web?

Our personal and institutional archive fevers are, in some ways, understandable. Content changes and disappears on the web all the time. The web is an ephemeral stream, and you won't step into the same one twice; Barnet equates web surfing with "channel surfing."[55] As users, we know the phenomenon as the dreaded `404 NOT`

[53] James Trevino and Doerte Doemeland, *Which World Bank reports are widely read?* WPS6851 (The World Bank, May 1, 2014), 1–34.

[54] Samuel Arbesman, *The Half-life of Facts* (New York, NY: Penguin, 2013).

[55] Barnet, "Pack-rat or Amnesiac?," 217.

FOUND. Researchers led by Jonathan Zittrain found that 30-50% of links in scholarly papers and legal opinions no longer work (a phenomenon known as "link rot"), and even when they work there's no telling how much they have changed since being cited (this is "reference rot").[56] The Hiberlink project has been researching reference rot in scholarly articles, finding that as many as 70% of linked web pages were irretrievable in the form they were originally cited.[57] Similarly, the NewsDiffs project discovered that the average breaking news story is edited and updated several times at the same URL; out of 28,000 New York Times articles, they found 44% that had changed at least once, and only one in four of these changes lead to official corrections.[58]

To combat link rot, Zittrain leads the Amber project, which supports a "mirror as you link" approach: for instance, if a blogger links to another site, Amber downloads the linked page directly to the blogger's site.[59] This petrifies the document, freezing it at the time that it was cited, and offers a fallback in case the link disappears or changes. This approach can be expanded by "spidering" out to the links of the target document in turn. Spidering is the default mode of web navigation for machines; this is how Google's random surfer surfs, and how the Internet Archive saves, always limited by the structure of the link network. Meanwhile, the Archive Team aims to preserve discussion forums and old blogs. The Library of Congress saves websites by manually choosing them, often taking an "aggregate the aggregators" approach and storing large text databases, such as Project Gutenberg and Twitter.

Most of these endeavors are publicly or institutionally funded, but the case of Twitter brings up the role of media and technology companies in archiving and organizing our history. In a 2014 talk at MIT, Tarleton Gillespie emphasized the responsibilities

[56] Jonathan Zittrain, Kendra Albert, and Lawrence Lessig, *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*, SSRN Scholarly Paper ID 2329161 (Rochester, NY: Social Science Research Network, October 1, 2013).

[57] "One in five online scholarly articles affected by 'reference rot'," Los Alamos National Laboratory, January 26, 2015, accessed April 20, 2015, `http://www.lanl.gov/discover/news-release-archive/2015/January/01.26-scholarly-articles-affected-by-reference-rot.php`.

[58] Jennifer Lee and Eric Price, "Version Controlling the News: How We Can Archive" (South by Southwest, Austin, TX, March 18, 2013), accessed April 20, 2015, `http://www.slideshare.net/jenny8lee/newsdiffs`.

[59] See http://amberlink.org; see also Jonathan Zittrain, *The Fourth Quadrant*, SSRN Scholarly Paper ID 1618042 (Rochester, NY: Social Science Research Network, May 30, 2010).

that companies like Google, Facebook, and Twitter have towards the contemporary public sphere, global information flow, and implicitly, history.[60] But even public and nonprofit efforts often come under scrutiny, as the Library of Congress' technology efforts were recently criticized for "digital neglect," and the Internet Archive has been accused of preserving "a specific set of values, a perspective," as limited and opinionated in its worldview as any traditional archive.[61]

Most link preservation efforts primarily rely on repeated mirroring, or copying—as the Stanford University Libraries' LOCKSS program acronym says, "Lots of Copies Keep Stuff Safe."[62] Whether or not it keeps stuff safe, lots of copies might keep stuff from being organized and traceable. While I will treat the historical implications of copying in the next chapter, Marlene Manoff implicitly attributes archive fever to the sheer ease of copying; if a user were in an actual archive she wouldn't scan every page, but she is happy to save and even re-host an entire database with just a few clicks.[63] This creates an arbitrary number of digital replicas online and explodes the notion of a "canonical" version at a specific URL.

In this chapter, I have aimed to broaden the scope and meaning of the word "archive" to encompass digital databases and personal note-taking systems alike. I have considered the web-as-archive in order to highlight the ways in which the web exerts influence as a particular type of archive and network, with its own blend of data structures and algorithms. The paradox of the online archive lies in the ways it both contains and links documents. It contains them in a traditional sense, reducing media to addresses, representations, and signs. But it another sense the new order of archive connects media, by highlighting each artifact's role in a broader network or ecosystem, with new potential paradigms for organization and reuse.

[60] Andrew Whitacre, "Podcast: Tarleton Gillespie: Algorithms, and the Production of Calculated Publics," CMS/W, May 2, 2014, accessed April 20, 2015, `http://cmsw.mit.edu/podcast-tarleton-gillespie-algorithms-production-calculated-publics/`.

[61] The Editorial Board, "Digital Neglect at the Library of Congress," *The New York Times* (April 4, 2015), accessed April 20, 2015, `http://www.nytimes.com/2015/04/05/opinion/sunday/digital-neglect-at-the-library-of-congress.html`; Amelia Abreu, "The Collection and the Cloud," The New Inquiry, March 9, 2015, accessed April 20, 2015, `http://thenewinquiry.com/essays/the-collection-and-the-cloud/`.

[62] Zittrain, *The Fourth Quadrant*, 2778.

[63] Manoff, "Archive and Database as Metaphor," 386.

Chapter 3

An Intertwingled History of Linking

The act of linking the archive is certainly aided by digital tools, but it is not a requirement. Many indexing and note-taking systems of the Renaissance and Enlightenment allowed for the interlinking of disparate ideas, and these offer useful inspirations and foils for examining the web and its related research tools today. Information overload is not a new phenomenon, and pre-digital knowledge systems had many techniques for what Ann Blair calls the four Ss: storing, summarizing, sorting, and selecting.[1] Moreover, the web is only one of many digital hypertext systems, and the hyperlink—the primary object and mechanic for network formation on the web—has its own limitations that early hypertextual systems bring into full relief, inviting close analysis of the web's archival affordances.

In section 2.1 I confessed that my use of the word "archive" might expand and contract in scope, signifying a token of preservation and access rather than a singular fixed artifact. In each succeeding section, I aim to hone in my definition of the archive, ultimately to the digital news publishers that form the primary case study of my inquiry. Here my definition remains broad, but I will take a historical rather than theoretical approach to the archive, especially encompassing the pre-digital and pre-web indexes, note-taking systems, bibliographies and encyclopedias that first forayed into networked information.

[1] Ann Blair, "Note Taking as an Art of Transmission," *Critical Inquiry* 31, no. 1 (September 1, 2004): 85.

Most histories of the web's origins begin with Vannevar Bush (and sometimes Paul Otlet before him), leading directly through hypertext pioneers Ted Nelson and Douglas Engelbart, and concluding with Tim Berners-Lee's World Wide Web in a direct line from past to present. I will look closely at these individuals and their goals, and even use this chronological lineage as a structuring point, but I will also break apart this history by introducing other systems and figures—whether they existed long before computers or after the rise of the web—that point towards three corresponding themes. These themes recurrently surface when dealing with digital archives and information management: *spatialization*, *intersubjectivity*, and *encyclopedism.*

3.1 Spatialization: The Radiated Library

Here I will examine the tendency to use visual metaphors for information retrieval, and the associations between memory and physical space. The spatial and dimensional nature of knowledge is at odds with the "flattening" effect of indexes and the collapsing of dimensional space that non-hierarchical linking affords. Cycling through Ephraim Chambers' *Cyclopaedia*, I will examine Paul Otlet's vision of the "radiated library" and his architectural inspirations.

Memory has a strong spatial component; even when we don't remember something, we often know where to find it. A 2011 Columbia University study asked participants to save statements into various folders with generic names (such as FACTS, DATA, INFO, and POINTS). Despite the unmemorable folder names, "participants recalled the places where the statements were kept better than they recalled the statements themselves." The researchers found that " 'where' was prioritized in memory," providing preliminary evidence that people "are more likely to remember where to find it than to remember the details of the item."[2] They conclude by suggesting that we may be using Google and Wikipedia as memory extensions that then rewire our own internal memory.

[2] Betsy Sparrow, Jenny Liu, and Daniel M. Wegner, "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips," *Science* 333, no. 6043 (August 5, 2011): 776–778.

But humans have relied on external memory since the origin of writing itself, and in the meantime we have developed scores of analog systems and techniques—Barnet might call them "memory machines," John Willinsky "technologies of knowing"—to help summarize, filter, sort, and select. Computer systems are only one piece of this longer history of tools and practices. David Weinberger's "three orders of order" suggest this continuum, while also pointing out the rupture that the digital creates. The first order consists of things themselves, such as books in a library. The second order is a physical set of indexes, pointers, and references to the things, like a library card catalog. Finally, the third order is the digital reference, made of bits instead of atoms.[3] The third order allows items to be in multiple categories at once, as if in multiple sections of the library—this is a phenomenon that information architects call *polyhierarchy*.[4]

A theme across all of these orders of order is a reliance on spatial memory (the "where to find it" in the Columbia study). Archival and classification schemes use terms like "border," "domain," and "kingdom" (is it a coincidence that these terms all carry connotations of politics and power struggle?). We visualize network schemes as trees and as rhizomes, represented on maps, graphs, and diagrams. It seems that proper spatial visualization of an archive might not only help us remember where something is saved, but also give a high-level understanding of the archive itself, improving browsing and serendipitous search.

The ancient practice of constructing "memory palaces" (and Giulio Camillo's memory theater of the Renaissance)—outlined in Frances Yates' *The Art of Memory*—strongly emphasizes memory's reliance on spatial orientation and fixed dimension.[5] In order to construct a memory palace, the first step is to imagine a series of *loci*, or places, to determine the order of the facts. Only after creating space can one then create the images that represent the facts themselves. The structure that these palaces

[3] David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder* (New York, NY: Macmillan, April 2008), 17-23.

[4] The second order supports polyhierarchy, but in an awkward way that still privileges a single interpretation—think of a "see also" footnote at the bottom of an index entry. The third order allows equal footing for all possible categories.

[5] Frances Amelia Yates, *The Art of Memory* (London: Routledge, 1966).

take on are up to the memorizer, but once fixed, they are rarely reordered—only added to. This completes a grander spatial metaphor that Peter Burke notices—that of the *course*, which a student must run, envisioning and memorizing *images* in *places* along the fixed route towards knowledge.[6] Such an exercise emphasizes the temporal as well as spatial, as items are better remembered in sequence (such as with a rhyming poem).

This reliance on spatial and temporal memory keeps us in just two or three dimensions; it does not escape the trappings of the physical archive. If our memories rely on a fixed visual referent to know where a book is in a library, then we cannot rearrange the library's stacks and expect to find it again. A similar concern arises with online reading and writing. Ted Nelson calls hypertext "multi-dimensional," and Stuart Moulthrop says it aims to be "writing in a higher-dimensional space,"[7] but some readers still prefer paper-imitating PDFs to websites and e-books, because PDFs maintain a layer of real-world dimensional reference (as in, "I remember reading that sentence near the top of the page in the left column..."). For all of the liberating power of the digital, computers still rely on physical metaphors to be usable, and so we use digital equivalents of desktops, files, folders, and cards. The web even nods to this with its hierarchical URL structure that asks us to "navigate" down "paths" in given "domains."

This last fact is surprising given that a common theme among hypertext's pioneers, including Berners-Lee, is a desire to break down traditional linear and hierarchical classification schemes. A hierarchical scheme—like Linnaeus's biological taxonomy or Dewey's decimal classification—immediately suggests a tree view, and we can find many old examples of tree graphs in the Renaissance and Enlightenment. On the other hand, an alphabetical scheme offers a linear view, one that flattens the brittle hierarchy of taxonomy, but dulls its potential for association. The linked hypertext view might be seen as a multi-dimensional graph, more nuanced and flexible but more

[6] Peter Burke, *A Social History of Knowledge: From Gutenberg to Diderot* (Cambridge: Polity, December 2000), 90.

[7] Stuart Moulthrop, "To Mandelbrot in Heaven," in *Memory Machines: The Evolution of Hypertext*, by Belinda Barnet (London: Anthem Press, July 15, 2013).

difficult to grasp. If the first two orders are in one (linear) and two (hierarchical) dimensions, how can we bring the third order of order into a still higher dimension? And can it complement the ways that our minds visualize information?

3.1.1 The linked encyclopedia

Some older, pre-digital systems and practices have hybrid hierarchical/linear structures that start to suggest a network. While not the first system to incorporate links, Ephraim Chambers' *Cyclopaedia* is one of the first reference works of its kind. The encyclopedia reads somewhat like a dictionary, but it expands into general knowledge and opinion as well, and it always suggests multiple views into its contents. Chambers wrote that his encyclopedia went beyond a dictionary because it was "capable of the advantages of a continued discourse."[8] The word "encyclopedia" literally means "circle of learning," calling into question the shape of such a knowledge structure. It may be organized linearly, but as a collection of words to describe words, it always strives to double back on itself and highlight its own circular logic.

The *Cyclopaedia* was organized alphabetically, a relatively bold form of classification in relationship to the traditional, hierarchical schemes. Most scholars seem to agree that alphabetical order was born out of sheer necessity, related to the "intellectual entropy" and "epistemological urgency" of the time.[9] New knowledge was simply being created too fast to systematize and order. But Michael Zimmer suggests that alphabetical order signaled the beginning of a shift to more distributed, networked, and "egalitarian" forms of knowledge organization.[10] For instance, religious topics would be placed alongside secular ones. Alphabetical organization also turned the system into more of a "quick reference" guide that favored brief digests over long forays into

[8] Ephraim Chambers, *Cyclopædia, or an Universal Dictionary of Arts and Sciences* (1728), i, `http://uwdc.library.wisc.edu/collections/HistSciTech/Cyclopaedia`; Richard Yeo, "A Solution to the Multitude of Books: Ephraim Chambers's "Cyclopaedia" (1728) as "The Best Book in the Universe"," *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 64.

[9] Daniel Rosenberg, "Early Modern Information Overload," *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 5.

[10] Michael Zimmer, "Renvois of the Past, Present and Future: Hyperlinks and the Structuring of Knowledge from the Encyclopédie to Web 2.0," *New Media & Society* 11, no. 1 (February 1, 2009): 100.

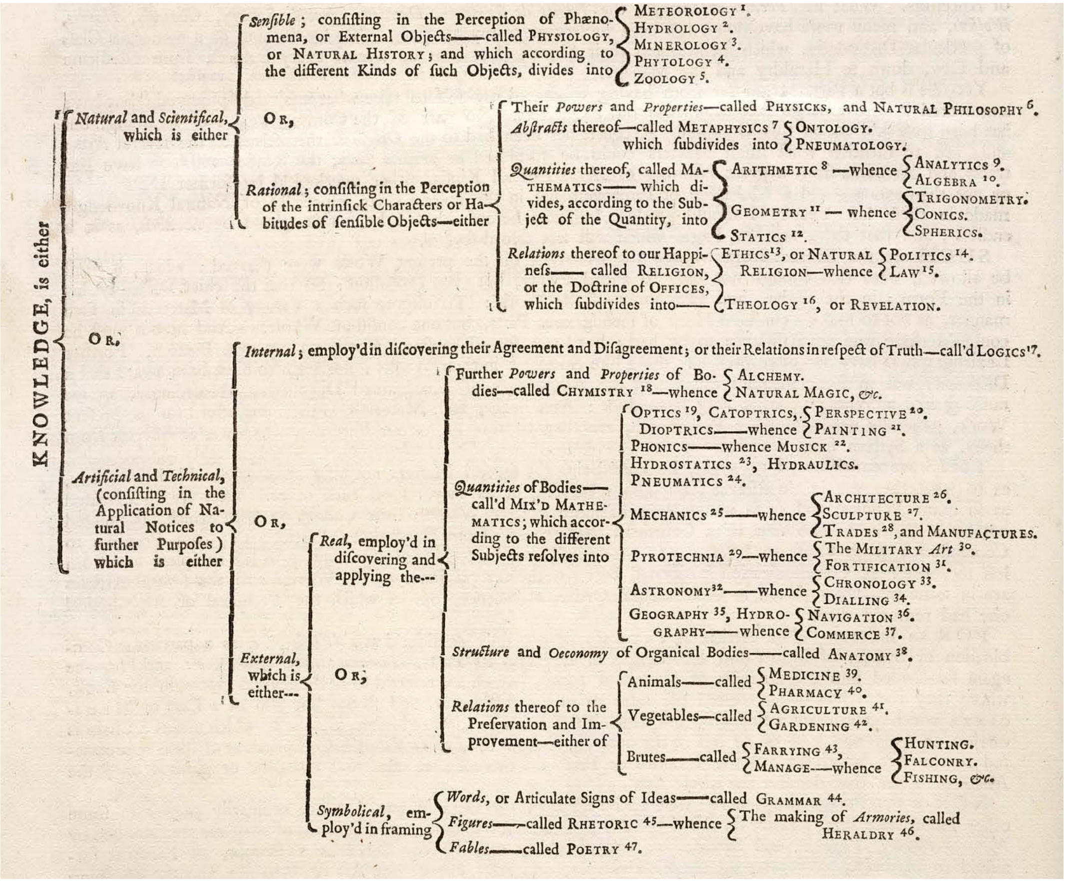

KNOWLEDGE, is either
Natural and Scientifical, which is either
Senſible; conſiſting in the Perception of Phænomena, or External Objects——called Physiology, or Natural History; and which according to the different Kinds of ſuch Objects, divides into
Meteorology 1.
Hydrology 2.
Minerology 3.
Phytology 4.
Zoology 5.
Or,
Rational; conſiſting in the Perception of the intrinſick Characters or Habitudes of ſenſible Objects—either
Their Powers and Properties—called Physicks, and Natural Philosophy 6.
Abſtracts thereof—called Metaphysics 7 which ſubdivides into
Ontology.
Pneumatology.
Quantities thereof, called Mathematics—— which divides, according to the Subject of the Quantity, into
Arithmetic 8 —whence
Analytics 9.
Algebra 10.
Geometry 11 — whence
Trigonometry.
Conics.
Spherics.
Statics 12.
Relations thereof to our Happineſs—— called Religion, or the Doctrine of Offices, which ſubdivides into——
Ethics13, or Natural Religion—whence
Politics 14.
Law 15.
Theology 16, or Revelation.
Or,
Artificial and Technical, (conſiſting in the Application of Natural Notices to further Purpoſes) which is either
Internal; employ'd in diſcovering their Agreement and Diſagreement; or their Relations in reſpect of Truth—call'd Logics17.
Or,
External, which is either---
Real, employ'd in diſcovering and applying the---
Further Powers and Properties of Bodies—called Chymistry 18—whence
Alchemy.
Natural Magic, &c.
Quantities of Bodies—— call'd Mix'd Mathematics; which according to the different Subjects reſolves into
Optics 19, Catoptrics, Dioptrics——whence
Perspective 20.
Painting 21.
Phonics——whence Musick 22.
Hydrostatics 23, Hydraulics.
Pneumatics 24.
Mechanics 25——whence
Architecture 26.
Sculpture 27.
Trades 28, and Manufactures.
Pyrotechnia 29—whence
The Military Art 30.
Fortification 31.
Astronomy32——whence
Chronology 33.
Dialling 34.
Geography 35, Hydrography——whence
Navigation 36.
Commerce 37.
Structure and Oeconomy of Organical Bodies——called Anatomy 38.
Relations thereof to the Preſervation and Improvement—either of
Animals——called
Medicine 39.
Pharmacy 40.
Vegetables—called
Agriculture 41.
Gardening 42.
Brutes——called
Farrying 43,
Manage——whence
Hunting.
Falconry.
Fishing, &c.
Or;
Symbolical, employ'd in framing
Words, or Articulate Signs of Ideas——called Grammar 44.
Figures——called Rhetoric 45—whence
The making of Armories, called Heraldry 46.
Fables——called Poetry 47.

Figure 3-1: The "branches of knowledge" in Ephraim Chambers' *Cyclopaedia.*

knowledge; the practices of browsing, skimming and summarizing were continuously honed during the Renaissance and Enlightenment as scholars coped with "a confusing and harmful abundance of books" as early as 1545.[11] Chambers called this complaint "as old as Solomon."[12]

All the same, Chambers felt he needed an overarching scheme. In the encyclopedia's preface, he included a diagram and listing of forty-seven categories (called Heads), complete with cross-references to the entries. In Chambers' words, "the difficulty lay in the form and oeconomy of it; so to dispose such a multitude of materials, as not to make a confused heap of incoherent Parts, but one consistent Whole."[13] In order to truly demonstrate a "continued discourse," Chambers needed a graph, a

[11] Ann Blair, "Reading Strategies for Coping With Information Overload ca.1550-1700," *Journal of the History of Ideas* 64, no. 1 (2003): 11–28.

[12] Yeo, "A Solution to the Multitude of Books," 65.

[13] Ibid., 67.

map. Each of the Heads in the diagram contains a footnote that lists that head's terms (known as Common Places).

Chambers' use of Heads and Common Places followed Phillipp Melanchthon's 1521 subject division into *loci* and *capita* (Peter Burke suggests that these would now be called "topics" and "headings," less strong and physical metaphors).[14] *Loci* ("places") bring to mind memory palaces, but also the "commonplace book"—to which Chambers was knowingly attaching himself. Many scholars used commonplace books as information management devices to store quotes, summaries, aphorisms, and so on, and these often had specialized systems for retrieval. Richard Yeo sees Chambers' use of the term as directly appealing to the popularity of commonplace books at the time.[15] Ann Blair also argues that note-taking and commonplacing were far more common than the memory palaces and theaters outlined by Frances Yates, and that the two traditions made "no explicit reference to one another."[16] Still they share a strong common thread: a reliance on *loci* as the root of knowledge retention, memory, and interconnection.

The *Cyclopaedia* was an ancestor to Diderot's celebrated *Encyclopédie* (Diderot started by translating Chambers). Diderot's work made further use of *renvois* (references) to question and subvert traditional knowledge structures and authorities—including the book's own authority as a reference work. Michael Zimmer argues that Diderot also used *renvois* to hide politically controversial topics in seemingly dry and tangential entries, "guiding the reader to radical or subversive knowledge" while evading the eyes of the censors.[17] Zimmer directly ties the *renvois* to the hypertext link, suggesting that Bush, Nelson, and Berners-Lee all "intended to free users from the hegemony of fixed information organization in much the same way that *renvois* did for the readers of the *Encyclopédie*."[18]

It is clear that Diderot fully recognized and built upon Chambers' developments in

[14] Burke, *A Social History of Knowledge*, 95.

[15] Yeo, "A Solution to the Multitude of Books," 65-66.

[16] Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (Yale University Press, November 2, 2010), "Note Taking as an Aid to Memory".

[17] Zimmer, "Renvois of the past, present and future," 103.

[18] Ibid., 104.

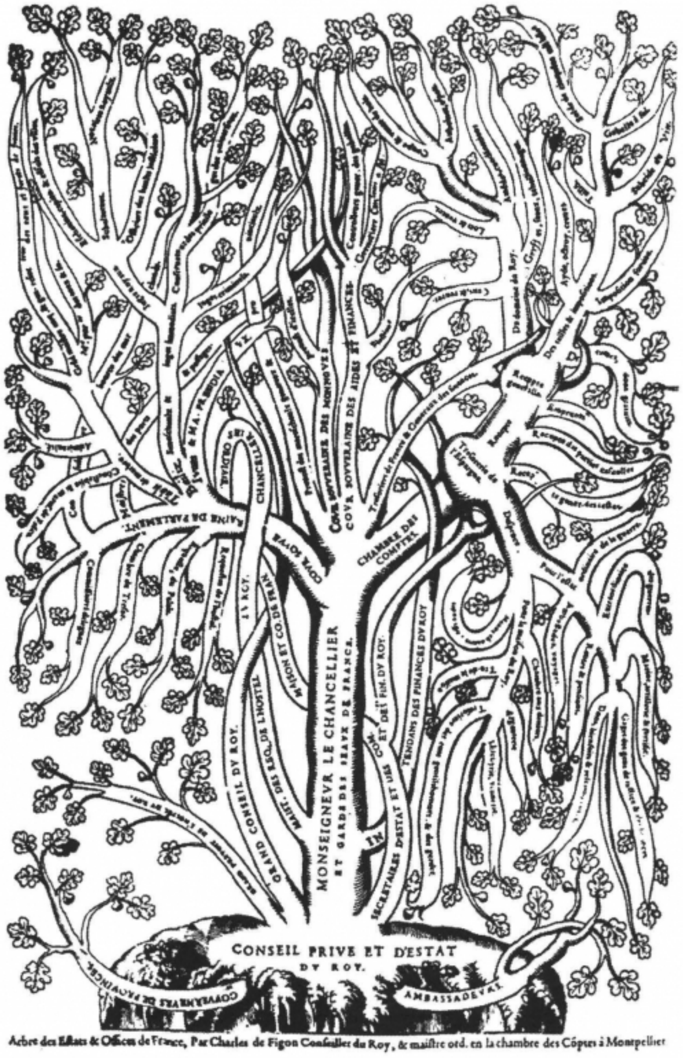

MONSEIGNEVR LE CHANCELLIER
ET GARDE DES SEAVX DE FRANCE.
CONSEIL PRIVE ET D'ESTAT
DV ROY.
GOVVERNEVRS DE PROVINCES.
AMBASSADEVRS
GRAND CONSEIL DV ROY.
CHAMBRE DES COMPTES.
COVR SOVVERAINE DES AIDES ET FINANCES.
SECRETAIRES D'ESTAT ET DES FIN. DV ROY
Acbre des Estats & Offices de France, Par Charles de Figon Conseiller du Roy, & maistre ord. en la chambre des Côptes à Montpellier

Figure 3-2: *Arbre des Etats et Offices de France*, by Charles de Figon (Paris 1579).

linking references, but I call into question the notion that the prior “fixed” organization systems had no detractors or provisional solutions (moreover, the *renvois* are “fixed” themselves). Carolus Linnaeus, the author of perhaps *the* prototypical taxonomy, knew well that classifications are “cultural constructs reflecting human ignorance.”[19] Leibniz also understood its limitations; his *Plan for Arranging a Library* included a “miscellaneous” section, a tacit acknowledgement that the system is in some way imperfect or incomplete.[20] Leibniz also praised his famous Note Closet, developed by Thomas Harrison, for this same ability: “A single truth can usually be put in different

[19] Zimmer, “Renvois of the past, present and future,” 99.

[20] Burke, *A Social History of Knowledge*, 106.

places, according to the various terms it contains. . . and different matters to which it is relevant."[21]

Moreover, multiple hierarchies can coexist and offer competing schemes. Some of these schemes were already organized not as much around content as con*text*. Peter Burke points out that Islamic classification systems were also tree-structured, but every element was organized based on its degree of separation from the Quran.[22] This is, crucially, an early citation-based network.

3.1.2 Paul Otlet and the dimensions of memory

Along with Vannevar Bush, Paul Otlet bridges the second and third orders of order. Born in Belgium in 1868, Otlet predated Ted Nelson's role as an obsessive encyclopedist and commonplacer. Between the ages of ages 11 and 27, he amassed 1400 pages of notes, and in his first move to Paris, he called it "the city where the world comes to take notes."[23] He liked to think big and in the aggregate, creating the Universal Decimal Classification and Universal Bibliographic Repertory. He also supported international politics associations like the League of Nations and the forerunner to UNESCO, going so far as to found the Union of International Assocations (which is, indeed, an international association of international associations) with his friend Henri La Fontaine in 1907.

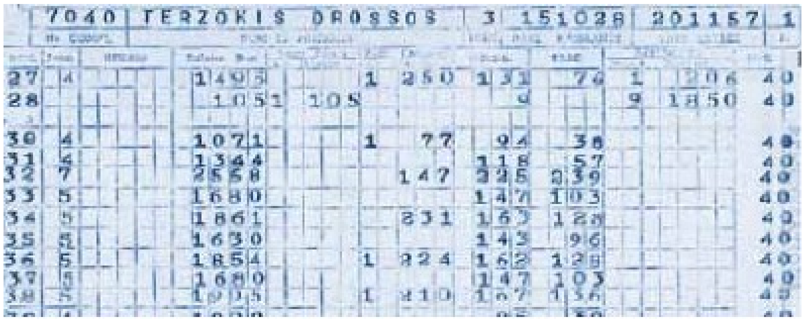
Due in part to the destruction of much of his work in World War II, Otlet was mostly forgotten for decades in favor of his American successors. However, the rise of the web and the efforts of several scholars—particularly his biographer Boyd Rayward—have given him a new life as a prescient predictor of a networked hypertext system. As one of the originators of information science, he envisioned (and even began to amass) a universal library to serve as the heart and central authority of the world's information. Following his belief that books were redundant and arbitrary ag-

[21] Blair, *Too Much to Know*, "Managing Abundant Notes".

[22] Burke, *A Social History of Knowledge*, 94.

[23] "A Limited Company of Useful Knowledge: Paul Otlet, the International Institute of Bibliography and the Limits of Documentalism," everything2, May 18, 2001, accessed September 23, 2014, `http://everything2.com/index.pl?node_id=1053046`.

glomerations that obscure the data held within (which is the object of a researcher's true inquiry), he suggested a universal decimal classification system that built on Dewey's system to incorporate an item's metadata, its references and constituent parts. Its entries read less like library call numbers and more like modern databases' structured queries. And in his most striking prediction, he proposed a "radiated library" that could handle remote requests from a centralized location by screen and telephone. He envisioned the screen with multiple windows for simultaneous document consultation, audiovisual data, and finally a full automation of the document request process: "Cinema, phonographs, radio, television, these instruments taken as substitutes for the book, will in fact become the new book."[24] While the phrase belies its roots in radiation and radio, Otlet's "radiated library" and "televised book" combine to suggest the networked multimedia of the web, more than 50 years before its creation.



7040 | TERZOKIS DROSSOS | 3 | 151028 | 201157 | 1

27	4	1495	[illegible]		1	250	131	76	1	206	40
28		1051	1	108			9		9	1850	40
30	4	1071	1		1	77	94	38			40
31	6	1344	4				118	57			40
32	7	2558	8			147	225	239			40
33	5	1680	0				147	103			40
34	5	1861	1			231	163	128			40
35	5	1630	0				143	96			40
36	5	1854	4		1	224	162	128			40
37	5	1680	0				147	103			40
38	5	[illegible]	5		1	210	167	156			40

Figure 3-3: Excerpt of the Universal Library's indexing system.

Otlet was an encyclopedist, but also an innovator in graphical and spatial representation. He frequently used architecture as a foil, metaphor, and inspiration for bibliographic structures, calling his main work *Traité de documentation* a study of the "architecture of ideas."[25] The first names for the Mundaneum—the universal repository Otlet and La Fontaine set out to build—were alternately "city of knowledge" and "World Palace." In the end, the Mundaneum—like the archive itself—bridged

[24] Ijsbrand van Veelen, *Alle Kennis van de Wereld (Biography of Paul Otlet)* (Noorederlicht, 1998), accessed December 19, 2014, `http://archive.org/details/paulotlet`.

[25] Charles van de Heuvel, "Building Society, Constructing Knowledge, Weaving the Web: Otlet's Visualizations of a Global Information Society and His Concept of a Universal Civilization," in *European Modernism and the Information Society* (Ashgate Publishing, Ltd., 2008), 129.

the physical and the digital or metaphysical, as Otlet called it at once "an idea, an institution, a method, a material body of work, a building and a network."[26] In his discussion of the architecting of knowledge, Otlet also crucially recognized that ideas are never so fixed as physical structures; as Charles van de Heuvel puts it, "For Otlet it was important to leave space for transformation and modification in response to the unforeseen and unpredictable."[27] Leibniz had conceived of the "library without walls" long before, but Otlet's radiated library went many steps further. As one of the fathers of information science, he is also one of its first information architects.

Otlet's resulting decimal classification and networked library is less bound by linear or hierarchical schemes. The architectural inspiration also may have helped him conceive of the radiated library, one that could transmit signals across space between screens, several decades before the first computers were linked together. All the same, it is hard to see Otlet's universal library project as anything but quixotic. The perpetual collection and detailed organization of the entirety of human history in one location, all managed by 3x5 index cards, is doomed to fail. Still, Otlet's system seems to have worked usefully for a time: the library had more than 17 million entries by 1934, handling 1500 research requests per year, all on the backbone of Otlet's Universal Decimal Classification.[28] The universal repository was, of course, never completed, but it came closer to fruition than the memex or Xanadu.

3.2 Intersubjectivity: The Memex

An individual's personal archive has markedly different properties and requirements than a group's or institution's, which in turn is different from a massive, aggregated universal archive for the public. At the same time, some archives sit in between these scopes, and each has different purposes and practices surrounding it. Linking and categorization schemes rely on individuals making connections between information, but different individuals might not make the same connections; how does linking be-

[26] Ibid., 130.
[27] Ibid., 131.
[28] "A Limited Company of Useful Knowledge."

come a collective and collaborative endeavor, a universal language? This phenomenon is both explicated and emphasized by a contemporary example: the web's algorithmic recommendation systems that conflate the individual and the collective as they traverse the links of the web.

The scrapbooks, commonplace books, and card catalogs of old usually belonged to an individual. He or she might share them and collaborate with others, or collect resources for children and grandchildren, but these early systems generally reflected and mimicked the scattered mind of a single person. A scholar's notes are likely to consist of many shorthands, mental leaps, and personal anecdotes that no one else would follow. Interestingly, most early hypertext systems focused on this individual scope, or at most on collaborative or collective research. Only Xanadu (and perhaps Otlet's Mundaneum) had the world-encompassing scope of the web.

Jeremias Drexel stated in 1638 that there is no substitute for personal note-taking: "One's own notes are the best notes. One page of excerpts written by your own labor will be of greater use to you than ten, even twenty or one hundred pages made by the diligence of another."[29] People forge connections and organizational schemes in unique and sometimes conflicting ways. As more and more people enter a system, it will encounter more and more possible definitions and connections.

The idiosyncratic connections formed by an individual's memory make it difficult to generalize categories. An individual's thought process might be reminiscent of Borges' Chinese encyclopedia, which offers a taxonomy of animals divided by absurd traits, such as "Those that belong to the emperor, embalmed ones, those that are trained, suckling pigs, mermaids, fabulous ones, stray dogs," and "those that are included in this classification."[30] These may be the trails that a mind follows, but the humor lies in calling it a taxonomy, in making the categories *intersubjective* and even official, *objective.* Borges' categories remind us of Bowker and Star's argument that classifications will always be compromises, between individuals and groups, or between groups and a collective whole.

[29] Blair, "Note Taking as an Art of Transmission."

[30] Jorge Luis Borges, "The Analytical Language of John Wilkins," in *Other Inquisitions 1937-1952*, trans. Ruth L.C. Simms (Austin, TX: University of Texas Press, 1993).

Markus Krajewski's *Paper Machines: About Cards and Catalogs* hinges on the difference and tension between a personal note-taking system and a universal library. We often use the same systems for organizing each (such as the card catalog or the SQL database), but they don't turn out to be for the same uses. Krajewski says "The difference between the collective search engine and the learned box of paper slips lies in its contingency."[31] Whenever we add a tag or make a connection in an archive, we are attempting to predict what will be searched for later. But it is easier to classify in a personal archive; we can predict our future selves better than we can predict the future.

As a result, personal note-taking tools might seem like an easier place to start with the challenge of hypertext. They are certainly technically easier, avoiding collaboration issues like version control. But an archive is almost never entirely personal. Thought may be idiosyncratic, but it follows common patterns. Users want the possibility of sharing documents, or of passing on entire collections to others. Ann Blair points out that successors would fight over their ancestors' notes in Renaissance wills, which suggests that any time a commonplace book is begun, it has some kind of common value.[32] In the case of historical figures, personal notes often become a literal part of an archive, then meant for public consultation. But we treat these archives differently than those that are constructed *for* us. For instance, Walter Benjamin's *Arcades Project* is a set of notecards, published as a sort of commonplace book that has become a prominent work to consult in its own right. Though Benjamin's cards were for his own use, he was aware of their potential value as a shared resource, entrusting them to his librarian friend Georges Bataille before fleeing Nazi-occupied France. So is the *Arcades Project* a book, an archive, or a database? Who is it for? What happens to Benjamin's memory as it becomes shared history?

This relationship between the personal and the collective is taking on new meaning on the web, where we expect personalized information, but rely on a massive collective of people in order to get it. Nick Seaver argues that recommendation sys-

[31] Markus Krajewski, *Paper Machines: About Cards and Catalogs* (Cambridge: MIT Press, 2011), 50.

[32] Blair, "Note Taking as an Art of Transmission," 104.

tems “algorithmically rearticulate the relationship between individual and aggregate traits.”[33] The communities and demographics that form around individuals can in turn be aggregated and intersected into a single, massive whole. At each stage, memory is abstracted further and further from us.

Today’s efforts to organize the web and its sub-archives (i.e. the web applications, tools, and platforms we use every day) tend to reflect this and aim to marry the best of both worlds: the individual and the mass. Clay Shirky and David Weinberger champion the folksonomy as a solution; let individuals tag however they want, and at the right scale everything will sort itself out.[34] The Semantic Web is similarly structured, by letting users define their own vocabularies for both pages and links, but strictly enforcing them once made. These approaches are certainly worth pursuing, but both still rely on fixed language rather than associative connection; tagging an item is undoubtedly an act meant to make connections between documents, but it is always mediated by language and structured according to certain systematic and linguistic conventions.

3.2.1 Vannevar Bush’s memory machine

Unlike Otlet’s radiated library, or Nelson’s Xanadu, Vannevar Bush’s memex was decidedly a machine designed for personal use. It did not build in weblike networked affordances. All the same, Bush suggests many intersubjective uses for the memex, adding to the confusion between personal archive and collective library.

Bush was perhaps best known as the director of U.S. military research and development during World War II, but he also made a lasting contribution to hypertext; a 1945 essay called “As We May Think” conceived of the memex machine, an automated microfilm device that could store an entire library in one drawer and retrieve any item within seconds.[35] Perhaps most crucially, Bush conceived of new ways to

[33] Nick Seaver, “Algorithmic Recommendations and Synaptic Functions,” *Limn*, no. 2 (2012): 44–47, `http://limn.it/algorithmic-recommendations-and-synaptic-functions/`.

[34] Clay Shirky, “Ontology is Overrated: Categories, Links, and Tags,” 2005, `http://www.shirky.com/writings/ontology_overrated.html`; Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder*, 165-8.

[35] Vannevar Bush, “As We May Think,” *The Atlantic* (July 1945), `http://www.theatlantic.`

connect items: through associative trails. Linda C. Smith analyzed the citation network of many hypertext articles and discovered, in Belinda Barnet's words, that "there is a conviction, without dissent, that modern hypertext is traceable to this article."[36]

Bush begins by arguing that, "The summation of human experience is being expanded at a prodigious rate," but suggests that our methods for retrieving such experience are hindered by "the artificiality of systems of indexing."[37] He points out the limitations of keeping data only in one place, and of using strict formal rules to access it: "the human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain." His proposed solution, the memex, aims to mechanize "selection by association, rather than by indexing."[38]

The memex is built for personal use; Bush's model is "the human mind," after all, and not "human minds" (as Barnet notes, he follows the cybernetic tradition of the time in modeling computation on human thought, along with Wiener, Shannon, Licklider, and others).[39] The idiosyncrasies of individual trails, and the challenges in developing a new language for a new invention, would suggest that the machine was strictly for individual use. However, Bush points immediately to its possibility for generalization as well; he envisions an example of a person sending his trail of research to a colleague "for insertion in his own memex, there to be linked into the more general trail."[40]

Bush goes on to suggest that the memex will hold new forms of encyclopedias and ready-made trails, along with "a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record."[41] This now seems prophetic, a prediction of contemporary culture's emphasis

com/magazine/archive/1945/07/as-we-may-think/303881/.

[36] Belinda Barnet, "The Technical Evolution of Vannevar Bush's Memex," *Digital Humanities Quarterly* 2, no. 1 (2008).

[37] Bush, "As We May Think."

[38] Ibid.

[39] Barnet, "The Technical Evolution of Vannevar Bush's Memex."

[40] Bush, "As We May Think."

[41] Ibid.

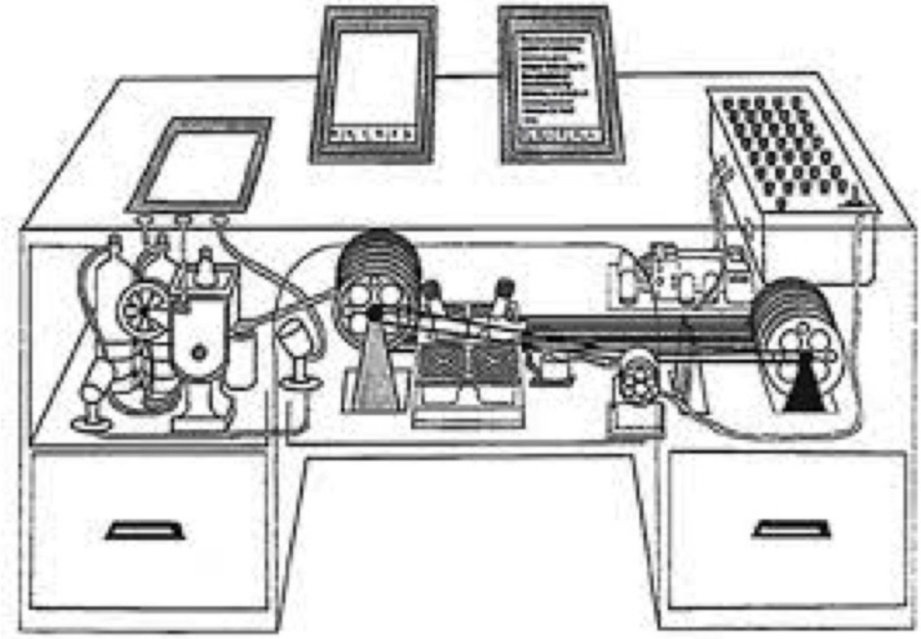


Figure 3-4: Vannevar Bush's memex diagram.

on curation, though it was predated by Otlet's assertion that we will need a new class of "readers, abstractors, systematisers, abbreviators, summarizers and ultimately synthesizers."[42] Bush does not dwell on this to consider where this "common record" will live, who will own and control it, and how individuals will tie these resources to their own idiosyncratic trails. The shift from subjectivity to intersubjectivity, and then in turn from intersubjectivity to some form of *objectivity*, makes each act of classification—or in Bush's case, each act of association—increasingly fraught.

Bush's work relies on the trail, a closely curated path where one document directly associates with another. Ted Nelson instead suggested "zippered lists," which would operate like trails but without Bush's emphasis on sequence.[43] In each of these cases they rely on a human curator to create the links. Bush envisions trails shared for personal, collaborative, and general use, but the connection itself remains person-to-person, intersubjective on the smallest scale. The trails and associations formed by

[42] Joseph Reagle, *Good Faith Collaboration: The Culture of Wikipedia* (Cambridge, Mass.: MIT Press, 2010), 23.

[43] Theodor H. Nelson, "Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate," in *Proceedings of the 1965 20th National Conference* (New York, NY, USA: ACM, 1965), 84, 89.

the memex always remain deeply human, and deeply individual.

In Bush's "Memex Revisited," he begins to tease out the possibility of the memex forming trails for a scholar, suggesting that it could "learn from its own experience and to refine its own trails."[44] Here the influence of Wiener's cybernetics and feedback theory are clear, and it begins to point to the machine learning and automated classification that occurs today. Most intriguing is Bush's suggestion that like the human mind, some well-worn trails would be kept in memory, reinforced and expanded, while other less-used trails would fall away. This conjures up the notion of a fluid archive, one that is constantly forming and re-forming its associations, dynamically linking the past.

But Bush's memex is not without its limitations. John H. Weakland offered two criticisms of the memex in response to "As We May Think." He asks "how personal associations of the general record could be generally useful," as well as how a researcher can find things they don't know about already.[45] It appears to me that the second challenge is an extension of the first: associative indexing may be more inherently fuzzy and idiosyncratic than content-based indexing systems like text search and tagging. It sacrifices fixity and consistency at the expense of individuality and nuance. So association may serve better as a supplement, rather than a replacement, for traditional classification schemes.

Another limitation of the memex, offered by Belinda Barnet, is that "Bush's model of mental association was itself technological; the mind 'snapped' between allied items, an unconscious movement directed by the trails themselves."[46] Bush himself recognized this, pointing out that the human memory system is a "three-dimensional array of cells" that can gather, re-form, and select relationships as a whole or a subset of a whole.[47] While later hypertext systems and the Semantic Web come closer to such a

[44] Vannevar Bush, "Memex Revisited," in *From Memex to Hypertext: Vannevar Bush and the Mind's Machine*, ed. James M. Nyce and Paul Kahn (San Diego, CA, USA: Academic Press Professional, Inc., 1991), 211.

[45] James M. Nyce and Paul Katin, "Innovation, Pragmaticism, and Technological Continuity: Vannevar Bush's Memex," *Journal of the American Society for Information Science* 40, no. 3 (May 1989): 217.

[46] Barnet, "The Technical Evolution of Vannevar Bush's Memex."

[47] Bush, "Memex Revisited," 209.

three-dimensional structure, like the memex they are often constrained to ‘snapping’ between associations.

Finally, even though Bush seems fully aware of the morphing state of collective knowledge and history, he assumed that the trails would not grow old. He envisions a father bequeathing a memex to his son, along with the myriad trails formed, as a fixed and locked document. Even Bush’s proposed adaptive memex would be modeled against the individual researcher; in machine learning terms, its “training set” would not be formed in the aggregate like modern-day recommendation systems, but rather from the unique trails formed by an individual.

3.3 Encyclopedism: Project Xanadu

This section analyzes the scale of knowledge systems, and specifically the constant striving to expand beyond the archive’s horizon. While the last section was based on the type and scale of *users* of the archive, this section concerns the type and scale of *information* or *content* within the archive. There does tend to be a relationship—an archive built for everyone is more likely to collect everything—but I divide them here to highlight the tendency for content to stretch towards complete and total comprehensiveness, or what I am calling *encyclopedism*. Many efforts to document, index, or link the world have truly attempted to map *the world*—every piece of information about everything—or have at least appealed to an impulse to do so. What leads to this encyclopedic impulse, and what are some of its promises and pitfalls? When building an archive, where do you stop?

Paul Otlet wanted to index every fact in every book. In his notes, he insists, “I write down everything that goes through my mind, but none of it has a sequel. At the moment there is only one thing I must do! That is, to gather together my material of all kinds, and connect it with everything else I have done up till now.”[48] This persistent, obsessive quest for comprehensiveness is part and parcel of the archive—you either want to collect and connect *everything*, or everything *worthwhile*, within a

[48] Reagle, *Good Faith Collaboration*, 20.

given scope.

Once again this conjures up a Borges story: his Library of Babel contains books with every permutation and combination of every letter. *Somewhere* in the library sits every great work ever written, and every great work that will be written. But the vast majority of these books are useless nonsense, and no great works will be found. Borges, a librarian himself, understood well the encyclopedic impulse and the noise and madness that results.[49]

Encyclopedism has its roots at least in the Renaissance, as Ann Blair notes: "it is reasonable to speak of encyclopedic ambition as a central ingredient of the Renaissance obsession with accumulating information."[50] Even in 1548, Conrad Gesner began compiling a "general bibliography" with the aim of indexing all known books; he ended with 10,000 works by 3,000 authors, which was surely an obsolete number even by the time he finished.[51] Some critics, like Jesuit scholars Francesco Sacchini and Antonio Possevino, recommended an "aggressively purged" rather than universal library, throwing out any redundant or misleading texts. Gesner disagreed, but his reasoning was telling: "No author was spurned by me, not so much because I considered them all worthy of being cataloged or remembered, but rather to satisfy the plan which I had set for myself."[52] He wanted to list all the books in order to leave others to be the judge, but first and foremost, he did it because it was his plan all along.

Some of today's technological language reflects this drive. Wikipedia's mission is "to give freely the sum of the world's knowledge to every single person on the planet,"[53] which is reminiscent of Google's: "to organize the world's information and make it universally accessible and useful."[54] The *world's* knowledge, *universally* accessible, to *every* person: the goal is impossible. Capturing "the sum of the world's knowledge" is akin to Borges' aleph—a point that contains all points—or his one-to-one map of

[49] Jorge Luis Borges, *Collected Fictions*, trans. Andrew Hurley (New York: Penguin, 1999), 112-18.

[50] Blair, *Too Much to Know*, "Information Management in Comparative Perspective".

[51] Burke, *A Social History of Knowledge*, 93.

[52] Blair, *Too Much to Know*, "Bibliographies".

[53] Reagle, *Good Faith Collaboration*, 18.

[54] "Company," Google, accessed April 20, 2015, `http://www.google.com/about/company/`.

the world. Still, Wikipedia knows well that "Regretfully, the vast majority of human knowledge is not, in actual fact, of interest to anyone, and the benefit of recording this collective total is dubious at best."[55]

All of these universal projects are destined to fail at their end goal, but the resulting collections can be useful. The book repositories and knowledge systems of today—Wikipedia, Google Books, Project Gutenberg, and Amazon—may have come closer than any previous efforts to capturing the world's knowledge, but they do so according to certain principles, conventions, demands and traditions. They also have something else in common: they must always adhere to the technical and conventional standards and limitations of the web itself.

3.3.1 Ted Nelson's endless archive

Ted Nelson, inventor of the term "hypertext," is a notorious collector, commonplacer, and self-documenter. He also always thinks big; he wants to collect *everything* and connect *everything* to *everything* ("everything is intertwingled," in his parlance), and only then will it all make sense. His project for doing so, called Xanadu, began work in 1960 and has inspired scores of hypertext acolytes, but after so many years of continuous development, it still has not been fully realized.

Nelson was deeply inspired by Bush's memex, referencing him frequently in presentations and even including the entirety of "As We May Think" in his book *Literary Machines.* Building on Bush's ideas, Nelson suggested "zippered lists" instead of trails, which could be linked or unliked as its creator desired, advancing beyond Bush's "prearranged sequences."[56] But his biggest development was to reintroduce the global ambition of Otlet into Bush's associative vision: the idea of a universal, networked, collectively managed hypertext system.

In Nelson's system, there would be no 404s, no missing links, no changes to pages forever lost to history. Links would be two-way, forged in both directions—imagine visiting a page and being able to immediately consult every page that linked *to* the

[55] Reagle, *Good Faith Collaboration*, 17.

[56] Theodor H. Nelson, *Computer Lib / Dream Machines* (Self-published, 1974), 313.

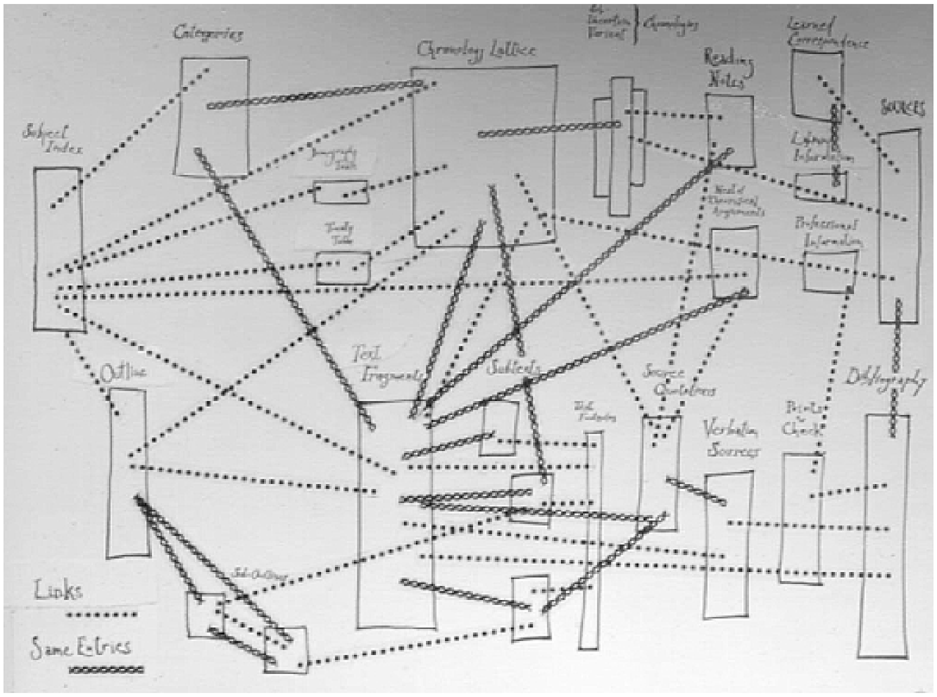


Figure 3-5: Schematic of a sample linked system in Xanadu.

page. And rather than copying, Xanadu operates on *transclusion*, a sort of soft link or window between documents that would allow new items to be quickly and easily constructed from constituent parts, readily pointing back to their source.

Nelson's idea for Xanadu might resemble Wikipedia; one of Wikipedia's core tenets is "No Original Research: don't create anything from scratch, just compile," reflecting the principle of Nelson's transclusions.[57] But on the web, where so much information is ripe for mash-up, remix, and reuse, the only option is to create from scratch. The links at the footer or the inside of a Wikipedia page are merely pointers and not true windows into the source documents. Nelson's transclusions are more akin to the Windows shortcut, Mac alias, or Linux softlink. The Web's default, on the other hand, is to *copy* rather than *link*. Jaron Lanier suggests that copying-not-linking is a vestige of the personal computer's origins at Xerox PARC, whose employer was quite literally in the business of copying, and was inherently wary of ideas that bypassed it.[58] While copying is certainly a useful safeguard against lost knowledge, it does so

[57] Reagle, *Good Faith Collaboration*, 11-12.

[58] Jaron Lanier, *Who Owns the Future?* (New York: Simon & Schuster, 2013), 221-232.

at the expense of added clutter and lost provenance.

One could look at the resulting Wikipedia, or any such aggregation of compiled knowledge, as a combination of two actions: *summarizing* and *filtering*. To summarize is to provide a shorter version of a longer text. To filter is to offer a verbatim excerpt of the text. Most knowledge systems that I am addressing here exist along a continuum between these two primary actions, and effective ones are able to elegantly balance both. Xanadu places more focus on filtering texts, while the web might lend itself better to summarizing; it is only through the web's hyperlinks that we get a glimpse of a filtering axis.

But unlike the web, Xanadu has still not been fully realized. It has lost, while the web has experienced an unprecedented, meteoric rise. Xanadu also has its share of detractors and challengers. Most of its biographies and summaries are fairly critical, most famously a 1995 *Wired* article that prompted a forceful response from Nelson.[59] There is a level of hubris in the encyclopedic impulse that Nelson doesn't hide. His proposed system is top-down and brittle in certain ways, including rigid security and identification systems. And his proposal for online "micropayments" per transclusion is interesting but controversial; Jaron Lanier and others have supported it, but many are skeptical, suggesting that it would stifle the sharing of knowledge and circulation of material.[60]

The Xanadu system is far from perfect, but its allure comes from the idea that it treats its contents with history and context in mind. Xanadu promised to treat its contents like an archive rather than making us build archives around it. Comparing it to the web raises interesting questions: how much structure, organization, and control should we place on our networked information systems? How much is desirable, and how much is technically and economically feasible? And if we consider the archival capabilities of each, how are they building, sorting, and selecting our information?

A skeletal version of Xanadu (still *without* its two-way links) was finally released

[59] Gary Wolf, "The Curse of Xanadu," *Wired* 3, no. 6 (June 1995).

[60] Jeff Atwood, "The Xanadu Dream," Coding Horror, October 12, 2009, accessed April 20, 2015, `http://blog.codinghorror.com/the-xanadu-dream/`; Lanier, *Who Owns the Future?*, Chapter 18.

on the web, after more than 50 years of development, in summer 2014.[61] It has joined the myriad archives and knowledge systems embedded inside the web. Many of the later, "second-generation" hypertext systems were geared towards personal and institutional uses (systems like NoteCards, Guide, WE, or Apple's HyperCard).[62] These likewise resemble the web platforms and tools we use today (such as Trello, Evernote, or Zotero). But these systems, like Xanadu itself, have been subsumed by the web. Hypertext systems can all interact with one another, but the encyclopedic, universal ones can only be in competition.

3.4 Conclusion

This long history of linked, indexed, and sorted archives would suggest that the current state of archives in the digital era has occurred as a result of a continuum of developments, rather than a radical leap into completely unknown territory. But in another sense, the digital does allow for a complete rupture. The "information overload" we experience today is a product of two factors, one old and one new. The *accumulation* of the archive is an age-old challenge that many tools, systems and practices have endeavored to solve. But the *linking* of the archive is a newer challenge. There has always been too much information, but now it can all be connected, quantified, broken down and aggregated as never before. As we sort through the webbed intersections of content and context, it will be crucial to keep in mind its long history; after all, it is what archives are fighting to preserve.

Archives' constant battle with issues of scope and dimensionality suggest a need to recognize and limit ambitions, to start small and build up rather than starting from the whole and breaking down. The linking of the archive requires knowing your

[61] Tim Carmody, "Pioneering hypertext project Xanadu released after 54 years," kottke.org, June 5, 2014, accessed April 20, 2015, `http://kottke.org/14/06/pioneering-hypertext-project-xanadu-released-after-54-years`; Alex Hern, "World's most delayed software released after 54 years of development," *The Guardian* (June 6, 2014), accessed April 19, 2015, `http://www.theguardian.com/technology/2014/jun/06/vapourware-software-54-years-xanadu-ted-nelson-chapman`.

[62] Frank Halasz, "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems," *Commun. ACM* 31, no. 7 (July 1988): 836–852.

archive—who is it for? How big is it, and how big do you want it to be? What visual and dimensional language can you employ to help the user navigate?

Looking to history can also temper the conclusions we attempt to draw from archives. The web's massive structure suggests total comprehensiveness—a true universal library—and understanding the limits of its scope *as well as* the limits of its context allows us to view its contents with greater nuance. This is a crucial question as our linked archives begin to link with one another, such as with linked data and APIs. These create new modes of analysis that suggest an inarguable universality: as danah boyd and Kate Crawford argue, "Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality."[63] A full understanding of the structures and challenges in network- and archive-building gives us one view into what boyd and Crawford call the "models of intelligibility" and "inbuilt limitations" of big data itself.

The web has evolved since its inception to support much more complex applications, structures, and graphics. But any new developments and platforms must be grafted onto the web rather than rethinking its core structure. I have aimed to suggest how historical context and understanding of the challenges and structures of early hypertext and information management systems can help to explain the powers and limitations of the web. These knowledge systems can also provide inspiration for new solutions: web-based digital archives could aim to mimic or approximate multiple linking, transclusions, or high-level graph views, all while keeping in mind their respective archive's size, shape, and scope.

[63] Kate Crawford and danah boyd, "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society* 15, no. 5 (June 2012): 662–679.

Chapter 4

Networking the News

In the previous chapters I outlined how archives, and critical readings of them, have expanded from fixed and graspable entities 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.

In the publishing world, recent years have seen two simultaneous trends that point to a fundamental shift in the function of mainstream news on the web, and a corresponding reformulation of the roles and practices of journalists. First, some new digital publishers are championing an "explainer" model of journalism. While explaining the news is nothing new (and there has been a Pulitzer for explanatory journalism since 1985), explainer outfits take it a step further; featuring headlines like "Everything you need to know about the government shutdown" or "40 charts that explain money in politics," explainer journalism aims to be "as good at explaining the world as it is at reporting on it."[2] Second, legacy publishers have led a new

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

[2] Ezra Klein, "Vox is our next," The Verge, January 26, 2014, accessed April 20, 2015, `http://www.theverge.com/2014/1/26/5348212/ezra-klein-vox-is-our-next`.

focus on renewing and reanimating their historical archives. Whether they're cherry-picking and republishing old curiosities, providing subscribers with an interface to dive in and leaf through, or leading automated and crowdsourced projects aimed at organizing and structuring old content, legacy media has jumped at the chance to offer something that their born-digital rivals can't: a new take on providing context, a source for research and remix, and a trove of institutional memory and brand history.

These two trends reflect a new online media landscape, one which has seen journalists adapt by amplifying their role as explainer, verifier, and context provider rather than news breaker or scooper. The newsrooms that employ these journalists must adapt in turn. Publishers have a newfound opportunity to thrive in the information industry, but in order to compete in an information ecosystem currently dominated by Silicon Valley and Wikipedia, this new mentality cannot be simply verbal or piecemeal. It requires updated technologies *as well as* cultural change.

Journalism has always hung in the balance between providing the facts and weaving a narrative. Michael Schudson frames the rise of popular newspapers in the 1890s as a pull between two journalisms, one telling "stories" and the other providing "information." A fast-paced breaking event requires quick information, while contextualized longer-form histories and multimedia are more likely to serve an "aesthetic" function.[3] Journalists sometimes draw a similar divide between "stock" and "flow": the constant stream of information built for *right now*, versus the durable, evergreen stuff, built to stand the test of time.[4] These are distinct but overlapping functions, and publishers often conflate the two. After all, longform stories also contain a great deal of information, even if that information is presented in a wall of text. In the digital age, the divide between stories and information can be blurred even further if we start to consider a news story as a curated, contextualized, and linked collection of facts and information. Through this framework, we might see structured and archive-oriented journalism begin to emerge as a hybrid of technology adoption and

[3] Michael Schudson, *Discovering the News: A Social History of American Newspapers* (New York: Basic Books, 1978), 89.

[4] Robin Sloan, "Stock and Flow," Snarkmarket, January 18, 2010, accessed April 20, 2015, `http://snarkmarket.com/2010/4890`.

journalistic practice.

As such, this chapter is divided into two sections, one cultural and the other technical. First I will outline the origins of archive-oriented and explainer journalism, suggesting that the rapid proliferation of new content and connections have fundamentally altered journalism's roles and practices. The second section will consider the ways that publishers can adopt a technical infrastructure to support these new roles, first by analyzing in detail the architecture of a digital news story, then offering frameworks, tools, and techniques that might help to link archives and structure stories for future archival value.

4.1 Context in context

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."[5] 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 endeavoring to use their rich histories to create something new. Back archives like *Harper's* and *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 its celebrated, leaked *Innovation* report of 2014, suggesting that a clever use of archives would enable the Gray Lady to be "both a daily newsletter and a library."[6] The report suggests that arts and culture content, more likely to be evergreen, could be organized by relevance instead

[5] Nicole Levy, "Time.com opens its 'Vault'," Capital New York, November 12, 2014, accessed April 20, 2015, http://www.capitalnewyork.com/article/media/2014/11/8556503/timecom-opens-its-vault.

[6] *Innovation* (New York Times, March 24, 2014), 28, accessed April 20, 2015, https://www.scribd.com/doc/224332847/NYT-Innovation-Report-2014.

of chronology, and that topic homepages should be more like guides than wires. It also enumerates successful experiments with repackaging old content in collections, organized by categories and themes; users could also create collections of archival stories, allowing reader interactivity without risk to the *Times* brand. The report argues that by creating "no new articles, only new packaging," the *Times* could easily give new life to old content.[7]

Along with the archival trend, we have recently witnessed an uptick in explainer journalism and an intense focus on providing context for online readers. Explainer journalism aims to take a step back from the immediate news event and place it in a larger phenomenon. It reflects a deep shift in the roles and practices of online journalists: as news is increasingly broken and scooped on blogs and social media, journalists are increasingly becoming summarizers, filterers, and context providers. In the archive and explainer movements, we see a pattern among some news outlets attempting to evade and reconsider the news cycle's obsession with speed and feeds, focusing instead on building structures and sustainability into journalism—in other words, on nurturing the archive in order to produce deep, contextualized stories. 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 comparing the challenges of legacy media archives and newer forms of explainer journalism, we can gain a sense not only of how print and digital media differ as objects and media texts, but also of how journalistic practice has changed across both sectors.

Journalism has, of course, changed drastically since it moved onto the web. When a newsworthy event occurs now, it is likely to appear on social media within seconds, broken by eyewitnesses with cell phones rather than a reporter with a mic and camera crew. Alfred Hermida calls this phenomenon "ambient journalism": news now happens continuously, and it comes from everyone.[8] This is especially true for distant events that are hard for newsrooms to cover in person; as foreign bureaus close

[7] *Innovation*, 34.

[8] Alfred Hermida, *Twittering the News: The Emergence of Ambient Journalism*, SSRN Scholarly Paper ID 1732598 (Rochester, NY: Social Science Research Network, July 8, 2010).

around the world, this is an increasingly normal reality. Some citizen journalism initiatives are attempting to codify and normalize this new form of news production;[9] but wherever it's published, this leaves mainstream journalists out of the breaking news loop. Reporters are relegated to picking up the digital pieces of stories as they are happening. Finding themselves in front of a computer for both reporting *and* writing, a journalist's job now depends on strong online research skills as much as following beats and interviewing sources.

Mark Coddington sees journalists as moving away from "what they consider the core of 'reporting,'" while non-professionals are moving towards it.[10] Gathering and breaking the news is a losing battle for mainstream journalists, and increasingly journalists are justifying their profession in different ways. Reviewing reporters' language around the WikiLeaks document trove, Coddington argues that journalists "cast themselves fundamentally as sense-makers rather than information-gatherers during an era in which information gathering has been widely networked."[11] In discussing their value in the WikiLeaks case, journalists emphasized "providing context, news judgment, and expertise."[12] Others noted similar trends and tendencies, even many years earlier; Kovach and Rosenstiel's *The Elements of Journalism* quotes Xerox PARC director John Seeley Brown's assertion that "what we need in the new economy and the new communications culture is sense making."[13] Journalist Amy Gahran agrees: "today's journalists can—and probably should—consciously shift away from jobs that revolve around content creation (producing packaged 'stories') and toward providing layers of journalistic insight and context on top of content created by others."[14]

[9] See Ethan Zuckerman, "International reporting in the age of participatory media," *Daedalus* 139, no. 2 (April 1, 2010): 66–75.

[10] Mark Coddington, "Defending judgment and context in 'original reporting': Journalists' construction of newswork in a networked age," *Journalism* 15, no. 6 (August 1, 2014): 682.

[11] Ibid., 678.

[12] Ibid., 689.

[13] Bill Kovach and Tom Rosenstiel, *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect* (Crown/Archetype, 2001), 19.

[14] Amy Gahran, "Swimming Lessons for Journalists," PBS Idea Lab, July 4, 2008, accessed March 15, 2015, `http://www.pbs.org/idealab/2008/07/swimming-lessons-for-journalists005/`.

Rather than being completely new, this trend is a sort of warped resurgence of traditional journalistic practices. Dan Gillmor praises legendary journalist I.F. Stone for his "old-fashioned reporting," which Victor Navasky describes: "To scour and devour public documents, bury himself in The Congressional Record, study obscure Congressional committee hearings, debates and and reports, all the time prospecting for news nuggets (which would appear as boxed paragraphs in his paper)... He lived in the public domain."[15] These are librarians' verbs—scour, devour, bury, study, prospect—and the techniques don't seem so old-fashioned now. Schudson noted a turn towards interviewing as the primary journalistic technique in the late 20th century, but new affordances in computer-assisted reporting—such as social media analysis tools and document repositories like DocumentCloud—have brought a resurgence of document-oriented sourcing. The "holy trinity" of newsgathering objects that media professor C.W. Anderson identifies—observation, documents, and interviews—has recently tipped back towards documents.[16]

But in addition to repackaging original documents, many journalists are now repackaging other journalism. In a 2013 study, Anderson compares the practices of "serious journalists" and the news aggregators who republish them, to look for the boundary line in professionalism between original and repurposed content. Journalism and aggregation are generally taken to be at odds when considering newsroom expertise, but Anderson finds specific and often journalistic skills needed to be an aggregator. Calling them "hierarchizers, interlinkers, bundlers, rewriters, and illustrators of web content," he suggests that *news judgment* is their fundamental skill.[17] News judgment is one of the core journalistic skills that Coddington identifies, and it undoubtedly requires an editorial eye. Meanwhile, "serious journalists" increasingly explain and contextualize breaking and ongoing news in an aggregational fashion. Joel Achenbach of the *Washington Post* wrote in 2014 that "Journalism is aggregation"; he noticed that in conducting rushed interviews with expert sources, he was

[15] Dan Gillmor, *We the Media: Grassroots Journalism By the People, For the People* (O'Reilly, 2006), 3-4.

[16] Coddington, "Defending judgment and context in 'original reporting'," 680.

[17] C. W. Anderson, "What aggregators do: Towards a networked concept of journalistic expertise in the digital age," *Journalism* 14, no. 8 (November 1, 2013): 1008–1023.

doing the same skimming, lifting, and summarizing of his sources that aggregators do with theirs.[18] In a Reddit conversation with New York Times media columnist David Carr, a reader asked if Carr was ever frustrated by news aggregator Gawker stealing traffic from his original writing; the *Innovation* report likewise laments a 161-year-old archival *Times* story that went viral on Gawker rather than the *Times*.[19] But Carr replied by pointing out that Gawker was starting to create original reporting, while the Times itself is dipping into types of aggregating.[20] These professional lines seem to be increasingly blurred.

From the perspective of a digital media archaeologist, a web historian, or a software engineer building search algorithms, the difference between journalism and aggregation is that aggregating leaves a trace. It does so usually in the form of a hyperlink (if the aggregator is being friendly), or a tag, quote, or snippet from the original article. This trace is crucial; it not only helps search engines understand influence, but it forms a dialogue and joins a network, allowing it to be resurfaced and mined in the future. This could be one reason that aggregators often win out over original sources in battles over traffic. This is not to say that aggregating is better or more influential than original reporting—only that it plays by the citation rules of the web, and is rewarded for it. Of course, sometimes journalists can't publish or link to their sources, and it would be unduly cumbersome for a reporter to publish hours of audio transcripts just to provide context, not to mention the complications of anonymous or off-the-record comments. Some bloggers and new journalists suggest that it is a best practice to publish every source in full context, but forcing a journalist to obsessively link to every source risks stifling journalistic practice. All the same, the hyperlink and the tag combine to allow for a new standard of citation, reference, and context provision for news on the web, and journalists would be well served to think about the explicit and latent links buried within their stories.

[18] Joel Achenbach, "Journalism is Aggregation," *Washington Post* (April 9, 2014), accessed April 16, 2014, http://www.washingtonpost.com/blogs/achenblog/wp/2014/04/09/journalism-is-aggregation/.

[19] *Innovation*, 28.

[20] David Carr, "IAmA columnist and reporter on media and culture for the New York Times.," reddit, accessed March 15, 2015, http://www.reddit.com/r/IAmA/comments/16k598/iama_columnist_and_reporter_on_media_and_culture/c7wt5ko.

While I will consider the more technical aspects of links and tags in section 4.2, this sets the stage for considering how the dual trend of archiving and explaining reflects a new focus in the work of digital journalists. While many have discussed the blending of journalist and computer scientist in so-called "data journalism" (for instance, consider the dual journalism and computer science degree now offered by Columbia University), I also posit a link between journalists and librarians/archivists in the current moment's focus on context and explanation. Emphasizing such a connection could lead to a journalistic focus on both more rigorous research and citation skills, and long-term information sustainability.

4.1.1 The archive drive

Legacy media's turn to archives is a reflection of many uncertainties about the future of journalism. After all, it seems antithetical for news to focus on history; news has forever been focused on the *new* and *now*. The pace of *now* has only accelerated online, as newsrooms shift from daily or weekly print schedules to continuous publishing and posting cycles. For a reporter, this creates a work environment of real-time frenzy; a reporter might have live statistics on her article's performance in one window, while she furiously scans a flurry of tweets looking for her next scoop in another. It's no wonder that newsrooms haven't traditionally focused on their archives; for today's reporter, media loses value with each passing second.

But this everyday practice—manifested in real-time feeds, streams, and notifications on Twitter, Reddit, and RSS—is at odds with the plethora of old stories, images, facts, and quotes that are perpetually accessible on a publisher's website. Some stories are part of larger story arcs that have lasted for years, while others lie dormant for decades before resurfacing due to new facts or a notable anniversary. These old stories are continuously updating too; if you click on one or add a comment, you register another pageview or annotation in the database. The past and present collide in the digital news database. Archived stories are consistently changing, acquiring new clicks, new links, new topics and tags to reflect ongoing events. Drawing on my discussions of containment and connection in section 2.2, this strikes a sort of paradox, where the archive contains the object but keeps forging connections. This paradox emphasizes the continual relevance of the archive to the present, and the need to balance the steady stream of news with a dose of history.

The web's archival affordances are crucial for media scholar Peter Dahlgren, who sees archivality as one of the five components of the "media logic" of cyberspace.[21] For Dahlgren, archivality forms a symbiosis with hypertextuality—another web logic—which enables new and more usable archives. The end result is that "users of cyberspace for journalism are in principle no longer so bound to the present."[22] Even as reporters are focused on *right now*, their users might be more interested in *back then.* But this is not necessarily borne out in publishers' insights about their readers. Most publishers are sure that archives are valuable, but they have vague answers for exactly *why.* Moreover, the answer could be different for each publisher. While the *Times* asserts a competitive advantage over born-digital upstarts like Vox and BuzzFeed, most do not go beyond suggesting that it is a unique and more or less rights-free repository of content. The University of Missouri's Reynolds Journalism Institute offers that "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."[23] They also suggest that archives "serve as a form of institutional knowledge," allowing new journalists to get up to speed on the historical context of an ongoing story. This is especially crucial for new hires at highly local or specialized publications, whose archives often contain a unique and unmatched amount of information on a certain community or genre.

The Reynolds Institute led a 2014 survey of 476 news websites, finding that 88–93 percent of them highly value their archives. But the origins of this study are

[21] Mark Deuze only names three "logics"—hypertextuality, multimediality, and interactivity—but Dahlgren adds "figurational" and "archival."

[22] Peter Dahlgren, "Media Logic in Cyberspace: Repositioning Journalism and its Publics," *Javnost - The Public* 3, no. 3 (January 1, 1996): 66.

[23] Edward McCain, "Saving the news: When your server crashes, you could lose decades of digital news content - forever," 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`.

telling; the school's *Columbia Missourian* paper lost 15 years of stories and seven years of images in a single 2002 server crash. According to the study, despite the nearly universal lip service paid to archives, about a quarter of news websites had lost significant portions of their archives due to technical failure. Missouri journalism professor Tom Warhover provides a series of examples to emphasize the devastation: "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."[24] These examples showcase the latent and often hidden value of the archive, not only as a product but as a deep journalistic resource and form of institutional memory.

The *Times* asserts that "our rich archive offers one of our clearest advantages over new competitors," though they have done little to mine its value, "largely because we are so focused on news and new features."[25] The *New Yorker* also found some success in summer 2014 by opening up its archives while it experimented with a new paywall. As readers raced to download classics written by Ariel Levy, Philip Gourevitch, and Kelefa Sanneh, the announcement got the public talking about the forthcoming redesign and the rich history of the *New Yorker* brand. While the archive summer saw decent engagement, the work has paid off since reinstating the paywall; web editor Nicholas Thompson says that in the summer "there wasn't a massive increase. . . [w]hat's weird is we launched the paywall, and *then* there was a massive increase."[26] While the *New Yorker*'s success is on the back of its redesign and paywall, its archive played a foundational and unmeasurable role.

Now the *New Yorker*'s archives are back under lock and key, as they are for most publishers. Newspapers see a synergy between digital archives and paywalls, considering that the archives could be a valuable draw for subscribers, and especially

[24] McCain, "Saving the news."

[25] *Innovation*, 28.

[26] Justin Ellis, "After the archive came down: The New Yorker's revamped paywall is driving new readers and subscribers," Nieman Journalism Lab, March 11, 2015, accessed April 20, 2015, http://www.niemanlab.org/2015/03/after-the-archive-came-down-the-new-yorkers-revamped-paywall-is-driving-new-readers-and-subscribers/.

an enticement for researchers who might be willing to pay for a subscription. The typical archive diver is seen as a dedicated browser or dedicated searcher; interest in the archive is already assumed at the outset. For the rest of the population, the archive is locked away. This runs counter to the linking and sharing economy of the web; as many publishers have experimented with the tension between linking and locking, the archives have remained firmly locked away. Joshua Rothman, the *New Yorker*'s archives editor, thinks that the most crucial element of enlivening the archive is to simply make it available in the first place; it is hard to predict or control when an old story will return to relevancy or gain new attention, but it cannot happen without users being able to find and access it.[27] This suggests a need to treat archival stories with a nuanced paywall approach. Most publishing archive landing pages hardly even allow even a "preview" of the archival content or interface, and the more nuanced paywalls (for instance, ones that let readers access six stories a month) are virtually nonexistent for archives. Some publishers, *Time* included, will lift the paywall on certain "whitelisted" stories, ones that might have jumped back into relevance and that new stories are linking to. This is a start, but defaulting to locking away limits the possibilities for serendipitous encounters with archives.

Still, archives don't directly drive clicks, and legacy publishers have well-documented financial troubles, so their archives have taken a hit. While most newsrooms have a library or research center, the ranks of newsroom librarians are dwindling. Over 250 news librarians lost their jobs in the U.S. from 2007 to 2010, and membership in the Special Libraries Association News Division has steadily dropped. Some news libraries and research centers have been completely shut down, outsourced to vendors like LexisNexis.[28] As newspapers endeavor to enter the information age, it is a shame to see them losing their information professionals. Librarians provide crucial research and fact-checking skills. Amy Disch, chair of the Special Libraries Association News Division, speaks to the traditional division of skills between reporter and librarian

[27] Liam Andrew, *Interview*, in collab. with Joshua Rothman, March 2, 2015.

[28] Craig Silverman, "Endangered Species: News librarians are a dying breed," Columbia Journalism Review, January 29, 2010, accessed April 20, 2015, `http://www.cjr.org/behind_the_news/endangered_species.php`.

in the newsroom: “We can find the information in a lot less time because we know how to drill down in a database. We know good sources to go to where you can quickly find information, so we can cut a lot of time for [reporters] and leave them to do what they do best, which is interviewing and writing. I have my specialty, and they have theirs.”[29] Librarians are also the shepherds of the archive, tagging and structuring new stories and continuously maintaining them as history develops. Most major news organizations no longer have the library manpower to manually review every new story for proper tagging and indexing; this also points to their limited capacity for maintaining the topic taxonomy, or experimenting with new indexing and metadata projects.

A few publishers have thrown up their hands altogether, relying on third-party services to organize and provide access to their own archives.[30] Reporters might suddenly see old stories disappeared from the web, locked away behind services like ProQuest and 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 story is not just text, and increasingly, an archive includes images, videos, charts, maps, interactives, facts, statistics, quotations, comments, and annotations. This will become increasingly important as media evolves in a “post-text” web; the next generation of media companies cannot rely alone on text search to access their past.[31]

Due to the ambiguous value of news archives, it seems to go without saying that old stories should be saved, but it’s harder to know exactly what to do with them. Saving them turns out to be the expensive part; the cost of digitizing, indexing, and hosting gigabytes of old content is no small investment. Some publishers stop here: they might offer a search interface for staff members and researchers, as well as a PDF-style flipbook view that allows obsessively curious subscribers to leaf through the “original” paper or magazine. These interfaces will serve those with a research

[29] Silverman, “Endangered Species.”

[30] Jim Romenesko, “U.S. News deletes archived web content published before 2007,” Romenesko, February 18, 2014, accessed April 20, 2015, http://jimromenesko.com/2014/02/18/u-s-news-deletes-content-published-before-2007/.

[31] Felix Salmon, “Why I’m joining Fusion,” Medium, April 23, 2014, accessed March 9, 2015, https://medium.com/@felixsalmon/why-im-joining-fusion-4dbb1d82eb52.

interest or innate curiosity, but this only scratches the surface of the archive's potential as a serendipitous window into past insights for journalists and casual readers alike.

A publisher's archive will often turn up in specific articles geared towards history. At *Time*, archive editor and curator Lily Rothman (no relation to Joshua Rothman at the *New Yorker*) digs out stories and quotes from the history of *Time*, ranging from historical interests ("Read TIME's Original Review of *The Catcher in the Rye*") to ephemeral oddities ("13 Weirdly Morbid Vintage News Stories"). Editor Samuel Jacobs likened Rothman to a radio D.J., highlighting singles from the archive to entice readers to pay for the whole collection.[32] Other historic deep-dives might occur in weekly columns or "long history" forays by individual journalists. A Sunday Times article, for instance, might take a historic look at a particular person, neighborhood, or community. These projects have a chance to draw attention to the past through curation as well, by drawing out and resurfacing old stories, photographs and statistics; such projects are a promising start, but they tend to be isolated endeavors, relegated to a single story. Sometimes willfully nostalgic, they do not bring the archive fully into dialogue with ongoing events, whether in the research or design process.

Topic pages suffer from lack of organization and explanation as a result. Imagine navigating to a topic page, such as one of *The New York Times'* over 5000 pages ranging from "A.C. Milan" to "Zimbabwe," and seeing a map or timeline of the most important stories in the *Times'* history. What about viewing a network of stories as a graph of influence and dialogue, or pulling out the most relevant quotes, images, videos, or facts from the story to highlight for the user? If the *Times* hasn't written about A.C. Milan in a while, they could suggest a story about a rival soccer team, or Italian football clubs in general. Search interfaces and topic pages can function like research librarians, retrieving information and context, instead of a list of stories and items. Like any good librarians, if the information is not at hand, they could point the reader towards where to find it.

Such a future requires smart use of indexing and metadata schemes, which I will consider in the following section; but it also requires a cultural shift in both the

[32] Levy, "Time.com opens its 'Vault'."

conception of the archive's audience and the ways that journalists think of their stories' value. For now, legacy publishers seem to know well that their archives are a competitive advantage, and an asset that sets them apart from born-digital upstarts. *The Nation*'s editor and publisher Katrina vanden Heuvel suggests that "a clever use of archives is kind of an explainer 2.0," directly placing the two movements in comparison and competition.[33] But ironically, publishers with short histories are the ones that have been emphasizing context. It's possible that in aiming to revitalize the archive and blend the present with the past, archive-oriented publishers might be bringing past conventions back into the present.

4.1.2 Explaining the news

The *Times' Innovation* report names Vox as one of its direct competitors and threats, a new digital upstart innovating with new models of news production. This was an auspicious start; at the time of the report, Vox had not even launched. Its parent company, Vox Media, already counted several special interest sites in its ranks, but Vox was its first general interest property. Vox also had a few big names behind it already, namely co-founder Ezra Klein, who left his position at the helm of the *Washington Post*'s Wonkblog to start this digital outlet. Vox has quickly become a fixture of digital start-up news, as well as the poster child of the explainer movement, along with Nate Silver's *FiveThirtyEight* and the *Times'* own Upshot venture. By taking a broad, data-oriented stance on reporting, Vox aims to infuse deep context and data into its journalism, technology, and business decisions alike. Vox's signature feature is its "card stacks": collections of reusable and editable snippets of facts, quotes, maps, and other media, collected and synthesized by a journalist. With titles like "Everything you need to know about marijuana legalization" or "9 facts about the Eurozone crisis," the stacks subdivide into question- and statement-driven "cards," like "What is marijuana decriminalization?" or "Debt didn't cause the crisis." Readers can navigate sequentially, or leap from card to card. The final option on each card is the same: the "Explore" button takes the reader back to the top of the stacks.

[33] Levy, "Time.com opens its 'Vault'."

The purpose of the card stacks is similar to many legacy institutions' reasons for protecting their archive: to avoid repeating work. For Klein, "the biggest source of waste is everything the journalist has written before today."[34] Through the card stacks, we see a steady accumulation of a new kind of publishing archive, one oriented around questions and statements rather than topics and tags. The card stacks are not only a public-facing form of archive, they are a core feature of the Vox product; Vox is not just selling stories (content), but a story *structure* (a container). This leads to a shift in the role of the journalists, who explain and contextualize as journalists always have, but with an eye towards updating the stacks as well as pushing new stories.

Klein has Wikipedia in his sights, suggesting in a *New Yorker* interview that "I think it's weird that the news cedes so much ground to Wikipedia. That isn't true in other informational sectors."[35] Wikipedia is the looming giant in context provision; publishers understandably ask how they could do any better without unlimited staff and budget. But Wikipedia is a generalist's resource, and its readings on more specialized and newsworthy topics can be dense and dry. Some topics, especially in science or medecine, assume a reader's technical knowledge from the outset. Linking to Wikipedia also, of course, takes the reader away from the publisher's site, at the risk of leading them down a rabbit hole of Wikipedia links in a quest for more context. Vox's card stacks are a sort of journalistic alternative to Wikipedia, a wiki "written by one person with a little attitude," as co-founder Melissa Bell puts it.[36] This allows for customized context, created "in-house."

Klein envisions a virtuous cycle of curiosity and information: "the card stacks add value to the news coverage. And the news coverage creates curiosity that leads people to the card stacks."[37] The symbiosis between news coverage and cards is clear

[34] Leslie Kaufman, "Vox Takes Melding of Journalism and Technology to a New Level," *The New York Times* (April 6, 2014), accessed April 20, 2015, `http://www.nytimes.com/2014/04/07/business/media/voxcom-takes-melding-of-journalism-and-technology-to-next-level.html`.

[35] Joe Coscarelli, "Ezra Klein on Vox's Launch, Media Condescension, and Competing With Wikipedia," *New York Magazine* (April 11, 2014), accessed April 20, 2015, `http://nymag.com/daily/intelligencer/2014/04/ezra-klein-interview-vox-launch.html`.

[36] Kaufman, "Vox Takes Melding of Journalism and Technology to a New Level."

[37] Coscarelli, "Ezra Klein on Vox's Launch, Media Condescension, and Competing With

for journalists who need to quickly fact-check and contextualize, but the magical formula of news coverage that "creates curiosity" for a reader is a vague and difficult goal. Some think that newsreaders are just looking for news, not information and context. Others believe that Vox is diving too deep into the information space; given that even teams of trained librarians struggle to organize information, how can a team of journalists hope to keep up? They have "a huge challenge, due to the rapid decay of facts," according to Craig Silverman, who suggests that Vox could use a librarian.[38] In order to be feasible, each card needs to be both independent and linked; the cards are adoptable as standalone pieces, but they also have to add up to a coherent narrative for linear reading. How can a journalist keep a card independently updated, and know that it is up-to-date? Consider, for instance, Vox's card on marijuana legalization, which has already been updated 40 times between April 2014 and February 2015.[39] Scientific research on marijuana's health effects is known to be limited; what happens to these preliminary studies when they are replaced by more rigorous ones? How will a journalist be alerted when a new research paper debunks an old fact or figure? What happens to such a card if marijuana becomes legalized across all 50 states?

But Vox doesn't say their cards contain everything, just "everything you need to know." While the card stacks might be Vox's technological foundation, this economical, itemized, somewhat glib approach to contextualizing the news is its journalistic signature. Some critics have lambasted the site for the hubris of its mission; it doesn't get much bigger than "explaining the news," and it might be best left to a distributed group of subject-matter experts (or the public, in Wikipedia's case) rather than a core of time-strapped journalists who are often distant from the topics that they cover. Vox has already gotten some things wrong, and Deadspin gleefully and sometimes unfairly picked apart a full 46 Vox corrections from its first year.[40] Others question

Wikipedia."

[38] Craig Silverman, "Why Vox (and other news orgs) could use a librarian," Nieman Journalism Lab, April 22, 2014, accessed April 20, 2015, http://www.niemanlab.org/2014/04/why-vox-and-other-news-orgs-could-use-a-librarian/.

[39] German Lopez, "Everything you need to know about marijuana legalization," Vox, accessed March 15, 2015, http://www.vox.com/cards/marijuana-legalization.

[40] Kevin Draper, "46 Times Vox Totally Fucked Up A Story," Deadspin, December 30, 2014, accessed April 20, 2015, http://theconcourse.deadspin.com/46-times-vox-totally-fucked-

the site's tone; music journalist David Holmes lamented that Vox's model gives the reader "just enough information. . . to *sound* smart," appealing primarily to "the type of person who's afraid of sounding gauche at your next dinner party."[41] "Everything you need to know" sounds final, running counter to the very idea that knowledge is interlinked, flexible, and intersubjective: everything *who* needs to know? Vox's style of explainer journalism can feel cold and distant; it's more likely to explain why people are excited about a new movie than to get *you* excited about the movie.

Vox is far from the first or only explainer, and the concept of explaining the news is a core journalistic principle. A 2001 Pew Center survey of newspaper editors concluded that they wanted to be "news explainers" first and foremost, ahead of "news breakers" or "investigative watchdogs."[42] But in a 2008 article called "National Explainer," Jay Rosen accused editors of not staying true to their mission: journalists are not good at explaining the news and providing context.[43] Instead, they focus too much on incremental and episodic updates, many of which go over the head of readers who haven't been following. Rosen likens the process to pushing software updates to a computer that doesn't have the software installed.

Journalists "don't do a very good job of talking about the beginning and what got us to this point where it became news," according to Alex Blumburg.[44] Even the occasional explainer that gets it right ends up in the flow of the same old information; Rosen argues that explainers like David Leonhardt's credit crisis piece in *The New York Times* "should have been a tool in the sidebar of every news story the Times did about the mortgage mess." The little "what's this?" link that pops up on occasional news websites is "not about web design. That's a whole new category in journalism

up-a-story-1673835447.

[41] David Holmes, "How Rap Genius and explainer sites are killing music journalism," PandoDaily, January 20, 2015, accessed March 16, 2015, http://pando.com/2015/01/20/how-rap-genius-and-explainer-sites-are-killing-music-journalism/.

[42] Pew Research Center, *Journalism Interactive: Survey Pinpoints a Sea Change in Attitudes and Practices; Engagement Defines New Era of Journalism*, July 26, 2001, accessed March 16, 2015, http://civicjournalism.org/about/pr_interact.html.

[43] Jay Rosen, "National Explainer: A Job for Journalists on the Demand Side of News," PressThink, August 13, 2008, accessed April 20, 2015, http://archive.pressthink.org/2008/08/13/national_explain.html.

[44] Ibid.

that I fear we do not understand at all." Rosen went on to create explainthis.org, a site for people to admit what they don't know; journalists, the site promises, are "standing by."[45] Now defunct, explainthis.org was like a library reference desk, staffed by the public and monitored by journalists. A peer of StackOverflow and ancestor to Quora, it is organized around questions rather than topics, discussed by the public and monitored by journalists. It requires someone to be curious enough to ask the question, however. Rosen and Klein tout the explainer's role as a driver of deeper interest in a topic, but here we're already expected to be interested.

At a 2010 South by Southwest conference panel called "Future of Context," Rosen outlined the reasons explanation is needed and why it wasn't taking off. He cited both design and institutional problems; the prestige and real-time excitement of breaking (rather than explaining) news, as well as the explainer format getting lost in the shuffle of other news items.[46] Metrics like clicking, watching, and even spending time on a site are not measuring the level of understanding or knowledge gained. Panelists like NPR's Matt Thompson and Apture CEO Tristan Harris turned more towards the systems and technologies that compound the problem. Harris offered an "object-oriented journalism" model, which asks journalists to "think like an engineer" and never do work they can't reuse. Thompson considered the potentials of a context-oriented website; how could you take a topic page and make it more than a random collection of links? What would a site look like if it were structured around systems instead of stories?[47]

Vox is not a complete shift into this territory, but it's a start. The card stacks might initiate a subtle shift in the ways that journalists think about the value of their

[45] "ExplainThis.org," March 5, 2010, accessed March 16, 2015, https://web.archive.org/web/20100305205057/http://explainthis.org/.

[46] Jay Rosen, "News Without the Narrative Needed to Make Sense of the News: What I Will Say at South by Southwest," PressThink, March 7, 2010, accessed April 20, 2015, http://pressthink.org/2010/03/news-without-the-narrative-needed-to-make-sense-of-the-news-what-i-will-say-at-south-by-southwest/.

[47] See Rosen, "News Without the Narrative Needed to Make Sense of the News"; Steve Myers, "Liveblogging SXSW: The Future of Context in Journalism," Poynter, March 15, 2010, accessed April 20, 2015, http://www.poynter.org/news/101399/liveblogging-sxsw-the-future-of-context-in-journalism/; Elise Hu, "Contextualizing Context," Hey Elise, March 15, 2010, accessed April 20, 2015, http://www.heyelise.com/2010/03/15/contextualizing-context/.

content. A reporter who is writing about a new study on marijuana's health effects could do so by editing an existing card, or by writing a new card; this card could also be published to Vox as a standalone story. If the existing card already featured a few other helpful background facts, or a nice map or infographic, the reporter could attach it to the new story too. This not only saves time and adds value to the story, but it produces a subtle shift in the journalist's practice; the story is itself a collection, and it's part of other, bigger collections. The story is here to stay.

Vox has also experimented with another simpler but novel use of content: republishing evergreen stories. In December 2014, during a typically slow news cycle, Vox edited, updated, and republished 88 old stories from their archive.[48] Some of these went through a heavy round of editing, while others were republished nearly as-is; in either case, no one seemed to notice that the content was old. Some stories that didn't draw readers the first time around experienced a second life. Executive editor Matthew Yglesias explains, "On the modern web, content tends to arrive via miscellaneous streams rather than coherent chunks. So the meaning of strict chronology is breaking down regardless of what publishers do."[49] This approach is a sort of workaround of the traditional stream-based publishing format; since we haven't found a good way to differentiate evergreen stories in our feeds and streams, Vox is just pushing them repeatedly. But the experiment was enough of a success that writers continue to refresh old content, united in Vox's quest to become a "persistent news resource."[50]

In this section I have aimed to outline the ways in which journalistic practice is changing on the web. Journalists are adopting the role of sense-maker, taking a step back from the immediate event and endeavoring to place it in a larger phenomenon. This is manifested in a corresponding reliance on documents more than interviews and observation, which blurs the line between journalist and aggregator and offers a new conception of a news story as a document-driven collection of linked information.

[48] Matthew Yglesias, "Refreshing the evergreen," Vox, January 15, 2015, accessed February 4, 2015, `http://www.vox.com/2015/1/15/7546877/evergreen-experiment`.

[49] Ibid.

[50] Ibid.

4.2 The size and shape of stories

Here I will turn to the technical challenges inherent in linking this information, first by examining existing practices in digitizing and researching in digital archives, then by looking to the ways that digital stories might be structured, treated, and archived as a result.

As I outlined in previous chapters, every archive has its own size, shape, and properties that the digital has complicated and exploded. The same goes for stories, too. The structure of a story and conception of its has rapidly changed. Even a typical newswire that delivers last night's sports scores—seemingly an unchanged art from its paper-news origins—now offers new context, such as an interactive chart with the box scores, or inline links to the teams' pages. Journalists have sometimes been slow to adopt these hypertextual and interactive elements, but as some more recent studies show, the pace is on the rise.[51]

Consider a typical old and archived newspaper story, such as *The New York Times'* March 14, 1969 feature "Lindsay Warned on School Budget," available on the TimesMachine. It appears on the paper's front page, sharing space with news of the safe return of the Apollo 9 mission, as well as eight other stories covering Washington, Russia, and local New York affairs. The story has several components: a title, a subtitle, a lead paragraph, and four distinct sections, spanning two pages. Its digitized version on TimesMachine exposes much of this structure: the title, subtitle, lede, byline, publish date, and page numbers all appear as separate fields in its metadata. TimesMachine also offers five "subjects" that the story belongs to: Education and Schools, Finances, New York City, State Aid, and Welfare Work.[52]

The TimesMachine is perhaps the most well-indexed publishing archive, but even these tags don't give a glimpse of the New York City public education system in

[51] See, e.g., Mark Coddington, "Building Frames Link by Link: The Linking Practices of Blogs and News Sites," *International Journal of Communication* 6 (July 16, 2012): 20; Mark Coddington, "Normalizing the Hyperlink," *Digital Journalism* 2, no. 2 (April 3, 2014): 140–155; Anders Olof Larsson, "Staying In or Going Out?," *Journalism Practice* 7, no. 6 (December 2013): 738–754.

[52] Charles G. Bennett, "Lindsay Warned on School Budget," *New York Times* (March 14, 1969): 1, 27, accessed March 16, 2015, `http://timesmachine.nytimes.com/timesmachine/1969/03/14/90065621.html`.

the 1960s. Today's readers might be wondering who "Lindsay" is (the answer: New York City mayor John Lindsay, who served from 1966 to 1973), or the subtitle's "Doar" (John Doar, prominent civil rights lawyer turned New York Board of Education president). These missing background details leave readers in the dark unless they have been following the story all along, whether the reader is a New York-based newspaper browser in 1969 or a Canadian TimesMachine diver in 2015. The story is replete with facts and background information of its own, but it is all buried deep within the story.

Comparing it to a more recent local education story reveals interesting differences. In a story called "New York Schools Chancellor Replaces 8 Superintendents," the *Times* discusses the latest actions by Carmen Fariña, the New York schools chancellor.[53] The structure is different even in the first two words: Fariña's name is hyperlinked, leading to a deeper piece from the archive on her background and management style. Other, more subtle shifts in language point to a focus on background and context as well; where the 1969 article simply referred to "Lindsay" or "John Lindsay," its 2015 counterpart refers to "former Mayor Michael R. Bloomberg" even though he needs no introduction to most of today's *Times* readers. A later reference to the "Education Department" links out to a Times Topic page about the New York City Education Department, allowing for curious or uninformed readers to see more articles and gain more background. These subtle shifts in writing style add new layers of structure and context to a news story, and allow it to be understandable both to less-informed readers and those who return to the story months or years later.

Also telling in these stories are the many original facts and figures: the names and districts of the eight superintendents being replaced, the names, titles, and relationships of major players in New York's education administration, and precise figures about the city's education budget and class sizes in 1969. But rather than entering this information into a database, reporters reveal them in their writing. Much of this information is then lost to future researchers and subsequent journalists. The

[53] Kate Taylor, "New York Schools Chancellor Replaces 8 Superintendents," *The New York Times* (October 21, 2014), accessed March 16, 2015, `http://www.nytimes.com/2014/10/22/nyregion/in-shake-up-new-york-schools-chief-changes-8-superintendents.html`.

challenge is to properly *structure* this information so that journalists' research can stay fresh and updated, but without requiring onerous data entry or duplicate work.

We can look to computer and information sciences for useful tools, techniques, and approaches for mining this latent structured information. For instance, historical figures like John Lindsay and John Doar are what a computer scientist calls Named Entities; most news stories mention a series of interactions and relationships between newsworthy people, places, organizations, dates, and times, and each of these can be a named entity. For instance, a phrase like "Barack Obama will speak to Congress on Wednesday," has two named entities—Barack Obama, and Congress—that can begin to give computers clues as to what a story is about, and what other stories are related to it. A story referring to his meetings with the Pentagon (another entity) is likely to be about federal defense and security, while another meeting with the Council on Environmental Quality is likely about clean energy. More subtly, a phrase like "Barack Obama will speak to Congress on Wednesday, July 15" adds another named entity: a date. This gives the phrase slightly more structure, which leads to more potential; an automated system could remind a reader on July 15 that Obama is speaking, or it could mine many stories' text for every time Obama has spoken to Congress, displaying it on a calendar or timeline. These small pieces of structure add up on a larger scale, offering many new uses for old news.

In this section, I will first trace the stages and techniques for exposing and gathering insight from archives of news. This will start to reveal how actors and facts interact in a news story, and the size and shape that modern stories can adopt as a result. Then I will consider the potentials for new forms of organization by taking advantage of the structures and systems buried within news stories.

4.2.1 The hypertext newspaper

Newspaper and magazine publishers prove an ideal study for examining the potentials of hypertext archives. Along with the subtle structure of stories, newspapers contain their own structures as well. 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 organized by general relevance that aim to draw the reader deeper. A given page can hold several articles organized by specific topics and subtopics, and an interested reader might be distracted or intrigued by a "related article" next to the one he 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, "jumpable" nature of newspapers, and the endless potential for insightful metadata; newspapers might seem to naturally lend themselves to the digital world.

The typical newspaper's current architecture 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."[54] 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 ledes. A newspaper's layout and seriality therefore scaffold a pseudo-hypertextual structure, one that can be computationally mined for insights.[55]

Given the print newspaper's proto-hypertextual status, it presents a unique metadata challenge for archivists. What metadata is worth saving and breaking down: the text, the subtext, the pictures? The photo or pullquote on the side? For some researchers, placement will be important (was an article's headline on the first page? Above or below the fold? Was there an photo, an ad, or a counterpoint article next to it?). Others could be focus on 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 dive into coverage of a particular

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

[55] Some libraries and cultural heritage institutions are leading these endeavors, such as the Library of Congress' National Digital Newspaper Program, Europeana, and Trove.

event across various journals. In each case, we can glean information from where and when it was published on the page. A newspaper is a very complex design object with specific archival affordances; their structure and seriality make them ripe for unique forms of automated analysis.

Old newspapers are rich archival documents for historians and scholars, because they store both ephemera and history. Newspapers also hold advertisements, classifieds, stock quotes, and weather diagrams. Many researchers rely on such ephemera—James Mussell calls it "a key instrument of cultural memory"—so from the traditional archivist's perspective, everything needs to be considered "stock," stored forever.[56] Paul Gooding, a researcher at University College London, sees digitized newspapers as ripe for analysis due to their irregular size and their seriality.[57] 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. He found that most researchers were not closely reading the newspapers 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. 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.[58] Richard Abel, a scholar of early U.S. and French cinema, breaks down the promises and challenges in generating archival "big data"

[56] James Mussell, "The Passing of Print," *Media History* 18, no. 1 (February 2012): 77–92.

[57] Paul Gooding, "Exploring Usage of Digitised Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online" (DH2014, Lausanne, Switzerland, July 9, 2014), accessed March 9, 2015, `http://www.slideshare.net/pmgooding/dh2014-pres`.

[58] See, e.g., Marlene Manoff, "Archive and Database as Metaphor: Theorizing the Historical Record," *portal: Libraries and the Academy* 10, no. 4 (2010): 388-389; James Mussell, "Elemental Forms," *Media History* 20, no. 1 (January 2014): 4–20.

in his research. Using archival resources like NewspaperArchive.com and GenealogyBank.com to access 1910s newspapers, he found "a wealth of unexpected documents," but he notes the unreliability of completeness and searchability, the collapse of community, and the "*having been there*."[59]

Old newspapers are also replete with images and advertisements, which offer a more vivid sense of history. Leafing through old pages of the *Times* or *Time*, it is the images and ads that jump out beyond the stilted text of news items. But tellingly, many news 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 lost, and many of them are being re-scanned by publishers, at great expense, in order to properly atomize the archive and capture the details that they ignored years ago. This doesn't account for historical problems with tagging images; University of Missouri photojournalism professor Keith Greenwood found that many 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.[60] In a 2014 special issue of *Media History* focusing on digital newspaper archive research, Nicole Maurantonio criticizes newspaper indexes for ignoring the visual in favor of text, "propelling scholars down a misguided path."[61]

Historical images are one of the greatest potential sources of engagement and revenue for news archives, and outlets like the New York Daily News, National Geographic, and Getty do provide and sell some old photographs with historic value. The *Times* is hoping to gain more value from its old images than that, though; its Madison project hopes to crowdsource insight about 1950s *Times* advertisements.[62] Outside of the publishing sphere, researcher Kalev Leetaru took an image-centric approach to

[59] Richard Abel, "The Pleasures and Perils of Big Data in Digitized Newspapers," *Film History* 25, no. 1 (January 2013): 6.

[60] Keith Greenwood, "Digital Photo Archives Lose Value As Record of Community History," *Newspaper Research Journal* 32, no. 3 (2011): 82–96.

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

[62] "New York Times Madison," accessed March 16, 2015, `http://madison.nytimes.com`.

the Internet Archive. The Internet Archive's text recognition (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."[63] Newspaper and journal images provide a richer glimpse of history, and one that might prove more engaging to digital readers than dated text. A photograph reveals a contingent 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 culture and history.

This examination of the digitized newspaper, both its construction and its reception, highlights how each newspaper article weaves a web of context, and it fits into a larger web of context in turn. The seriality and architecture of newspapers and magazines can teach us a great deal about how to gain insight from them, and how to contextualize and contain the new digital stories of today. Stories and publications always had structure, but now we can do more with it.

4.2.2 Atoms of news

In technical terms, a story is usually an object in a database that has 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. While more metadata and structure requires more investment at the outset, smart use of such metadata prepares a story for archival reuse. 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 *Innovation*

[63] 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.

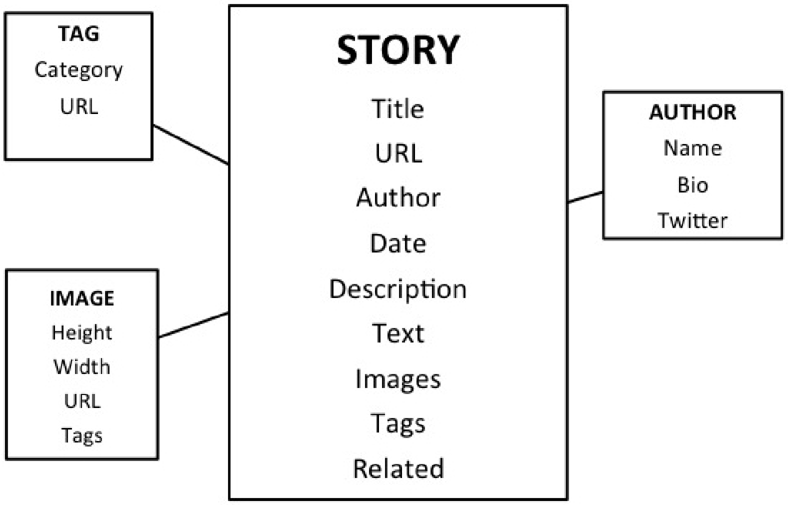


Figure 4-1: Simplified metadata scheme for a typical news story.

report announces that the Times produces "more than 300 URLs a day," using URL as a sort of "thing" word, their default unit of work.[64] 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.[65]

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 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 Circa's atoms or Vox's cards aim to create a baseline repository to build upon rather than recreate from scratch every time.

The sustainability and commercial viability of such approaches is still unclear, but the excitement around them speaks to a fundamental rethinking of 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,

[64] *Innovation*, 27.

[65] 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.

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.

Adrian Holovaty, co-creator of the Django web framework and its now-retired "benevolent dictator for life," sees a deep future in this new, structured approach to journalism. Holovaty wrote in 2006 that "newspapers need to stop the story-centric worldview"; each story, he argues, contains a vast amount of structure that is being thrown away with every click of the "publish" button.[66] 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*, 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 photos taken at a historic site, or the happy hours in a neighborhood? You might already have that information buried in your stories.

One example comes from the *Boston Globe*'s March 2015 coverage of the Boston Marathon bombing trial. Data editor Laura Amico knows well that trials are far from

[66] 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/`.

linear stories; since they are told by lawyers, they unfold as a series of conflicting arguments. Trials present a proliferation of stories: the chronological narrative of the trial, the pieced-together narrative of the original event, and fragments of tangential narratives from witnesses and documents. Moreover, the Globe already knows about, and has written about, many of the key players in the bombing trial, through covering the bombing itself and its many witnesses, victims, and pieces of evidence. Amico and her team knew there was more than one way to tell this story, so they decided to focus on the facts: each witness, exhibit, and argument is entered into a spreadsheet, which can generate snippets of facts and entities—designed as cards—for later review.[67] These cards can embed and contain other cards, or be combined to form a story. Such a framework is built for easy reuse; it recognizes the interlinked nature of events, and the stories that summarize and filter events. The coverage also highlights that "structured journalism" does not need to be adopted wholesale; it can work for one-off stories, events, and experiments as well.

Still, despite the news story's rigid structure and the limitations of indexing, the article is not disappearing anytime soon; it remains the core and canonical unit of online work. Link-based platforms and services like RSS and social media feeds still rely on conventional, stable, and consistent resources at given URLs. 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 need to restructure the news story completely; only that we rethink how they work under the hood. Stories are not uniform resources, and they should not be uniformly tagged and categorized.

4.2.3 From tags to links

On the web, stories gain a new structure and unprecedented source of new insights through hyperlinking. Links—and their siblings, linked tags—allow for a new stan-

[67] Benjamin Mullin, "How The Boston Globe is covering the Boston Marathon bombing trial," Poynter, 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/.

dard of citation, reference, and context provision for news. When used internally and indexed as first-order metadata on a story, a link can 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" between websites, and the same principles can be adopted and expanded within news websites.[68]

Organizing the web by link and tag has often proven more effective than trying to fit its contents into an overarching taxonomy or ontology. Google's PageRank algorithm was the lifeblood that made it the dominant search engine over rivals like Yahoo! and HotBot.[69] When Yahoo! began in 1994 as a hierarchical directory of useful websites, it seemed like a natural step. Computer users were accustomed to the tree-like document and file structure of computer systems, and the web replicated this in turn. Taxonomy allowed Yahoo! to build relationships between categories into its structure—parents, children, and siblings—which readily enabled features like "related categories" and "more like this." But Google succeeded by crawling in the weeds rather than commanding from on high. For Google, the links sort everything out. Berners-Lee proved that networks could work with many links—and in fact, if you had a lot of links, as Clay Shirky puts it, "you don't need the hierarchy anymore. There is no shelf. There is no file system. The links alone are enough."[70]

This is not a completely new phenomenon; long before Google's PageRank, scholars long relied on the footnote and bibliography to systematically track influence and dialogue, and networks of citations can be created through organizing and ranking

[68] Chakrabarti, *Mining the Web*, 12. See also section 5.1.

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

[70] Ibid.

by reference. 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, weights, and details which online newspapers can take advantage of, both in measuring impact on the web and inside their own archives.

The tag emerges somewhere in between the category and the link, as a hybrid hierarchical/networked organizational structure. On one hand, tagging relies on an individual singularly classifying 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.[71] These systems, sometimes called "folksonomies," emerge as manifestations of the "boundary infrastructures" proposed by Bowker and Star.[72]

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. Even controlled vocabularies don't always work: tags like "George Bush" are more challenging for a machine than "Barack Obama." Some research has also shown that the first tags on an item are likely to influence future tags in turn, resulting in a sort of ontological groupthink.[73] 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. This suggests that tagging is a crutch, a tacit recognition that all associations must be mediated through language. Taxonomists sometimes employ "synonym rings" to link

[71] 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.

[72] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT Press, August 28, 2000), See also section 2.1.

[73] Cattuto, Loreto, and Pietronero, "Semiotic dynamics and collaborative tagging."

related terms, but Shirky considers the possibility that syonymous terms aren't always 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."[74] This radical suggestion uses minor semantic differences as signals, but sometimes hierarchies are crucial; a user searching for movies set in Massachusetts would also want movies tagged "Boston."

The *Times* sees tagging as core to its business, and the reason it has remained the "paper of record" for so long.[75] This title was bestowed to them largely on the back of the legacy New York 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 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; for instance, it took seven years for the *Times* to start tagging stories "September 11." Other times, topic relevancy changes; early stories about solar power might have been tagged "science," but now the paper could have a dedicated "environment" or "clean energy" section or tag. Along with keeping up with topical drift, the *Times'* 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."[76]

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

[74] Shirky, "Ontology is Overrated."

[75] *Innovation*, 41.

[76] Ibid., 41-42.

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. Most stories are tagged manually, by an impatient reporter or editor on a tight deadline and according to a certain context. By combining modern natural language processing tools, a story can instead 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.

While tagging practices can be enhanced in a variety of ways, Stijn Debrouwere thinks that even "tags don't cut it."[77] 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 and ideas that carry value in their own right. For Debrouwere "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, and they give little indication of the meaning behind the relationship being formed. Debrouwere advocates for indexing these relationships: "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."

Debrouwere also suggests using *linked entities* instead of tags and labels. Natu-

[77] Stijn Debrouwere, "Information architecture for news websites," stdout.be, April 5, 2010, "Tags don't cut it", accessed April 20, 2015, `http://stdout.be/2010/04/06/information-architecture-for-news-websites/`.

ral language processing tools can find the entities—the people, places, organizations, events, and themes—in a given text with reasonable accuracy, and then link these entities in turn to knowledge bases like Wikipedia, IMDB, or any number of digital libraries, museums, and research institutions. This allows machines to infer the broader topic and detailed relationships from a given text. The result is 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, linked to other unambiguous entities (like Obama, and the POTUS) in turn.[78] While this research has a long way to go, it is a promising start towards extending the notion of events as relationships between entities, and stories as relationships between their fragments.

Every story—whether a blog post, a map, a listicle, or an interview—has its own structures and patterns, each of which can be indexed and mined for future archival value. One of the most obvious and underexplored of such 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 linked tags and 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."[79] It is better, Debrouwere imagines, 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."

[78] Joel Nothman, "Grounding event references in news" (PhD diss., University of Sydney, 2013), accessed April 20, 2015, `http://hdl.handle.net/2123/10609`.

[79] Stijn Debrouwere, "Taxonomies don't matter anymore," stdout.be, December 20, 2011, accessed March 10, 2015, `http://stdout.be/2011/12/19/taxonomies-dont-matter-anymore/`.

Another under-mined source of insight lies in the plethora of "human-generated lists" (as Google calls them in one patent) around the web.[80] 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 type 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 leverage readers' collective wisdom by asking users to specify the context behind their lists as well.

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 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. These play to a journalist's strengths, and augment rather than replace the human touch that comes with inline links and curated collections.

These web-native and polyhierarchical approaches to classification reflect the growing need for newsrooms to find weblike ways to organize their stories. Shirky is a champion of the tag, but he recognizes that organizing by taxonomy and ontology is sometimes preferable; namely, with a small corpus and expert catalogers.[81] This could have described a pre-web newsroom, but no longer: the publisher's corpus has

[80] Discovering and scoring relationships extracted from human generated lists (US8108417 B2, filed January 31, 2012), accessed March 9, 2015, `http://www.google.com/patents/US8108417`.

[81] Shirky, "Ontology is Overrated."

expanded and linked beyond measure, and its ranks of expert catalogers are rapidly dwindling. Meanwhile, readers expect an increasingly rich, contextualized, and personalized approach to information. This suggests a need to adopt new schemes, ones that leverage automatic and dynamic tagging, linked entites, image recognition, and the knowledge of experts and crowds.

4.3 Conclusion

The linked archive, when considered on a massive scale, 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. Likewise, no automated system is perfect, and a computer cannot fully follow many of the nuances and changes that languages and topics undergo. Linking the archive demands a symbiosis of content, business, and technology; it requires considering journalist-as-archivist and technology-as-prosthesis. Linking the archive requires making more structured and explicit, and subsequently mining the web of references that already reside in the stories; it requires a rethinking of the practices of journalists and editors, as well as an adoption of new tools and techniques.

The linked archive borrows from, but is distinct from the notion of "link journalism" or "networked journalism." As a term popularized by Jeff Jarvis to refer to the growing citizen journalism movement, link journalism has also led to Jarvis's succinct motto of "Cover what you do best, link to the rest."[82] Building on this idea, Charlie Beckett posits that linking between sources leads to editorial diversity, connectivity and interactivity, and relevance.[83] A linked archive turns the conversation inward—as Mark Deuze and others note, inlinks are vastly different from those that point out—but they can adhere to the same principles of diversity, connectivity, and relevance.

[82] Jeff Jarvis, "New rule: Cover what you do best. Link to the rest," BuzzMachine, February 22, 2007, accessed April 20, 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/.

[83] 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/.

While inlinking may seem nepotistic and selfish, this is not the case if the archive itself links out in turn. 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?

It is unhelpful to have a massive, borderless archive, but linked archives can expand their borders strategically through clever use of APIs. Publishing websites rarely operate alone; they rely on a plethora of third-party platforms and services for analytics, sharing, commenting, and recommendations. One could similarly integrate with APIs that offer archival resources from around the web, such as results from Google, Wikipedia, Flickr, or resources from digital libraries like Europaeana and the Digital Public Library of America—not to mention partnering with other publishers to merge archives or indices. If a user is searching a publisher's website instead of Google's, it is likely because she wants more context than a mere list or index of items. 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. 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.

Chapter 5

Tracing the Links

In each chapter I have increasingly honed in on the publishing archive and the conceptions of journalists and newsmakers of the archival value of their work. Here I will shift from prescriptive to descriptive, tracing the links themselves to determine what publishers are *actually doing* with their archives already. I will do this first by examining existing qualitative and quantitative research around the conception and use of hyperlinks by journalists and news outlets, then by closely examining the "inlinking" that journalists already do within their publications' stories. Hyperlinks are curated windows into archival material, and a sign of publishers' historical research and institutional memory at work. By determining which articles journalists are pointing to, we can start to understand how stories become canonized within a media institution, as well as how the links themselves can begin to form new categories of their own.

In the first section, I will review journalists' current considerations of the role of hyperlinking in their work, consisting of both qualitative interviews and quantitative link analysis. The second section examines the hyperlinking practices of a variety of publications with the aid of a custom open-source software tool. The tool aids in the exploration and examination of hyperlinks within news stories via two possible paradigms; by broader category or newsroom desk on the one hand, and by crawling the inlinks to form a network graph on the other. In the first instance, I consider the URL structure as a proxy for the hierarchical and institutional categories in a newsroom, in order to see if hyperlinking practices differ across categories as well as publications. In the second instance, I aim to ask whether the hyperlinks that journalists create can begin to form a new paradigm of categorization on their own; are journalists linking across traditional desks and categories, or are their hyperlinks reinforcing existing institutional structures?

5.1 Links and journalism

Hypertextuality is one of the primary new affordances of online journalism.[1] It seems to naturally lend itself to journalistic use; Juliette De Maeyer identifies several aspects of traditional journalistic practice that are amplified by the hyperlink, such as providing context and credibility, allowing "multiperspectival" journalism by linking to contradicting sources, initiating connection with other news outlets, and strengthening the process of gatekeeping.[2]

Journalists tend to agree that hyperlinking lends itself to core tenets of journalism. James C. Foust writes that "choosing links to include in your story gets to the very essence of what it means to be a journalist."[3] In 1995, Poynter's Nora Paul coined the term "annotative journalism" to describe judicious use of linking by journalists; she suggested that this might even become a whole new category of newsroom employee.[4] Hyperlinks also have a hand in allowing journalists to take a step back from a news event and add context and interpretation, rather than constantly providing facts they've provided before. Instead of rehashing yesterday's news, why not link to it?

While editors will explain that their role often involves changing the syntax or

[1] See 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`; Juliette De Maeyer, "Towards a hyperlinked society: A critical review of link studies," *New Media & Society* 15, no. 5 (August 1, 2013): 737–751, accessed December 12, 2013, `http://nms.sagepub.com.libproxy.mit.edu/content/15/5/737`; Tanja Oblak, "The Lack of Interactivity and Hypertextuality in Online Media," *Gazette* 67, no. 1 (February 1, 2005): 87–106, accessed December 12, 2013, `http://gaz.sagepub.com/content/67/1/87`.

[2] Juliette De Maeyer, "Methods for mapping hyperlink networks: Examining the environment of Belgian news websites" (Austin, TX, 2010), 9-10.

[3] James C. Foust, *Online journalism: principles and practices of news for the Web*, 2nd ed (Scottsdale, Ariz: Holcomb Hathaway, 2009), 161.

[4] Nora Paul, "Content: A Re-Visioning" (Interactive Newspapers '95, Dallas, TX, February 6, 1995), accessed February 8, 2015, `http://www.lehigh.edu/~jl0d/J366-99/366npaul.html`.

function of hyperlinks in a reporter's work, there is little about the journalistic practice of hyperlinking that has been standardized or codified. Some publications have written standards or internal memos for hyperlinking, but even these have changed in a short amount of time; for instance, major publications like *The Dallas Morning News* and even *National Public Radio* initially prohibited "deep linking" to their content (i.e. linking directly to a story, rather than a website homepage).[5] Similarly, one 2003 study also found latent signs of institutional outlinking standards; in a sampling of stories about Timothy McVeigh's execution, the *Times* only linked out to ".gov" and ".org" websites, avoiding ".com" entirely, while the *St. Petersburg Times*, only ever linked to ".org."[6] Some news organizations would even warn users when they were leaving the site, making clear that an outside link was not an endorsement. These practices seem quaint and obsolete in today's web, where blogs have inspired new linking practices and homepages are increasingly seen as an afterthought in favor of social media and sidebars; but this highlights the real fear in taking readers off site and towards the open web in the early 2000s, and many vestiges of this hesitance persist today.

Other publishers have explicitly codified linking policies and best practices. In a 2013 article, Mark Coddington summarizes a series of in-depth interviews that track the "normalization" of the hyperlink across different publishers. Coddington finds that institutional influences from professional journalism have blended with the political blogosphere to change the standards and practices of linking in recent years, as publishers and blogs alike "mutually adapt towards new norms."[7] He finds that reporters and editors "overwhelmingly expressed philosophies of openness" towards link sources, which is a marked change from publishers' original hesitation.[8] This

[5] Mark Tremayne, "Applying Network Theory to the Use of External Links on News Web Sites," in *Internet Newspapers: The Making of a Mainstream Medium*, ed. Xigen Li (Routledge, September 13, 2013), 49.

[6] Daniela V. Dimitrova et al., "Hyperlinking as Gatekeeping: online newspaper coverage of the execution of an American terrorist," *Journalism Studies* 4, no. 3 (2003): 410.

[7] Mark Coddington, "Normalizing the Hyperlink," *Digital Journalism* 2, no. 2 (April 3, 2014): 140.

[8] Mark Coddington, "Building Frames Link by Link: The Linking Practices of Blogs and News Sites," *International Journal of Communication* 6 (July 16, 2012): 2017.

can be considered as a product of blogging culture and the changing habits of web users, even down to individuals' web research skills and the design of browsers (for instance, many users are more comfortable navigating simultaneous browser tabs than in previous decades of the web). It would seem that in embracing links, traditional publishers are borrowing a page from the blogosphere, recognizing that linking is a two-way street, and a form of "paying it forward" for long-term gain.

However, Coddington also finds that this spirit is not always borne out in the actual links that publishers forge. Traditional publishers still tend to link internally, which he suggests is "almost anti-conflict, devoid of nearly all of the perspectival distinctiveness and boldness that characterizes the Web's discourse, but that might be perceived as a threat to the norm of journalistic objectivity."[9] Bloggers are more social with their linking, and frame their links in a way that he describes as more "episodic" (linking to recent events in an ongoing story) rather than "thematic" (linking to provide deeper background, context, or resources). Traditional publishers are also limited by institutional oversight, style guidelines, and inherited technological setbacks, such as content management systems that make it cumbersome to link.[10]

Coddington notes that organizations like *The New York Times* have linking style guides that include "who should add links and how," whereas bloggers' linking practices are free from oversight; however, these style guides and policies are not always strongly enforced, and some journalists discovered that their organizations had link policies they were not aware of.[11] Public media like the BBC and PBS use linking to further their own distinct institutional goals as well; the BBC requires at least one outlink per story, while PBS arranges "a careful balance of opinion through links" so as to maintain a sense of neutrality.[12]

[9] Coddington, "Building Frames Link by Link," 2021.

[10] Coddington, "Normalizing the Hyperlink," 148-9; Igor Vobič, "Practice of Hypertext: Insights from the online departments of two Slovenian newspapers," *Journalism Practice* 8, no. 4 (August 8, 2013): 363, accessed December 12, 2013, http://www.tandfonline.com/doi/abs/10.1080/17512786.2013.821325.

[11] Coddington, "Normalizing the Hyperlink," 149.

[12] Coddington, "Normalizing the Hyperlink," 151; BBC, *Putting Quality First - Inside the BBC*, 2010, 35-7, accessed February 7, 2015, http://www.bbc.co.uk/aboutthebbc/insidethebbc/howwework/reports/strategy_review.html.

These examples point to the varied goals for linking, and the divergent practices around it as a result. While digital news media is increasingly embracing the link as a powerful journalistic and storytelling tool, varying and slow-moving institutional forces keep some of these organizations from fully adopting them; one 2014 study of Swedish news found that, "The general impression is that a plateau of hyperlink use has been reached – a plateau rather lower than the potential."[13]

5.1.1 Qualitative or quantitative?

Research around hyperlink usage divides between quantitative and qualitative methods. Quantitative studies tend to be more frequent, as it is easier to mine the links in a publication (seeing what they're doing) than to gain access and interviews with the journalists who are doing the linking (seeing what they're thinking). But most scholars would argue for a healthy combination, and many quantitative analyses conclude with the suggestion that their findings would be well served by supplementing with interviews and newsroom observation.

For Brazilian researcher Suely Fragoso, the success of network analysis, webometrics, and Google's PageRank algorithm have "favoured macroscopic approaches focusing on the structure and topology of the Web," which "overshadow the fragile conceptualisation of individual hyperlinks they are based on."[14] The success that network analysis has found with hyperlinks makes it difficult to progress from description to interpretation. For Fragoso, the crucial step is to consider the web "in terms of cascading categories": considering the web as a medium allows Fragoso to see websites, webpages, and hyperlinks as successive media texts that encompass sub-types of media artifacts themselves.[15] This approach treats each layer as a distinct but overlapping media text, closely mapping to the layers of containment that I outline

[13] Michael Karlsson, Christer Clerwall, and Henrik Örnebring, "Hyperlinking practices in Swedish online news 2007-2013: The rise, fall, and stagnation of hyperlinking as a journalistic tool," *Information, Communication & Society* (2014): 13, accessed February 8, 2015, `http://dx.doi.org/10.1080/1369118X.2014.984743`.

[14] Suely Fragoso, "Understanding links: Web Science and hyperlink studies at macro, meso and micro-levels," *New Review of Hypermedia and Multimedia* 17, no. 2 (2011): 164.

[15] Ibid., 193.

in section 2.2.

Fragoso also finds that in-line links are more likely to function as traditional citations than other types of links. These links are also usually framed in a different way, as their anchor text flows as part of the diegesis of the text rather than existing outside of it as a URL or headline. Few studies have examined hyperlinking from a close-reading perspective—such a study is overdue, as a fruitful avenue for correlating the meaning or function of the hyperlink with its semantics and style. One 2009 study analyzed academic weblogs and finds patterns in the stylistic use of anchor text, some of which could be incorporated as signals into link analyses. A verb-oriented linked phrase like "X claims that" or "X responds" functions differently from a noun-oriented one, like linking to a "story," "editorial," or "letter."[16]

But generally speaking, links have too many possible meanings and too little inherent structure to automatically infer a motive. Luzon enumerates eight major types of links on blog pages even before diving into subcategories within them: links to comment pages, permalinks, trackbacks, archival categories, blog reactions, social shares, and so on.[17] The anchor text is a useful but limited signal for each of these possibilities, and the many uses of the hyperlink clearly vary across institutional boundaries, whether small blogs, major publications, or public and nonprofit outlets. It seems that quantitative approaches to link analysis carry great promise because they are relatively easy to do at scale, but they do not have enough information to infer the meaning of any single link with great accuracy.

Juliette De Maeyer's "overview of link studies" suggests that while studies of hyperlink networks employ a hodgepodge of methods, "a unified framework exists. It combines quantitative link counts, qualitative inquiries and valuation of field expertise to support link interpretation."[18] Such an approach would make logical sense, taking advantage of the scale of link crawling while always approaching this data with a critical eye. However, De Maeyer does note a conceptual divide between two

[16] María José Luzón, "Scholarly Hyperwriting: The Function of Links in Academic Weblogs," *Journal of the American Society for Information Science and Technology* 60, no. 1 (January 2009): 80-85.

[17] Ibid., 79.

[18] De Maeyer, "Towards a hyperlinked society," 737.

major approaches to link analysis. The first stems from the network science world, and aims to describe the structure and properties of the network being created, while the second originates from social science, and looks at the "information side-effect" of the link as an indicator of other phenomena.

The former, network theory-oriented approach is led by researchers like Albert-Lázló Barabási; it relies on complex mathematical methods and models to predict the size and shape of networks. Such an approach takes a high-level view of the web and examines its properties in the context of other networks in the world (such as transportation and power grids, or epidemics). This risks glossing over the idiosyncratic linking practices of a particular community, but it brings several helpful frameworks and ideas that journalists and publishers should keep in mind when considering the linking properties of the web. First, network theory finds the web to be a *scale-free* network, which follows a *power law* distribution in terms of connecting pages. In short, this means that 80% of the links on the web point to just 15% of its pages.[19] Second, network theory brings the idea of *preferential attachment*, which suggests that pages that already have links are more likely to acquire new links. The second leads to the first, in a sort of "rich get richer" phenomenon that exacerbates today's concerns about filter bubbles and cyberbalkanization. While network theory also addresses softer influences like trends and content quality, it shows that networks by their very nature will favor older and more popular websites, which have had more time to spend acquiring new links.

The latter, social-scientific approach treats hyperlink formation as a proxy for other influences, such as economic and political ties, international communication flow, or prestige and presence. As Han Woo Park and Chien-leng Hsu wrote in a 2003 study, "hyperlinks are not merely connectives between texts but mediators of a wide range of associative relations between producers of Web materials."[20] Much research has found a correlation between virtual hyperlink status and real-world net-

[19] Albert-Laszlo Barabasi, *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life* (Plume, April 29, 2003), "The 80/20 Rule".

[20] Chien-leng Hsu and Han Woo Park, "Sociology of Hyperlink Networks of Web 1.0, Web 2.0, and Twitter: A Case Study of South Korea," *Social Science Computer Review* 29, no. 3 (September 21, 2010): 357.

work influence, and such research is backed up by Google's PageRank algorithm, which similarly uses links as such a proxy.

PageRank is just one of hundreds of algorithms that Google simultaneously employs to determine credibility and influence on the web (others include the professionalism of the design and markup of the page, or the anchor text of the hyperlinks that link to the page). PageRank relies on a relatively simple underlying concept; a page's rank is a function of the sum of the rank of all the pages that link to it. But PageRank is only one way to measure it, and other citation-based search rankings offer alternative methods for graphing the web. One compelling example is Jon Kleinberg's HITS (Hyperlink-Induced Topic Search) algorithm. HITS is a sort of byproduct of the shape of the web as network theory conceives of it; generally speaking, webpages can be divided into "hubs" (whose purpose is to link out to other resources, like an index or topic page) and "authorities" (whose purpose is to be linked to as a reference for information, like a typical news article). The HITS algorithm computes a separate hub and authority score for each webpage; the latter determines the value of its content, while the former determines the value of its links.[21] Two pages that are linked by the same hub are said to have "cocitation," a common proxy for topical relevancy; the fewer the links from the hub, or the closer they are to each other on the page, the greater the confidence that the pages are related. In their treatment of the news ecosystem, Matthew Weber and Peter Monge add a third type of information actor to this flow; that of the "source," such as wire services like Reuters and the Associated Press that supply many authorities with the original information. This information is in turn directed from authorities to hubs, as publications like *The New York Times* feed to aggregators and indexes such as Google News and the Huffington Post.[22] Jeffrey Dean and Monika Henzinger likewise adopted a HITS-like method to create a hyperlink-oriented related page suggestion system. They turned to two algorithms: a

[21] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM* 46, no. 5 (September 1999): 604–632, accessed February 7, 2015, http://doi.acm.org/10.1145/324133.324140.

[22] Matthew S. Weber and Peter Monge, "The Flow of Digital News in a Network of Sources, Authorities, and Hubs," *Journal of Communication* 61, no. 6 (2011): 1062–1081, accessed February 8, 2015, http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1460-2466.2011.01596.x/abstract.

cocitation algorithm as outlined above, and a "companion" algorithm, which extends Kleinberg's HITS to account for the hyperlinks' order on a given page.[23] This example, along with the qualitative studies and anchor text analyses discussed earlier, shows that determining influence and relevancy on the web is more than a matter of how-many-links, or even who-links-to-whom.

Other research considers the ability to automatically add hyperlinks to news stories, whether pre-digital ones that are in the process of being digitized, or born-digital stories that could simply use some additional context.[24] Such systems can recognize entities in text and link them to other entities in the archive; for instance, a system might find the phrase "The Pittsburgh Steelers defeated the New England Patriots" and automatically hyperlink the text, pointing to the original article that announced the game result. Outlets like *The Washington Post* and *The New York Times* have used such automated hyperlink creation methods before, but Coddington finds that they are phasing them out in most cases.[25] This is not to suggest that automated linking carries no promise, but that simplistic and fully automated methods do not seem as effective. Others have proposed "link apprentices" that work with journalists to find the relevant links, combining automated search and human curation.[26] This approach seems to follow the "prosthetics" model that Stijn Debrouwere advocates.[27]

Still other research has examined readers' responses to hyperlink navigation. Such studies tend to rely on tech-savvy students as test subjects, tempering the conclusions we can draw; but the authors point to an interesting discrepancy in the link's function. Readers tended to prefer links in sidebars of news articles rather than in the anchor text of the article itself; they also preferred seeing additional metadata, like a lede or

[23] Jeffrey Dean and Monika R Henzinger, "Finding related pages in the World Wide Web," *Computer Networks* 31, no. 11 (May 17, 1999): 1467–1479, accessed February 7, 2015, `http://www.sciencedirect.com/science/article/pii/S1389128699000225`.

[24] Ioannis Arapakis et al., "Automatically embedding newsworthy links to articles: From implementation to evaluation," *Journal of the Association for Information Science & Technology* 65, no. 1 (January 2014): 129–145.

[25] Coddington, "Normalizing the Hyperlink," 148.

[26] M. Bernstein et al., "An apprentice that discovers hypertext links," in *Hypertext: Concepts, Systems and Applications* (Cambridge University Press, 1990).

[27] See section 4.2.3.

image, along with it.[28] In this sense, sometimes the hyperlink exists to entice the user to click and dive into more information, but other times, the link's metadata is all we need to understand the reference. This points to a need to consider the content being embedded when a link is made; embedding a title, description, and image may sacrifice a curiosity-driven click, but it is likely to be more user-friendly at the outset.

5.1.2 Staying in or going out?

Most research tends to divide the "inlink" and the "outlink," or those links that stay within the same website versus pointing out to others around the web. Using network theory to map external hyperlinks, Mark Tremayne observes that "news stories on the Web are more heavily linked every year," but he also notes that the percentage of external links from publishers has steadily decreased.[29] This, he suggests, is a result of publishers building up an increasingly robust archive of their own, as they no longer need to link out to provide contextual information: "As news sites' digital archives grow. . . there are more opportunities for linking."[30] He calls the siloed publications that result "gated cybercommunities," predating Philip Napoli's consideration of online media as "walled gardens" that have undergone "massification."[31] Here Tremayne assumes that external linking is altruistic and fosters global communication, while internal linking is nepotistic, narcissistic, and stifling for information flow. Mark Deuze calls inlinks and outlinks "quite different types of hypertextuality," with one opening up new content, and the other leading to a "spiraling down" of content. He suggests that "if a site only refers to documents to be found on that site, it actually tells us

[28] See William P. Eveland and Sharon Dunwoody, "User Control and Structural Isomorphism or Disorientation and Cognitive Load? Learning From the Web Versus Print," *Communication Research* 28, no. 1 (February 1, 2001): 48–78; Mark Tremayne, "Manipulating interactivity with thematically hyperlinked news texts: a media learning experiment," *New Media & Society* 10, no. 5 (October 1, 2008): 703–727.

[29] Tremayne, "Applying Network Theory to the Use of External Links on News Web Sites," 49.

[30] Mark Tremayne, "The Web of Context: Applying Network Theory to the Use of Hyperlinks in Journalism on the Web," *Journalism & Mass Communication Quarterly* 81, no. 2 (2004): 241.

[31] This also points to terms like "cyberbalkanization" and the "splinternet," which each similarly suggest that the web is increasingly dividing and splintering due to rising business and governmental interests. Here my scope is limited to news publishing on the web, but these terms are increasingly used in the context of applications and proprietary platforms beyond the web.

that the 'worldwide' Web does not exist."[32]

Such a conclusion is understandable given the current state of publishing archives and the traditional metrics for success online; editors, of course, want to boost their own clicks, search results, and time-on-site, so it would seem that they may be linking for the wrong reasons. But linking in could also allow for *better* context that is curated and controlled by one's own institution and brand. The content has no risk of changing without notice, and it adds an additional wrinkle of insight to a publishing archive. Linking to the archive can be beneficial to readers, as long as the archive itself altruistically links out when needed.

Moreover, news outlets weren't linking much *at all* prior to building up their own archives. The 2003 study of McVeigh's execution found that hyperlinks function as an additional layer of gatekeeping for web news editors. Gatekeeping has long been a purpose of the news media, introduced by David Manning White in 1950 with a study of "Mr. Gates," whose role in the newsroom was to choose which of myriad possible news stories to report on and publish. Mr. Gates rejected about 90 percent of proposed stories, and while he aimed to channel his readers when making these decisions, instinct and personal bias surely played a substantial role. Gatekeeping studies focus on the selection and framing of what defines newsworthy content, which has immense consequences for what topics and communities become amplified or marginalized.

Given the volume of content online and the unlimited space to fill, the online form of gatekeeping could be considered something more like floodgate-letting. The McVeigh study argues that hyperlinks play a crucial role, since linking is, in a sense, opening the gate, whether it's to your source (such as an Associated Press story) or a competing or contrasting article. In their analysis of about 500 stories from 15 news outlets, the study found that newspapers did not link often in general; less than 1% of a story's links came from the text itself, with most emerging in the sidebar. In-text hyperlinks are markedly different from sidebar or "blogroll" links. These tend to be

[32] Mark Deuze, "Online journalism: Modelling the first generation of news media on the World Wide Web," *First Monday* 6, no. 10 (October 1, 2001), accessed February 20, 2015, `http://ojphi.org/ojs/index.php/fm/article/view/893`.

produced for different reasons and carry markedly different functions; today's sidebar links tend to be algorithmically generated and always linking in, while blogrolls link out and serve more of a social than a topical function.[33]

In a 2013 paper titled "Staying in or Going Out?," Anders Olof Larsson examines how a handful of Swedish newspapers use internal and external hyperlinks. He finds that the vast majority of external links are found in the text of the article itself, while most internal hyperlinking occurs in the navigational sidebars (think "Recommended for you" or "Most emailed"). He concludes that "we are mostly seeing what could be labeled an automated approach to employing hyperlinks."[34] The archival links tend to be algorithmically generated, and related on the level of the entire article; the external links, built into the fabric of the writing, are hand-curated and tied to a specific part of the story. This may be one reason external links are considered an improvement over internal ones; in general, more thought has gone into them.

So there is a clear distinction between inlinks and outlinks, as well as between in-text links and sidebar links, but it does not stop here; each type of story might maintain a slightly different link profile based on its topic and content. Tremayne posits that stories about international affairs might be more heavily linked than stories about less complex topics closer to home.[35] The Times' *Innovation* report suggested starting with arts and culture coverage in enabling archives, in part because they offer more contextual links to well-known actors, films, and artists.[36] These two approaches highlight the varying function of the hyperlink, too; an international affairs story might offer more links to recent news stories, while an arts and culture story is more likely to link to topic pages or persistent resources, such as IMDB or Wikipedia. Moreover, arts and culture stories tend to be longer pieces, rather than quick informational updates such as a sports recap. Longer stories lead to more hyperlinks, which facilitates network formation.

[33] De Maeyer, "Towards a hyperlinked society," 745.

[34] Anders Olof Larsson, "Staying In or Going Out?," *Journalism Practice* 7, no. 6 (December 2013): 738.

[35] Tremayne, "The Web of Context."

[36] *Innovation* (New York Times, March 24, 2014), accessed April 20, 2015, `https://www.scribd.com/doc/224332847/NYT-Innovation-Report-2014`.

5.2 Link analysis

With these potentials and limitations in mind, I began considering approaches for conducting a link analysis of my own. As a software developer advocating for reuse of work, I was especially interested in building a tool or framework that would facilitate nuanced quantitative and qualitative analyses of linking within and between news websites. Such a tool would allow others to conduct their own link studies, and account for many of the signals and pitfalls of existing studies by storing and indexing detailed data about the links.

I first aimed to determine how news publishers are using internal, archival hyperlinks within their own website; I was especially interested in how these might vary from the linking practices of bloggers and aggregators. Many hyperlink studies discard internal links, but I was interested specifically in these links as windows into the archive. How often are publishers linking to articles from weeks ago, or even years ago? How many of them make use of topic pages, and how heavily do they link to them? Where on the page do these links tend to occur, and what sort of anchor text is used on them? What sort of additional metadata or markup travels with the archival link?

I also hoped to infer whether these internal, archival links can lead to alternative modes of classification, relevancy, and discovery. Are large legacy publishers linking within their own categories, or do they also link across newsroom desks and institutional structures? Addressing this question would begin to suggest whether a link-oriented classification scheme could enhance or replace the traditional categories, topics, and beats that structure news websites.

I suspected that at this stage, the current models, institutional structures, and norms of hyperlink use would reinforce many of the existing categories and perhaps lead to some glaring omissions. Unless reporters and editors are thinking of hyperlinks as units of organization and classification themselves, it is unlikely that they will be satisfactory in providing a complete or nuanced contextual picture of an ongoing event. However, I likewise was curious about any promising glimpses in existing practices; some specific categories of articles, or specific publications, might already be creating useful networks inadvertently, simply by judicious use of internal links.

As such, I began with a larger data set to examine high-level inlinking practices, in order to hone in on smaller, promising potential cases for network-oriented classification. While I expected that most articles and publications would not have sufficient internal linking to generate networks, certain categories of news articles or certain types of publications might show more promise. This roughly corresponds to a balance of quantitative and qualitative research approaches, as well as hierarchical-versus-networked classificational schemes.

Because several studies have noted an increased use of internal hyperlinks over time, as well as a qualitative shift in the approaches and mentalities of publishers towards hyperlinking, I was also interested in gathering longitudinal data about these links, in order to determine whether the number and characteristics of internal linking have changed in recent years. This would allow me to determine whether a change was already happening.

5.2.1 Methodology

I aimed to gather three data sets, which traversed from larger to smaller data and from high-level to low-level analysis frameworks:

- A large sampling of major news websites and blogs, for an aggregate view of linking practices across the field.
- A comprehensive sample of certain publishers, for specific comparative analyses of those publishers' linking practices.
- A few specific, crawler-generated networks of links surrounding a particular article, topic, or event.

I obtained the first two data sets through Media Cloud, a project jointly run by Harvard's Berkman Center for Internet & Society and MIT's Center for Civic Media. Media Cloud offers a nearly-comprehensive and well-indexed repository of news articles from recent years, which can be downloaded and processed programmatically for research use; it also contains a series of tags and "media sets," which broadly define and classify different types of media. These have been used to map online controversies, ranging from the Trayvon Martin case to debates around net neutrality.[37]

Before downloading the data sets from Media Cloud, I considered the software architecture of the tool. Building on Media Cloud's API Client, I created a "link extractor," which indexes details about each inline link in each story across the collections. The link extractor stores the target URL, the anchor text, the paragraph number, and whether or not it is an inlink (if the source and target URL are the same domain, it is considered an inlink). In case a researcher is looking for custom, publisher-specific link metadata, it also stores all of the attributes of the link's HTML element. Lastly, the extractor indexes story-wide metadata, such as word and paragraph counts. This enables a variety of research questions based on the text, placement, target, and markup of the inline hyperlinks in a story, all with the story's overall length in mind. In this case I aim for the inlinks within a publisher's archive, but one could readily track and analyze links between websites as well.

The first, large sample set honed in on two of Media Cloud's media sets: publishers in the Top 25 Mainstream Media, and publishers in the Top 1000 Popular Blogs.[38] Comparing these two sets allows an understanding of the quantitative differences in linking practices between major publishers on the one hand, and smaller blogs and aggregators on the other (I will note, however, that mainstream media is not the same thing as legacy media). For each set, I selected a week's worth of stories from February 2015, and a single day of stories from March 2013.[39] This gave me a total of nearly 42,000 mass media stories, and 23,000 stories from blogs.

For the second set, I downloaded all articles published by *The New York Times* and the *Guardian* for a single month in 2013, 2014, and 2015 each; for the purposes of comparison to a digital aggregator, I also collected stories from the Huffington Post

[37] See Yochai Benkler et al., *Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate*, SSRN Scholarly Paper ID 2295953 (Rochester, NY: Social Science Research Network, July 19, 2013); Erhardt Graeff, Matt Stempeck, and Ethan Zuckerman, "The Battle for 'Trayvon Martin': Mapping a media controversy online and off-line," *First Monday* 19, no. 2 (January 28, 2014).

[38] See Appendix A for examples of sources found in these two Media Cloud media sets.

[39] See appendix for details on the data sets.

for a single month in 2014 and 2015. This resulted in approximately 76,000 stories.

The third data set, consisting of the crawler networks, does not rely on Media Cloud. Here I used a more directed approach and selected sample stories to start a crawl; using the link extraction software, the crawler follows each inlink in a given article, in order of distance from the source, and extracts those inlinks' inlinks in turn, until it has no more inlinks to crawl. This results in a set that is necessarily not comprehensive like Media Cloud's, but one that is able to traverse across Media Cloud's time-bound link network. I considered three sample stories here, one each from the *Times*, the *Guardian*, and the *Washington Post*.

The link extractor and crawler will be available as open source tools.

5.2.2 Results

Here I will discuss some early findings that I've extracted with the help of Media Cloud and my link analysis software. The results outlined in this section are preliminary, and not purporting to rigorous statistical significance; the primary goal is to begin a conversation, and to outline the potential of the software for future link analysis endeavors.

With the first data set I aimed to compare mainstream media and popular blogs, as well as gain a sense of the linking practices between mainstream publishers. In a sense, I was looking for a typical "link profile" for a publisher or a media organization. I found that blogs maintain an advantage, and mainstream media hasn't yet caught up to their linking practices; blogs averaged 3.3 links per story to mainstream media's 2.6. When controlling for word count, the disparity between blogs and mainstream media grows larger; blogs have an average of over 60% more inlinks per word. Interestingly, both blogs and mainstream media also had a nearly equal number of inlinks and outlinks; this seems to run counter to Tremayne's observation that mainstream publishers tend to link in where blogs link out. Perhaps this is a result of changing practices, but it could also be because of varying methodologies (Tremayne counted links in the sidebars, where I am only counting in-text links). It is also worth noting that mainstream publications are more likely to have topic pages; for instance, over 40%

of the Times' total inlinks point to one of their topic or entity pages, rather than an old article. These pages signify a pledge to continually organize and update the archival record on this topic.

I also noticed that the link extractor sometimes found false hits; some publishers place related articles in between paragraphs of the text. Whether or not these are algorithmically generated or manually curated, they artificially increase the link profile of some publishers. Other times, the extractor would find author biography pages or other signals; given the structure of some of these websites, such misses are unavoidable, and they point to the need to temper conclusions and supplement them with qualitative inquiry.

Initial comparisons between major news websites found a wide range of linking practices; the percent of stories with inlinks ranged from less than 1% (the Daily Mail) to 86% (the San Francisco Chronicle). It's clear from these results that no unified linking framework or norm exists amongst mainstream publishers. However, one can find patterns in subsets of these mainstream media outlets; for instance, most of the the titans of legacy publishing (the Times, the Guardian, Reuters) maintain an above-average amount, with 60-75% of stories containing inlinks.

The second data set honed in specifically on these major figures, by extracting a month's worth of data from 2013, 2014, and 2015 from *The New York Times* and the *Guardian.* I then examined the URL structures of each publication, finding a proxy for the section or newsroom desk within the url (such as `/world/` for global stories, or `dealbook.nytimes.com` for NYT DealBook stories). This allowed me to begin to determine whether certain news desks were linking more or differently, whether due to the content or the institutional culture. Honing in on these publishers gave a few additional advantages: not only was I able to check for changes in linking practice over time, but I was also able to verify the significance of the week-long sample set, at least in terms of these two publishers. Both of their 2015 results were quite close to the results from the first data set.

Comparing the Times to the Guardian, I found a stark difference: the Guardian deployed an average of 2.6 inlinks per story, far more than the Times' 1.4. The

Guardian also appeared to focus more on inlinks, with 57% of their stories staying inside the domain, as opposed to just 46% from the Times; this is an especially stark difference given that the 50/50 split between inlinks and outlinks is one of the few common patterns across publications.

However, looking at linking practices across time reveals a different story. The Times has nearly doubled its number of links—both in and out—since 2013. Meanwhile, the Guardian's link quantities have steadily *reduced* in each consecutive year. It seems as though the Times is catching up, and both are nearly at parity in terms of total links in 2015; but perhaps the Guardian's loss of links is part of a similar plan, such as honing in on the most relevant links or using fewer fully-automated links. All the same, Coddington found that the Times was also cutting back on its automated links. This, of course, gets into the *quality* rather than quantity of links, which I will leave for future study.

Finally, I broke each story into categories based on the category found in its URL. This allowed me to compare, first of all, whether linking practices changed radically across topics and desks; and second of all, whether certain topics were markedly different between the two publications. The Times' blogs stood out as the most often linked, but closer examination found that the increase was primarily due to outlinking. This would seem to reflect the institutional structure of the Times, as their blogs are intended to be somewhat free from the Times' usual approach, while adopting some of the spirit of online blogs. The Guardian's desks featured no similar pattern, but ranged from between 1 and 5 inlinks per story based on the desk.

While the Times and Guardian organize their content differently, I hoped to directly compare their practices across desks; as such, I manually merged the Times and Guardian data wherever overlaps between categories were apparent (for instance, each one had a `science` category and a `film` or `movies` category). In general, the Guardian featured consistently more links across desks, without great variation; however, some patterns emerged. Movies and theater were both relatively well-represented, suggesting that arts and culture content lends itself to linking. Each outlet's politics section likewise came out above average in inlinking. But some desks were noticeably be-

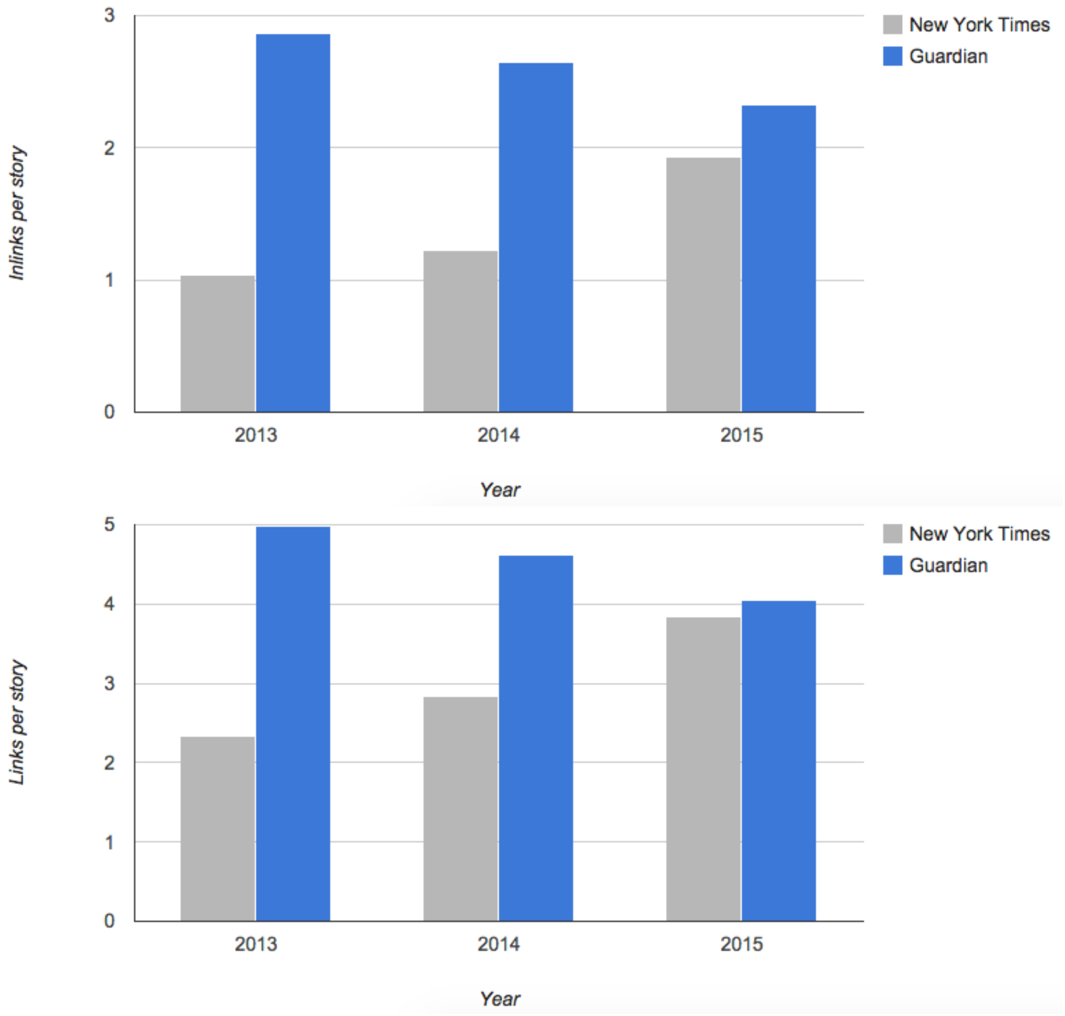


Figure 5-1: New York Times and Guardian inline link data over time.

low average for both, most strikingly their "world" sections, a topic that Tremayne suggested would require a great deal of context.

The third data set, gathered by crawling out from individual links to generate a network, was looking for linking deep into the past, or across traditional categories. I started with three incremental-update articles – one each from the *Times*, the *Guardian*, and the *Washington Post* – about a single event: a March 2015 plane crash in the Alps. For each, the crawler generated a network; this resulted in a few dozen articles for the Guardian and the Post, but the Times crawler found thousands of articles and would have continued to crawl indefinitely. Visualizing the resulting networks, I found that the Guardian maintained the most robust network,

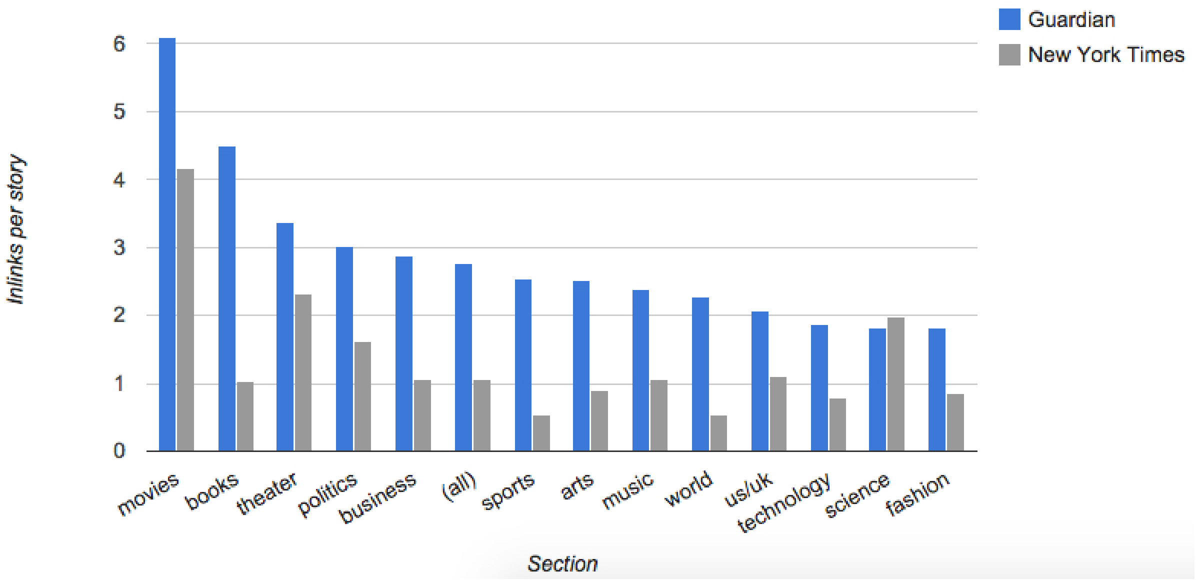


Figure 5-2: New York Times and Guardian inline link data by section.

more clustered than the Post's and featuring almost entirely relevant articles. For the Guardian, the crawler found a full 32 articles about the crash – this is close to all of the relevant articles, and they self-organized through links without human intervention. However, the Post's articles dip farther back into the past, referencing articles about related plane crashes from decades earlier. These older articles benefit most from a link-oriented classification scheme, as they would otherwise need to be retroactively tagged in order to turn up as relevant, unless part of a complex hierarchy.

Tellingly, one day after running the Times' crawler (on March 27), I tried it again and found that the crawler stopped after gathering just three articles, rather than exploding into infinity. Looking into the article, I discovered that the article had been updated, and its crucial link removed. This points to the fragility of such a classification scheme unless reporters and editors are writing with links in mind. If the editor simply did not want to present such a link to the user, the editor could simply hide the link element in the text, making it invisible except to machines.

These results are preliminary, and more sample data is needed to determine how far a network-oriented scheme can work. Topic pages—or their modern equivalent, like Vox's card stacks—function as "hubs" that point to the stories themselves, the "authorities." But authorities can be hubs on their own, measured by the quality of their links as much as their content; sometimes they link back to topic pages, and other times they link to one another. A link-oriented classification scheme might first need to establish the relationship between these new forms of hubs and authorities—and this relationship might be different for each publisher.

Chapter 6

Conclusion

Stories are increasingly becoming multifarious media texts, consisting of numerous interlinking fibers that each constitute different narratives of their own. Some of these texts are traditional—photographs, audio, video, maps, charts, and graphs—while others are something a bit new, emerging as a combination of these media. Journalism has always relied on data, but data can now flow with the narrative arc of a story in new ways. Interactive pieces, special features, and standalone applications often expand the rules of the hyperlink and bypass the URL, taking advantage of the new capabilities of HTML5 and modern JavaScript frameworks. As stories increasingly become networked and interactive, they explode into something more like apps.

An app can take many forms: it can live on the Apple or Google app store, whether for a phone or a tablet (known to developers as a "native" app); or an app can simply be a web product that is a bit *more* than a static webpage, or a single story. Its information might update on the fly, or it might feature advanced interactive elements inspired by documentary film and gaming; some of these take advantage of traditional hypertextuality, but others rely on hovering, swiping, watching, and listening rather than clicking and reading. There is a sort of continuum between traditional narrative stories and interactive applications, and most emerge as a sort of hybrid. The New York Times' Pulitzer-winning 2012 piece "Snow Fall" is the most canonical example of such a mixture of story and application.[1] These stories stretch the traditional link-

[1] John Branch, "Snow Fall: The Avalanche at Tunnel Creek," *The New York Times* (December 20,

and URL-based navigational scheme of the web to its limits. Native apps go even farther than interactive stories, sometimes offering views and glimpses of the web (for instance, think of clicking on a link in your Facebook or Twitter app and being led to a webpage), but excluding search or URL-based navigation from the experience.

Modern publishing platforms, such as Medium or Storify, can serve as factories for expanded news stories, which blur the divide between content and context and begin to veer into app territory. Such platforms allow writers to incorporate a variety of media, and take as a given that online stories combine original creation with remix and curation. Tools that facilitate annotation, such as Genius, further complicate the picture, tying texts to links, tags, and other commentary. Here stories are treated as systems from the outset, consisting of a plethora of media types. There is no doubt that modern web browsers and applications offer compelling new user experiences, and some parts of the web have essentially outgrown their need for traditional links. The link has limitations of its own, after all, which I have discussed at length in prior chapters. But in terms of archiving, storage, and face-value navigation, URLs and links are often all we have to go by; the URL, after all, serves as the default archival unit for the Internet Archive. While dynamically-generated web pages have been around since the early days of the web, recent years have seen an increasing reliance on browser-side scripts to generate webpages; this approach offers far greater flexibility in presentation and user experience, but it limits its historical potential, reducing the page's data and structure to the technologically contingent lines of code that generate it. Modern JavaScript not only bypasses the link's function, but often runs explicitly counter to it; the anchor tag is sometimes repurposed for customized use, and JavaScript's commonly-used `preventDefault` function actively stops the link from performing its usual task (that is, jumping you to another part of the web).

In this conclusion, I will consider the expanding future of the hyperlink. How is the traditional, default, cornerstone "jumping" action of the web—so crucial to the web's size, shape, and political economy—being bypassed or made obsolete in the next generation of stories? How are theorists, designers, and platforms alike

2012), accessed April 22, 2015, http://www.nytimes.com/projects/2012/snow-fall/.

rethinking the nature of the link, and therefore the roles of attribution and citation? Finally, how is the resulting change affecting the user's experience of the web, and its future archivability? I will begin with a typology of what I consider "news apps," and the challenges with preserving the process and results behind them. I will then look to certain designers' reconceptualizations of the hyperlink, and their potentials for alternative use; here I will also consider linking within mixed media such as audio and video, afforded by HTML5. Finally, I will conclude by considering an artistic treatment of digital archives, which offers a window into a user's experience of the link amidst this information deluge.

6.1 From links to apps

Some newsrooms are beginning to form dedicated internal teams for the production of interactive applications. Public, private, and nonprofit media alike are creating "news apps teams" that straddle the line between journalism and software development. A 2014 panel at the Online News Association, co-led by the Wall Street Journal and the Texas Tribune, closely considered how to build such a team in a newsroom, and similar outfits can be found at the Chicago Tribune, ProPublica, NPR and the Guardian.[2] Featuring journalists who are equally comfortable writing in English and JavaScript, or calling sources and wrangling datasets, these desks signal a new future where editorial and tech are fully integrated and managed by multitalented generalists.

Given the increasing proliferation of context, multimedia, and interactivity built into today's stories, this turn towards apps seems a logical next step. News apps aim to blend the bundle of story and information, by creating bespoke platforms and containers that serve to perfectly complement the right story with the right navigation and presentation. Sometimes these applications are even built outside of the content management system of the organization; while the CMS is a factory for stories, it cannot always contain apps. Apps come in a variety of shapes and sizes, and each

[2] Kelly McBride, "How to build a news apps team," Poynter, September 29, 2014, accessed April 22, 2015, `http://www.poynter.org/news/mediawire/271823/how-to-build-a-news-apps-team-hint-if-you-dont-have-a-lot-of-money-settle-for-scrappy/`.

brings its own challenges for standardization, organization, archivality, and reuse. This makes summarizing their future and associated challenges difficult; here I will aim to define a sort of typology of apps and platforms that challenge or expand the hyperlink's basic function, through the lens of the difficulties around archiving such works.

First, news apps come in the form of aggregation, personalization, and recommendation tools. These often take the form of native mobile apps, and have usually been the purview of information technology companies; some examples include Flipboard, Prismatic, and LinkedIn's Pulse. As publishers begin increasingly aggregating and linking out, though, there is a new potential niche for such products tied to newsrooms. BuzzFeed's news app features stories from around the web, as "part of a collaborative project with other news outlets."[3] While it is a publisher's product, the app is taking cues from aggregators like Yahoo! News Digest and Circa for design and recommendation decisions; similar efforts are coming from the New York Times, whose NYT Now app and front-page "Watching" section both reflect a desire to link out more often. Moreover, while these publishers have long focused on the importance and role of social media in their business, they are increasingly creating content directly for outlets like Facebook and Snapchat.[4] These social ranking and recommendation tools seem to promote open linking across different parts of the web, but just as often they may encourage continuous scrolling, browsing, and watching rather than clicking and reading.

A second type of news app to consider are standalone sites, usually data-driven and self-updating, that offer compelling statistics and narratives with design playing a crucial role. Often these state a specific role of adding context and explanation to a story. Sometimes these are not affiliated with any mainstream legacy publisher, such as Laura and Chris Amico's Homicidewatch.org, which combines stories and data to

[3] Caroline O'Donovan, "BuzzFeed is building a new mobile app just for news - theirs and everyone else's," Nieman Journalism Lab, July 31, 2014, accessed April 15, 2015, http://www.niemanlab.org/2014/07/buzzfeed-is-building-a-new-mobile-app-just-for-news-theirs-and-everyone-elses/.

[4] Ravi Somaiya Isaac Mike and Vindu Goel, "Facebook May Host News Sites' Content," *The New York Times* (March 23, 2015), accessed April 22, 2015, http://www.nytimes.com/2015/03/24/business/media/facebook-may-host-news-sites-content.html.

track every homicide in certain major urban areas. New York-based News Deeply aims to create platforms that enhance the user experience of news as well, combining statistics, maps, and timelines to tell a larger narrative instead of a single story; early experiments have tracked the Ebola pandemic and the Syrian civil war. Perhaps the most telling example is Timeline, a mobile app whose tagline succinctly summarizes the contemporary turn towards context: "The news is the short tail of a very long string of events. *Timeline* weaves those events into rich, compelling stories."[5] These apps and standalone sites are evolving repositories of code in their own right, and they call into question the dominant role of stories, and the content management systems that generate them.

The final type of news app that challenges the conventions of hyperlinking is the multimedia storytelling platform. Applications like Medium, Storify, and MIT Center for Civic Media's FOLD fall under this rubric. Like most content management systems, these applications still produce stories; but their notion of what a news story *is* is reflected in the types of media that can be collected and embedded within them. These platforms provide a system for easily incorporating tweets, videos, maps, and graphs; they are thus deeply reliant on hypertext, assuming that writers are going to link and embed as a core part of their work. While the platforms push the boundaries of the story in greater or lesser degrees, they still follow certain conventions and rules; in FOLD's case, for instance, the text proceeds vertically, while contextual multimedia elements are presented horizontally to offset it. This results in a balance between convention and innovation in storytelling; such systems offer the promise of allowing new forms of expression without breaking or challenging archival standards.

In each of these cases, and as news increasingly adopts elements of documentary film and gaming, the language and formats of storytelling are changing. This leads to particular challenges in storage and archiving. In one sense, these apps form a symbiotic relationship with archives; an application like Timeline aims to keep archives alive and dynamic in the context of today's stories, while data-oriented apps tend to be built with information sustainability in mind. All the same, their many forms

[5] "Timeline," Timeline, accessed April 22, 2015, `http://www.timeline.com`.

lead to difficulty in standardized storage. Scott Klein, luminary of news apps at the *New York Times*, brings up Adrian Holovaty's ChicagoCrime.org, which Holovaty described as "one of the original map mashups."[6] Launched in 2005, 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 is no doubt that some apps must be retired when their event passes or their function is completed. Software constantly races against obsolescence, and code is an ephemeral and evolving artifact. With some level of forethought, apps can be saved, stored, and archived; but what is important and worth saving about a given application? Is it the visual design, the data, or even the process of building it? In March 2014, a group of attendees of the NICAR conference gathered at Washington, D.C.'s Newseum to brainstorm the challenges and potentials of preserving news apps, suggesting more collaboration with libraries, museums, and cultural heritage institutions.[7] The Library of Congress has pledged to aid in digital preservation of newspapers, and some institutions are offering novel ways of preserving and maintaining digital work for the future. At the Cooper Hewitt Smithsonian Design Museum, a team led by Sebastian Chan has been preserving a defunct Silicon Valley app called Planetary as "a living object." For the Cooper Hewitt, preserving an app is more like running a zoo than a museum: "open sourcing the code is akin to a panda breeding program."[8] They'll preserve the original, but also shepherd the open-source continuation of app development, thereby protecting its offspring and suggesting new applications for old frameworks. While the Cooper Hewitt is currently guarding Silicon Valley technology, overlaps and partnerships between newspapers and cultural heritage institutions could lead to similar experiments.

These data-driven apps demand a great deal of work at the outset, which makes them high-investment projects for publishers, with high return expected as a result.

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

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

[8] 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/`.

This means that the most high-profile, engaging interactives tend to pair with "slow news" or predictable events with guaranteed audience, such as the World Cup or the Oscars. Unless such platforms become increasingly standardized, it will be difficult to produce data-driven and interactive stories about breaking and unpredictable events. What are the implications of this for archivality, storage, and history? How can we balance the creative adventurousness of new forms of storytelling with a particular set of frameworks and standards that will allow them to be legible to the public, and archivable for history?

News sites are, after all, increasingly linked to one another, and saving a single page out of the internet is itself an exercise in futility. The work of the members of the Amsterdam-based Digital Methods Initiative especially brings home this point; Anne Helmond's research into the Internet Archive examines the "boundaries of a website," exploring where content ends and context begins. She finds, echoing Niels Brügger's work into web archives, that pages (content) are privileged over sociotechnical context. Richard Rogers' *Digial Methods* laments the inability of current digital archiving methods to highlight "the references contained therein (hyperlinks), the systems that delivered them (engines), the ecology in which they may or may not thrive (the sphere) and the related pages, profiles and status updates on platforms."[9] Helmond then examines the phantom, broken links to dozens of advertising and analytics platforms on a typical archived New York Times page. The Open Knowledge Foundation approaches the same phenomenon from an online privacy angle; in their 2014 project "The News Reads Us," they find that German news outlets like *Die Welt* generate links to up to 60 other websites under the hood, with much of their data feeding straight to analytics platforms at Google and Facebook.[10] Here we see a small-scale example of the larger information flows suggested by the source-authority-hub

[9] Anne Helmond, "Exploring the Boundaries of a Website: Using the Internet Archive to Study Historical Web Ecologies" (MIT8, Cambridge, MA, 2013), accessed April 19, 2015, `http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-using-the-internet-archive-to-study-historical-web-ecologies/`; Richard Rogers, *Digital Methods* (The MIT Press, May 10, 2013).

[10] 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`.

model introduced by Weber and Monge.[11] Thus content is not the only information consistently flowing from publishers to platforms; user data is also aggregated in turn.

6.2 Standardizing stories

The online publishing platforms of today still tend to read linearly, but special features and apps are beginning to expand stories into two dimensions. The linear feed or stream is the simplest to understand, but limits storytelling capabilities and audience literacy as a result. Users tend to read digital stories in two dimensions: the X-axis is for browsing, while the Y-axis can be better used for deep dives.[12] A platform like FOLD reflects this standard, with linear cards supplemented by context along the Y-axis. FOLD therefore imposes certain creative limitations, giving a storyteller more linking potential than a single webpage, but less than an entire website. This not only allows for easier development and reuse, it lets audience literacy to develop over time, and creative exploration of the form. FOLD will have the ability to embed other FOLD cards, allowing for links between stories themselves. FOLD is a sort of factory for a certain type of linked story, a content management system on a higher plane. Best of all, it allows storytellers a different dimension of expression with little to no technical effort; while "Snow Fall" was rightly lauded, Quartz's Kevin Delaney quipped "I'd rather have a Snow Fall builder than a Snow Fall."[13] Rather than continuously building apps and special features, some publishers are striving to turn their CMS into a factory for rich interactives as well as simple stories.

Other standards are also changing the web and its narratives in enticing new ways. HTML5 was completed in October 2014, the first official update to the web in decades.

[11] Matthew S. Weber and Peter Monge, "The Flow of Digital News in a Network of Sources, Authorities, and Hubs," *Journal of Communication* 61, no. 6 (2011): See also section 4.1.2. Accessed February 8, 2015, `http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1460-2466.2011.01596.x/abstract`.

[12] Peter Samis and Tim Svenonius, "The X, Y, and Z of digital storytelling: Dramaturgy, directionality, and design," in *Museums and the Web* (Chicago, February 10, 2015), accessed April 26, 2015, `http://mw2015.museumsandtheweb.com/paper/the-x-y-z-of-digital-storytelling-dramaturgy-directionality-and-design/`.

[13] *Innovation* (New York Times, March 24, 2014), 36, accessed April 20, 2015, `https://www.scribd.com/doc/224332847/NYT-Innovation-Report-2014`.

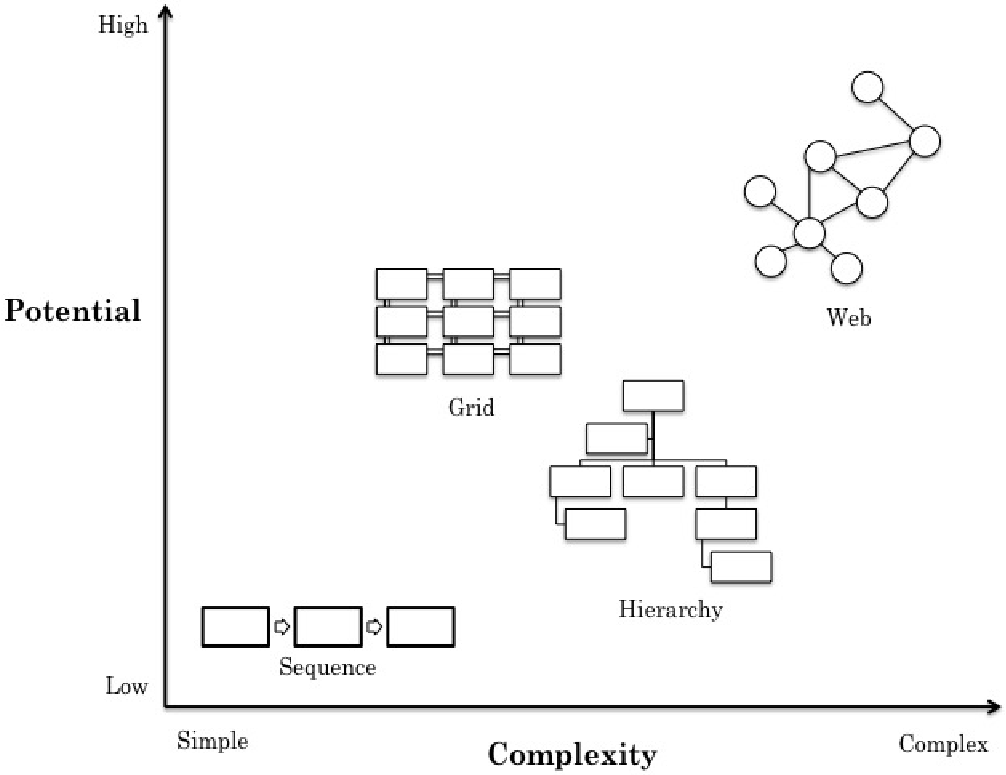


Figure 6-1: Diagram of standard link structures. Adapted from Blustein 1999.

It brings with it powerful features across a range of devices. Some of these aim to enrich the semantic meaning of content, with new standardized tags like `<section>` and `<article>`; others bring hypertext to other media, such as `<audio>`, `<video>`, and `<canvas>`. All of these tags allow for richer forms of automatic navigation and data collection, giving multimedia some of the jumping and contextualizing powers of the hyperlink. Some also offer new forms of linking; with HTML5, sound and image elements can themselves be made interactive, and a lot of them adopt the mechanics of the link and the semantic backup of hypertext.

Consider, for instance, the `<audio>` tag, which by default places an audio player on a page. Previously, audio required a plugin (such as Adobe's Flash), and did not adhere to any standards; now, audio files can contain associated text, variable play modes, and detailed time and playback information. This not only improves the user experience and enables new modes of presentation for developers, but it also lets audio be more searchable and discoverable online. For example, a radio interview about climate change is no longer a meaningless block of audio to a web browser; if a user were only interested in the sections about solar power, or wanted to hear all of the questions but not the answers, HTML5 might enable the user to "jump" to different sections of the audio. San Francisco-based startup Pop Up Archive automatically transcribes spoken word audio, aiming to provide precisely the sort of discoverability and research assistance through audio that both expands on and bypasses hypertext.

As we have seen, the story does not end with the story itself, and the myriad contextual elements—from auxiliary photos, to charts and graphs, to ads and popups—are as much part of the experience of online journalism as the text itself. Yet another layer is added to the archive through the social metadata that travels with a given story, as well, both on and off the site; when a reader registers a comment, click, or Facebook like on a story, this too is archived along with it. Recent years have seen a focus on digital annotations, or comments that are tied to specific sub-pieces of text rather than comments that appear at the bottom of a page. The goal of multimillion dollar company Genius is to "annotate the world," while more scholarly startup Hypothes.is asks you to "annotate with anyone, anywhere" (and a promotional video on their homepage prominently features Vannevar Bush's memex). Publishers like Atlantic Media's Quartz and platforms like Medium are also incorporating side-by-side annotations, rather than end-of-page comments. This approach privileges commentary, placing it on a similar level to the original content.

The W3C—the web's standards organization—is developing a series of standards for annotation, which will allow for multimedia, as well as annotations *of* annotations. This again points to the recursive nature of knowledge, and a need for similar structures when archiving and storing it. It also emphasizes the inherent challenges in providing concrete commentary on ephemeral content; many of its technical challenges—such as what to do with an annotation if the source text is deleted—allude to deeper and more philosophical challenges innate to the web itself. The effort to standardize annotation also reflects the primacy afforded to online commentary—more often found on social media platforms than end-of-page comment sections—and a need to consider these socially contextual elements as a potential part of the archived story as well.

6.2.1 Future design

Some designers and institutions are also reconsidering the look and feel of the hyperlink. These approaches assume that this basic jumping mechanic is not going away anytime soon, so the trick is to give it better interactive capabilities and differential nuance. Australian researcher and founder of NewsCubed, Skye Doherty, has especially considered these limitations of the link, also noticing parallel limitations in the treatment of hyperlinks by researchers. While online link analysis considers the location and direction of links, and qualitative reviews address the use of links by information gatherers and providers, there is little research that combines these broad treatments with user experience and narrative structure.

Indeed, hypertext found perhaps its first ally in fiction and art; scholars like George Landow, Jay David Bolter, and Lev Manovich have traced the use of hypertext as a narrative device that champions a proliferation of narrative and collapses linear forms of storytelling. Early experimenters in hypertext were often not information technologists, but writers and artists looking for new forms of expression that integrated image, text, and sound.

These digital literary scholars have focused on hypertext's artistic and narrative potential, and often highlighted pre-echoes of nonlinear writing that hypertext brings to the fore, especially in modernist literature. The "inter-cutting" techniques that eschew chronological and single-voiced narrative found in works like James Joyce's *Ulysses* and Virginia Woolf's *The Waves* point to the limitations of language and an inspiration from narrative techniques that were new at the time, borrowing from film, radio, and the clipped, decontextualized print of newspapers. In these cases too, writers and thinkers like Joyce and Walter Benjamin saw the creative value of obsessive collection and curation. Justin Hall, one of the web's first popular "weblog" writers, wrote in 1993 that his blog stemmed from "a deep geek archivist's urge to experiment with documenting and archiving personal media and experience. In college I realized that Proust and Joyce would have loved the web, and they likely would have tried a similar experiment—they wrote in hypertext, about human lives."[14]

Since hypertext's early pioneers were writers and artists experimenting with a new form, it is surprising to see this body of work and research ignored in favor of quantitative link analyses, whether beginning from network science or social science. This is not to say that such analyses are not useful, but that they are only part of the picture; Doherty asserts that hypertext "is under-researched in journalism practice, particularly as a narrative device."[15] As journalism continues to balance warily on the fence between "stories" and "information," analyses of journalism and hyperlinks could combine these divergent research paths, both understanding hypertext's effect on narrative structure on a small scale, and its foundational role in contextual network formation at a larger scale. In order to bring in new design and storytelling potentials to hypertext scholarship, journalism can serve as a crucial bridge between big data and small story.

Much of the utopian rhetoric around the web has fallen away in recent years, perhaps replaced by similar discussion around big data. Early-web writings on hypertext by Manovich, Bolter, Landow, and others sometimes treated the hyperlink as a tool of liberating power and narrative freedom, but this rhetoric has since been tempered by its role as a device for tracking and directing user behavior. Doherty's aim is to bring back the early web spirit of the hyperlink's potential, and expand it into the contemporary web. In considering new models and interactions for the hyperlink, Doherty emphasizes its spatiality, which I address historically in section 3.2.2. He notes that few modern tools offer graphical maps of a document's or website's structure, despite that some such views used to play "an important role in authoring (and also navigating) hypertext."[16] Journalism literature does not consider how "hypertext exists in space."[17] Spatial, dimensional expressions of hypertext can help users and

[14] Dan Gillmor, *We the Media: Grassroots Journalism By the People, For the People* (O'Reilly, 2006).

[15] Skye Doherty, "Hypertext and Journalism," *Digital Journalism* 2, no. 2 (April 3, 2014): 124, accessed February 8, 2015, `http://dx.doi.org/10.1080/21670811.2013.821323`.

[16] Skye Doherty, "Hypertext and news stories" (November 1, 2013): 1–10, accessed February 8, 2015, `http://espace.library.uq.edu.au/view/UQ:313369`.

[17] Ibid.

writers alike see the overarching structure of a story. Stories can take on a variety of graphical structures, some of which can be standardized, while others expand or break a typical model of story complexity. The applications, features, and factories for new stories are changing the traditional notion of the hyperlink as a blue, underlined string of text, and playing to the strengths of the web as never before; Larrondo Ureta suggests that online special features exhibit "one of the maximum expressions of hypertextuality."[18] The challenge lies in maintaining this creative potential while still standardizing them enough to visualize, coherently read, and archive them.

Doherty concludes by proposing three models of ongoing stories in journalism, which offer two competing narrative approaches. The "martini glass" emphasizes author-driven narratives, while the "drill-down story" prioritizes a reader-driven approach. A third type, the interactive slideshow, promotes a dialogue between the two approaches, balancing author control and user navigation. While this typology is surely too-simple, it points to a desire to begin to standardize and develop a language around interactive features, which break current boundaries and therefore current methods of storage, archivality, and reuse.

6.3 The art of the archive

The artist Lara Baladi does not want people to forget about Tahrir Square and #Jan25. Although it was "the most digitally documented and disseminated event in modern history," the 2011 Egyptian revolution is slipping into the past, and Baladi has gathered and stored thousands of media artifacts surrounding it.[19] Her resulting project, *Vox Populi: Archiving a Revolution in the Digital Age*, combines the latest news and ancient history, featuring documents ranging from a printout of the Declaration of the Rights of Man to a dynamic visualization of Twitter activity at the moment that president Hosni Mubarak resigned. In spring 2014 I attended an early

[18] Doherty, "Hypertext and Journalism"; Ainara Larrondo Ureta, "The Potential of Web-Only Feature Stories," *Journalism Studies* 12, no. 2 (April 1, 2011): 188–204.

[19] Lara Baladi, "Vox Populi: Archiving a Revolution in the Digital Age," Open Documentary Lab at MIT, accessed April 26, 2015, `http://opendoclab.mit.edu/lara-baladi-vox-populi-archiving-a-revolution-in-the-digital-age`.

study for Baladi's project at Harvard's Graduate School of Design. Baladi had taken over a room and assembled an overwhelming array of media artifacts, surrounding the participants with such a variety of formats that sensemaking seemed impossible. But soon patterns started to emerge. The event featured coffee and popcorn, and photos of popcorn machines at Tahrir Square gave a sense that we were experiencing the smells of the protests. Hard-copy printouts of tweets mixed with ancient artifacts projected onto TV screens, melding the ancient and modern, the digital and physical. It seemed to reflect the simultaneous revolutionary activity on the streets and online. These many representations combined with my own memory of reading about the protests from afar, and many of the artifacts on display intentionally reminded me that this was only a sliver of the real thing.

Baladi invited us to sit in a circle, drinking coffee and discussing how the exhibit around us made us rethink the flow of history and its digital representations. As we talked, we were being constantly documented ourselves; photographers and videographers surrounded us snapping pictures and videos. Our discussion of the archive was being folded back into it. The urgent news of the eighteen-day revolution is sliding towards history, but Baladi's act of archiving it is far from an exercise in cold storage. By talking about the archive, we became part of it, reanimating the past and the people and ideas behind the Egyption revolution. *Vox Populi* conjures up the exuberance and the fear at Tahrir Square in 2011, the "unstuck-in-time" qualities of social media, the age-old drive to collect and remember, and the bitter irony of struggling to hold onto memory in the age of information. As all of these traces remind us, we had to be there—but at least we have all of these traces.

Appendix A

Media Cloud and data sets

Media Cloud is a joint project between the Berkman Center for Internet & Society at Harvard, and the Center for Civic Media at MIT. An open source, open data project, it provides nearly comprehensive access to a wide variety of news media, allowing researchers "to answer complex quantitative and qualitative questions about the content of online media." Along with the extensive repository of stories, Media Cloud researchers have organized and classified many of its sources into groupings of "media sets." For this project, I focused on two of Media Cloud's media sets: the Top 25 Mainstream Media, and Top 1000 Popular Blogs.

Top 25 Mainstream Media

- New York Times
- San Francisco Chronicle
- CNET
- New York Post
- Boston Herald
- CBS News
- FOX News
- Los Angeles Times
- Time

- NBC News
- New York Daily News
- Reuters
- The Guardian
- Washington Post
- The Huffington Post
- CNN
- USA Today
- The Telegraph
- BBC
- Daily Mail
- The Examiner
- Forbes

Sample of top 1000 Popular Blogs

- Yahoo! News
- U.S. News
- ESPN.com
- BuzzFeed
- Reddit
- Hacker News
- Mashable
- Hype Machine
- Christian Science Monitor
- MTV News
- E! Online
- People.com
- TechCrunch
- Gizmodo
- ScienceDaily

- Laughing Squid
- MetaFilter
- Boing Boing
- Techmeme
- Stereogum
- Engadget
- Economist.com
- GigaOM
- Gothamist
- PCWorld
- Lifehacker
- Daily Kos

I selected the sample data sets based on a combination of variation and technical feasibility. Some Media Cloud downloads were made difficult due to corruptions or issues with the files, sometimes making obtaining data for consecutive dates impossible.

For the first data set, I gathered link data from the Top 25 Mainstream Media in February 1-7, 2015, and March 1, 2013; I also collected from the Top 1000 Popular Blogs on February 1, 2, 6, 7, 11, and 12 of 2015 (this simulates a full week of stories, as it staggers the sample by the day of the week), and April 1, 2012.

For the second data set, I collected all New York Times and Guardian stories from January 2015, February 2014, and March 2013. I also downloaded all Huffington Post stories from January 2015 and February 2014. This allowed for a nearly-annual set while avoiding any potential macro-level link pattern changes based on events in a given month.

The custom link extraction software is called "sausage" and will be released under an open source license. Written in Python for a MongoDB database, it is an extension of MIT Center for Civic Media's Media Cloud API Client and incorporates several third-party libraries and tools, such as Newspaper, BeautifulSoup, and NetworkX. It is available as of May 8, 2015 at `http://github.com/mailbackwards/sausage`.

Bibliography

"A Limited Company of Useful Knowledge: Paul Otlet, the International Institute of Bibliography and the Limits of Documentalism." Everything2. May 18, 2001. Accessed September 23, 2014. `http://everything2.com/index.pl?node_id=1053046`.

Abel, Richard. "The Pleasures and Perils of Big Data in Digitized Newspapers." *Film History* 25, no. 1 (January 2013): 1–10.

Abelson, Brian, Stijn Debrouwere, and Michael Keller. "Hyper-compensation: Ted Nelson and the impact of journalism." Tow Center for Digital Journalism. August 6, 2014. Accessed April 19, 2015. `http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/`.

Abreu, Amelia. "The Collection and the Cloud." The New Inquiry. March 9, 2015. Accessed April 20, 2015. `http://thenewinquiry.com/essays/the-collection-and-the-cloud/`.

Achenbach, Joel. "Journalism is Aggregation." *Washington Post* (April 9, 2014). Accessed April 16, 2014. `http://www.washingtonpost.com/blogs/achenblog/wp/2014/04/09/journalism-is-aggregation/`.

Anderson, C. W. "What aggregators do: Towards a networked concept of journalistic expertise in the digital age." *Journalism* 14, no. 8 (November 1, 2013): 1008–1023.

Arapakis, Ioannis, Mounia Lalmas, Hakan Ceylan, and Pinar Donmez. "Automatically embedding newsworthy links to articles: From implementation to evaluation." *Journal of the Association for Information Science & Technology* 65, no. 1 (January 2014): 129–145.

Arbesman, Samuel. *The Half-life of Facts*. New York, NY: Penguin, 2013.

Atwood, Jeff. "The Xanadu Dream." Coding Horror. October 12, 2009. Accessed April 20, 2015. `http://blog.codinghorror.com/the-xanadu-dream/`.

Baladi, Lara. "Vox Populi: Archiving a Revolution in the Digital Age." Open Documentary Lab at MIT. Accessed April 26, 2015. `http://opendoclab.mit.edu/lara-baladi-vox-populi-archiving-a-revolution-in-the-digital-age`.

Barabasi, Albert-Laszlo. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, April 29, 2003.

Barnet, Belinda. *Memory Machines: The Evolution of Hypertext*. London: Anthem Press, July 15, 2013.

———. “Pack-rat or Amnesiac? Memory, the archive and the birth of the Internet.” *Continuum: Journal of Media & Cultural Studies* 15, no. 2 (July 2001): 217–231.

———. “The Technical Evolution of Vannevar Bush’s Memex.” *Digital Humanities Quarterly* 2, no. 1 (2008).

Benton, Joshua. “That was quick: Four lines of code is all it takes for The New York Times’ paywall to come tumbling down.” Nieman Journalism Lab. March 21, 2011. Accessed April 20, 2015. `http://www.niemanlab.org/2011/03/that-was-quick-four-lines-of-code-is-all-it-takes-for-the-new-york-times-paywall-to-come-tumbling-down-2/`.

Berners-Lee, Tim. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. HarperBusiness, 2000.

Bernstein, M., N. Streitz, A. Rizk, and J. Andre. “An apprentice that discovers hypertext links.” In *Hypertext: Concepts, Systems and Applications*. Cambridge University Press, 1990.

Blair, Ann. “Note Taking as an Art of Transmission.” *Critical Inquiry* 31, no. 1 (September 1, 2004): 85–107.

———. “Reading Strategies for Coping With Information Overload ca.1550-1700.” *Journal of the History of Ideas* 64, no. 1 (2003): 11–28.

———. *Too Much to Know: Managing Scholarly Information before the Modern Age*. Yale University Press, November 2, 2010.

@blippoblappo and @crushingbort. “3 Reasons Benny Johnson Shouldn’t Call Out Plagiarism: He’s A Plagiarist, He’s A Plagiarist, and He’s A Plagiarist.” Our Bad Media. July 24, 2014. Accessed April 19, 2015. `https://ourbadmedia.wordpress.com/2014/07/24/benny-johnson-probably-shouldnt-call-people-out-for-plagiarism/`.

———. “More Plagiarism From "One Of The Web’s Deeply Original Writers".” Our Bad Media. July 25, 2014. Accessed April 19, 2015. `https://ourbadmedia.wordpress.com/2014/07/25/more-plagiarism-from-one-of-the-webs-deeply-original-writers/`.

Blustein, William James. *Hypertext Versions of Journal Articles: Computer-aided linking and realistic human-based evaluation*. 1999.

Borges, Jorge Luis. *Collected Fictions.* Translated by Andrew Hurley. New York: Penguin, 1999.

———. “The Analytical Language of John Wilkins.” In *Other Inquisitions 1937-1952*, translated by Ruth L.C. Simms. Austin, TX: University of Texas Press, 1993.

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

Branch, John. “Snow Fall: The Avalanche at Tunnel Creek.” *The New York Times* (December 20, 2012). Accessed April 22, 2015. `http://www.nytimes.com/projects/2012/snow-fall/`.

Brin, Sergey, and Lawrence Page. “The Anatomy of a Large-scale Hypertextual Web Search Engine.” In *Proceedings of the Seventh International Conference on World Wide Web 7*, 107–117. WWW7. Amsterdam: Elsevier Science Publishers B. V., 1998.

Brown, John Seely, and Paul Duguid. *The Social Life of Information.* Harvard Business Press, 2002.

Burke, Peter. *A Social History of Knowledge: From Gutenberg to Diderot.* Cambridge: Polity, December 2000.

Bush, Vannevar. “As We May Think.” *The Atlantic* (July 1945). `http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/`.

———. “Memex Revisited.” In *From Memex to Hypertext: Vannevar Bush and the Mind’s Machine*, edited by James M. Nyce and Paul Kahn, 197–216. San Diego, CA, USA: Academic Press Professional, Inc., 1991.

Carr, David. “IAmA columnist and reporter on media and culture for the New York Times.” Reddit. Accessed March 15, 2015. `http://www.reddit.com/r/IAmA/comments/16k598/iama_columnist_and_reporter_on_media_and_culture/c7wt5ko`.

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.

Center, Pew Research. *Journalism Interactive: Survey Pinpoints a Sea Change in Attitudes and Practices; Engagement Defines New Era of Journalism*, July 26, 2001. Accessed March 16, 2015. `http://civicjournalism.org/about/pr_interact.html`.

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

Chambers, Ephraim. *Cyclopædia, or an Universal Dictionary of Arts and Sciences.* 1728. http://uwdc.library.wisc.edu/collections/HistSciTech/Cyclopaedia.

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/.

Chen, Adrian. "Remix Everything: BuzzFeed and the Plagiarism Problem." Gawker. July 28, 2012. Accessed April 19, 2015. http://gawker.com/5922038/remix-everything-buzzfeed-and-the-plagiarism-problem.

Chung, Chung Joo, George A. Barnett, and Han Woo Park. "Inferring international dotcom Web communities by link and content analysis." *Quality & Quantity* 48, no. 2 (April 3, 2013): 1117–1133.

Coddington, Mark. "Building Frames Link by Link: The Linking Practices of Blogs and News Sites." *International Journal of Communication* 6 (July 16, 2012): 20.

———. "Defending judgment and context in 'original reporting': Journalists' construction of newswork in a networked age." *Journalism* 15, no. 6 (August 1, 2014): 678–695.

———. "Normalizing the Hyperlink." *Digital Journalism* 2, no. 2 (April 3, 2014): 140–155.

Coscarelli, Joe. "Ezra Klein on Vox's Launch, Media Condescension, and Competing With Wikipedia." *New York Magazine* (April 11, 2014). Accessed April 20, 2015. http://nymag.com/daily/intelligencer/2014/04/ezra-klein-interview-vox-launch.html.

Crawford, Kate, and danah boyd. "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society* 15, no. 5 (June 2012): 662–679.

Dahlgren, Peter. "Media Logic in Cyberspace: Repositioning Journalism and its Publics." *Javnost - The Public* 3, no. 3 (January 1, 1996): 59–72.

De Maeyer, Juliette. "Methods for mapping hyperlink networks: Examining the environment of Belgian news websites." Austin, TX, 2010.

———. "Towards a hyperlinked society: A critical review of link studies." *New Media & Society* 15, no. 5 (August 1, 2013): 737–751. Accessed December 12, 2013. http://nms.sagepub.com.libproxy.mit.edu/content/15/5/737.

Dean, Jeffrey, and Monika R Henzinger. “Finding related pages in the World Wide Web.” *Computer Networks* 31, no. 11 (May 17, 1999): 1467–1479. Accessed February 7, 2015. http://www.sciencedirect.com/science/article/pii/S1389128699000225.

Debrouwere, Stijn. “Information architecture for news websites.” Stdout.be. April 5, 2010. Accessed April 20, 2015. http://stdout.be/2010/04/06/information-architecture-for-news-websites/.

———. “Taxonomies don’t matter anymore.” Stdout.be. December 20, 2011. Accessed March 10, 2015. http://stdout.be/2011/12/19/taxonomies-dont-matter-anymore/.

Derrida, Jacques. “Archive Fever: A Freudian Impression.” Translated by Eric Prenowitz. *Diacritics* 25, no. 2 (July 1, 1995): 9–63.

Deuze, Mark. “Online journalism: Modelling the first generation of news media on the World Wide Web.” *First Monday* 6, no. 10 (October 1, 2001). Accessed February 20, 2015. http://ojphi.org/ojs/index.php/fm/article/view/893.

———. “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.

Dimitrova, Daniela V., Colleen Connolly-Ahern, Andrew Paul Williams, Lynda Lee Kaid, and Amanda Reid. “Hyperlinking as Gatekeeping: online newspaper coverage of the execution of an American terrorist.” *Journalism Studies* 4, no. 3 (2003): 401–414.

Discovering and scoring relationships extracted from human generated lists US8108417 B2, filed January 31, 2012. Accessed March 9, 2015. http://www.google.com/patents/US8108417.

Doherty, Skye. “Hypertext and Journalism.” *Digital Journalism* 2, no. 2 (April 3, 2014): 124–139. Accessed February 8, 2015. http://dx.doi.org/10.1080/21670811.2013.821323.

———. “Hypertext and news stories” (November 1, 2013): 1–10. Accessed February 8, 2015. http://espace.library.uq.edu.au/view/UQ:313369.

Drucker, Johanna. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly* 5, no. 1 (2011). http://www.digitalhumanities.org/dhq/vol/5/%201/000091/000091.html.

Ellis, Justin. “After the archive came down: The New Yorker’s revamped paywall is driving new readers and subscribers.” Nieman Journalism Lab. March 11, 2015. Accessed April 20, 2015. http://www.niemanlab.org/2015/03/after-the-archive-came-down-the-new-yorkers-revamped-paywall-is-driving-new-readers-and-subscribers/.

Eveland, William P., and Sharon Dunwoody. “User Control and Structural Isomorphism or Disorientation and Cognitive Load? Learning From the Web Versus Print.” *Communication Research* 28, no. 1 (February 1, 2001): 48–78.

Foucault, Michel. *Archaeology of Knowledge.* London: Tavistock, 1972.

Foust, James C. *Online journalism: principles and practices of news for the Web.* 2nd ed. Scottsdale, Ariz: Holcomb Hathaway, 2009.

Fragoso, Suely. “Understanding links: Web Science and hyperlink studies at macro, meso and micro-levels.” *New Review of Hypermedia and Multimedia* 17, no. 2 (2011): 163–198.

Freedman, Jonathan, N. Katherine Hayles, Jerome McGann, Meredith L. McGill, Peter Stallybrass, and Ed Folsom. “Responses to Ed Folsom’s "Database as Genre: The Epic Transformation of Archives".” *PMLA* 122, no. 5 (October 2007): 1580–1612.

Gahran, Amy. “Swimming Lessons for Journalists.” PBS Idea Lab. July 4, 2008. Accessed March 15, 2015. http://www.pbs.org/idealab/2008/07/swimming-lessons-for-journalists005/.

Gillespie, Tarleton. “The Politics of ’Platforms’.” *New Media & Society* 12, no. 3 (May 1, 2010): 347–364.

Gillmor, Dan. *We the Media: Grassroots Journalism By the People, For the People.* O’Reilly, 2006.

Gooding, Paul. “Exploring Usage of Digitised Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online.” 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.

Halasz, Frank. “Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems.” *Commun. ACM* 31, no. 7 (July 1988): 836–852.

Hargittai, Eszter. “The Role of Expertise in Navigating Links of Influence.” In *The Hyperlinked Society*, 85–103. Ann Arbor, MI: University of Michigan Press, May 23, 2008.

Helmond, Anne. “Exploring the Boundaries of a Website: Using the Internet Archive to Study Historical Web Ecologies.” MIT8, Cambridge, MA, 2013. Accessed April 19, 2015. http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-using-the-internet-archive-to-study-historical-web-ecologies/.

———. “The Algorithmization of the Hyperlink.” *Computational Culture* 3 (November 2013).

Hermida, Alfred. *Twittering the News: The Emergence of Ambient Journalism.* SSRN Scholarly Paper ID 1732598. Rochester, NY: Social Science Research Network, July 8, 2010.

Heuvel, Charles van de. “Building Society, Constructing Knowledge, Weaving the Web: Otlet’s Visualizations of a Global Information Society and His Concept of a Universal Civilization.” In *European Modernism and the Information Society*, 127–153. Ashgate Publishing, Ltd., 2008.

Himelboim, Itai. “The International Network Structure of News Media: An Analysis of Hyperlinks Usage in News Web sites.” *Journal of Broadcasting & Electronic Media* 54, no. 3 (August 17, 2010): 373–390.

Holmes, David. “How Rap Genius and explainer sites are killing music journalism.” PANDODAILY. January 20, 2015. Accessed March 16, 2015. http://pando.com/2015/01/20/how-rap-genius-and-explainer-sites-are-killing-music-journalism/.

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/.

Hsu, Chien-leng, and Han Woo Park. “Sociology of Hyperlink Networks of Web 1.0, Web 2.0, and Twitter: A Case Study of South Korea.” *Social Science Computer Review* 29, no. 3 (September 21, 2010): 354–368.

Hu, Elise. “Contextualizing Context.” Hey Elise. March 15, 2010. Accessed April 20, 2015. http://www.heyelise.com/2010/03/15/contextualizing-context/.

Innovation. New York Times, March 24, 2014. Accessed April 20, 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/.

Jarvis, Jeff. “New rule: Cover what you do best. Link to the rest.” BUZZMACHINE. February 22, 2007. Accessed April 20, 2015. `http://buzzmachine.com/2007/02/22/new-rule-cover-what-you-do-best-link-to-the-rest/`.

Karlsson, Michael, Christer Clerwall, and Henrik Örnebring. “Hyperlinking practices in Swedish online news 2007-2013: The rise, fall, and stagnation of hyperlinking as a journalistic tool.” *Information, Communication & Society* (2014): 1–17. Accessed February 8, 2015. `http://dx.doi.org/10.1080/1369118X.2014.984743`.

Kaufman, Leslie. “Vox Takes Melding of Journalism and Technology to a New Level.” *The New York Times* (April 6, 2014). Accessed April 20, 2015. `http://www.nytimes.com/2014/04/07/business/media/voxcom-takes-melding-of-journalism-and-technology-to-next-level.html`.

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. “How Vox aggregates.” Vox. April 13, 2015. Accessed April 14, 2015. `http://www.vox.com/2015/4/13/8405999/how-vox-aggregates`.

———. “Vox is our next.” The Verge. January 26, 2014. Accessed April 20, 2015. `http://www.theverge.com/2014/1/26/5348212/ezra-klein-vox-is-our-next`.

Kleinberg, Jon M. “Authoritative Sources in a Hyperlinked Environment.” *J. ACM* 46, no. 5 (September 1999): 604–632. Accessed February 7, 2015. `http://doi.acm.org/10.1145/324133.324140`.

Kopytoff, Igor. “The Cultural Biography of Things: Commoditization as Process.” In *The Social Life of Things: Commodities in Cultural Perspective*, edited by Arjun Appadurai, 64–91. Cambridge University Press, 1986.

Kovach, Bill, and Tom Rosenstiel. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect.* Crown/Archetype, 2001.

Krajewski, Markus. *Paper Machines: About Cards and Catalogs.* Cambridge: MIT Press, 2011.

Lanier, Jaron. *Who Owns the Future?* New York: Simon & Schuster, 2013.

Larsson, Anders Olof. “Staying In or Going Out?” *Journalism Practice* 7, no. 6 (December 2013): 738–754.

Lee, Jennifer, and Eric Price. “Version Controlling the News: How We Can Archive.” South by Southwest, Austin, TX, March 18, 2013. Accessed April 20, 2015. `http://www.slideshare.net/jenny8lee/newsdiffs`.

Levy, Nicole. “Time.com opens its ’Vault’.” Capital New York. November 12, 2014. Accessed April 20, 2015. http://www.capitalnewyork.com/article/media/2014/11/8556503/timecom-opens-its-vault.

Luzón, María José. “Scholarly Hyperwriting: The Function of Links in Academic Weblogs.” *Journal of the American Society for Information Science and Technology* 60, no. 1 (January 2009): 75–89.

Manjoo, Farhad. “How To Make a Viral Hit in Four Easy Steps.” *Slate* (June 26, 2012). Accessed April 19, 2015. http://www.slate.com/articles/technology/technology/2012/06/_21_pictures_that_will_restore_your_faith_in_humanity_how_buzzfeed_makes_viral_hits_in_four_easy_steps_.single.html.

Manoff, Marlene. “Archive and Database as Metaphor: Theorizing the Historical Record.” *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398.

Manovich, Lev. “Database as Symbolic Form.” *Convergence* 5, no. 2 (June 1, 1999): 80–99.

Markham, Annette N. “Undermining ’data’: A critical examination of a core term in scientific inquiry.” *First Monday* 18, no. 10 (September 21, 2013). http://firstmonday.org/ojs/index.php/fm/article/view/4868.

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.” 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.

Moulthrop, Stuart. “To Mandelbrot in Heaven.” In *Memory Machines: The Evolution of Hypertext*, by Belinda Barnet. London: Anthem Press, July 15, 2013.

Mullin, Benjamin. “How The Boston Globe is covering the Boston Marathon bombing trial.” Poynter. 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.

Myers, Steve. “Liveblogging SXSW: The Future of Context in Journalism.” Poynter. March 15, 2010. Accessed April 20, 2015. http://www.poynter.org/news/101399/liveblogging-sxsw-the-future-of-context-in-journalism/.

Napoli, Philip. "Hyperlinking and the Forces of "Massification"." In *The Hyperlinked Society: Questioning Connections in the Digital Age*, edited by Lokman Tsui and Joseph Turow. Ann Arbor, MI: University of Michigan Press, May 23, 2008.

Nelson, Theodor H. "Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate." In *Proceedings of the 1965 20th National Conference*, 84–100. New York, NY, USA: ACM, 1965.

———. *Computer Lib / Dream Machines*. Self-published, 1974.

———. "Ted Nelson's Computer Paradigm, Expressed as One-Liners." 1999. Accessed April 19, 2015. `http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html`.

Noguchi, Yuki. "A New Internet Domain: Extortion Or Free Speech?" NPR. April 7, 2015. Accessed April 20, 2015. `http://www.npr.org/blogs/alltechconsidered/2015/04/07/397886748/a-new-internet-domain-extortion-or-free-speech`.

Nothman, Joel. "Grounding event references in news." PhD diss., University of Sydney, 2013. Accessed April 20, 2015. `http://hdl.handle.net/2123/10609`.

Nyce, James M., and Paul Katin. "Innovation, Pragmaticism, and Technological Continuity: Vannevar Bush's Memex." *Journal of the American Society for Information Science* 40, no. 3 (May 1989): 214–220.

Oblak, Tanja. "The Lack of Interactivity and Hypertextuality in Online Media." *Gazette* 67, no. 1 (February 1, 2005): 87–106. Accessed December 12, 2013. `http://gaz.sagepub.com/content/67/1/87`.

"One in five online scholarly articles affected by 'reference rot'." Los Alamos National Laboratory. January 26, 2015. Accessed April 20, 2015. `http://www.lanl.gov/discover/news-release-archive/2015/January/01.26-scholarly-articles-affected-by-reference-rot.php`.

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

Parikka, Jussi. "Archival Media Theory: An Introduction to Wolfgang Ernst's Media Archaeology." In *Digital Memory and the Archive*, by Wolfgang Ernst, 1–22. Minneapolis: University of Minnesota Press, 2012.

Paul, Nora. "Content: A Re-Visioning." Interactive Newspapers '95, Dallas, TX, February 6, 1995. Accessed February 8, 2015. `http://www.lehigh.edu/~jl0d/J366-99/366npaul.html`.

Reagle, Joseph. *Good Faith Collaboration: The Culture of Wikipedia*. Cambridge, Mass.: MIT Press, 2010.

Romenesko, Jim. “U.S. News deletes archived web content published before 2007.” Romenesko. February 18, 2014. Accessed April 20, 2015. http://jimromenesko.com/2014/02/18/u-s-news-deletes-content-published-before-2007/.

Rosen, Jay. “National Explainer: A Job for Journalists on the Demand Side of News.” PRESSTHINK. August 13, 2008. Accessed April 20, 2015. http://archive.pressthink.org/2008/08/13/national_explain.html.

———. “News Without the Narrative Needed to Make Sense of the News: What I Will Say at South by Southwest.” PRESSTHINK. March 7, 2010. Accessed April 20, 2015. http://pressthink.org/2010/03/news-without-the-narrative-needed-to-make-sense-of-the-news-what-i-will-say-at-south-by-southwest/.

Rosenberg, Daniel. “Data Before the Fact.” In *"Raw Data" is an Oxymoron*, edited by Lisa Gitelman, 15–40. Cambridge, MA: MIT Press, 2013.

———. “Early Modern Information Overload.” *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 1–9.

Samis, Peter, and Tim Svenonius. “The X, Y, and Z of digital storytelling: Dramaturgy, directionality, and design.” In *Museums and the Web.* Chicago, February 10, 2015. Accessed April 26, 2015. http://mw2015.museumsandtheweb.com/paper/the-x-y-z-of-digital-storytelling-dramaturgy-directionality-and-design/.

Schoenberger, Viktor. *Useful Void: The Art of Forgetting in the Age of Ubiquitous Computing* Working Paper RWP07-022. Cambridge, MA: John F. Kennedy School of Government, Harvard University, April 2007. Accessed April 21, 2015. http://ksgnotes1.harvard.edu/Research/wpaper.nsf/rwp/RWP07-022.

Schudson, Michael. *Discovering the News: A Social History of American Newspapers*. New York: Basic Books, 1978.

Seaver, Nick. “Algorithmic Recommendations and Synaptic Functions.” *Limn*, no. 2 (2012): 44–47. http://limn.it/algorithmic-recommendations-and-synaptic-functions/.

Shafer, Jack. “Who Said It First?” *Slate* (August 30, 2010). Accessed April 19, 2015. http://www.slate.com/articles/news_and_politics/press_box/2010/08/who_said_it_first.html.

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

Silverman, Craig. “Endangered Species: News librarians are a dying breed.” Columbia Journalism Review. January 29, 2010. Accessed April 20, 2015. http://www.cjr.org/behind_the_news/endangered_species.php.

———. “Why Vox (and other news orgs) could use a librarian.” Nieman Journalism Lab. April 22, 2014. Accessed April 20, 2015. http://www.niemanlab.org/2014/04/why-vox-and-other-news-orgs-could-use-a-librarian/.

Sloan, Robin. “Stock and Flow.” Snarkmarket. January 18, 2010. Accessed April 20, 2015. http://snarkmarket.com/2010/4890.

Smith, Ben. “Editor’s Note: An Apology To Our Readers.” BUZZFEED. July 25, 2014. Accessed April 19, 2015. http://www.buzzfeed.com/bensmith/editors-note-an-apology-to-our-readers.

Smith, D.A., R. Cordell, and E.M. Dillon. “Infectious texts: Modeling text reuse in nineteenth-century newspapers.” In *2013 IEEE International Conference on Big Data*, 86–94. October 2013.

Sparrow, Betsy, Jenny Liu, and Daniel M. Wegner. “Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips.” *Science* 333, no. 6043 (August 5, 2011): 776–778.

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

The Editorial Board. “Digital Neglect at the Library of Congress.” *The New York Times* (April 4, 2015). Accessed April 20, 2015. http://www.nytimes.com/2015/04/05/opinion/sunday/digital-neglect-at-the-library-of-congress.html.

Thompson, Derek. “Upworthy: I Thought This Website Was Crazy, but What Happened Next Changed Everything.” *The Atlantic* (November 14, 2013). Accessed April 19, 2015. http://www.theatlantic.com/business/archive/2013/11/upworthy-i-thought-this-website-was-crazy-but-what-happened-next-changed-everything/281472/.

Tremayne, Mark. “Applying Network Theory to the Use of External Links on News Web Sites.” In *Internet Newspapers: The Making of a Mainstream Medium*, edited by Xigen Li. Routledge, September 13, 2013.

———. “Manipulating interactivity with thematically hyperlinked news texts: a media learning experiment.” *New Media & Society* 10, no. 5 (October 1, 2008): 703–727.

———. “The Web of Context: Applying Network Theory to the Use of Hyperlinks in Journalism on the Web.” *Journalism & Mass Communication Quarterly* 81, no. 2 (2004): 237–253.

Trevino, James, and Doerte Doemeland. *Which World Bank reports are widely read?* WPS6851. The World Bank, May 1, 2014.

Trotter, J.K. “Don’t Ask BuzzFeed Why It Deleted Thousands of Posts.” Gawker. August 14, 2014. Accessed April 19, 2015. `http://gawker.com/don-t-ask-buzzfeed-why-it-deleted-thousands-of-posts-1621830810`.

———. “Over 4,000 BuzzFeed Posts Have Completely Disappeared.” Gawker. August 12, 2014. Accessed April 19, 2015. `http://gawker.com/over-4-000-buzzfeed-posts-have-completely-disappeared-1619473070`.

———. “Plagiarist of the Day: Mic News Director Jared Keller.” Gawker. February 11, 2015. Accessed April 19, 2015. `http://tktk.gawker.com/plagiarist-of-the-day-mic-news-director-jared-keller-1684959192`.

Tufekci, Zeynep. “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.” *arXiv:1403.7400 [physics]* (March 28, 2014).

Veelen, Ijsbrand van. *Alle Kennis van de Wereld (Biography of Paul Otlet)*. Noorederlicht, 1998. Accessed December 19, 2014. `http://archive.org/details/paulotlet`.

Vobič, Igor. “Practice of Hypertext: Insights from the online departments of two Slovenian newspapers.” *Journalism Practice* 8, no. 4 (August 8, 2013): 357–372. Accessed December 12, 2013. `http://www.tandfonline.com/doi/abs/10.1080/17512786.2013.821325`.

Walker, Jill. “Links and Power: The Political Economy of Linking on the Web.” In *ACM Hypertext Conference*. Baltimore, MD, June 2002.

Waterton, Claire. “Experimenting with the Archive: STS-ers As Analysts and Co-constructors of Databases and Other Archival Forms.” *Science, Technology & Human Values* 35, no. 5 (September 1, 2010): 645–676.

Weber, Matthew S., and Peter Monge. “The Flow of Digital News in a Network of Sources, Authorities, and Hubs.” *Journal of Communication* 61, no. 6 (2011): 1062–1081. Accessed February 8, 2015. `http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1460-2466.2011.01596.x/abstract`.

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`.

Weinberger, David. *Everything Is Miscellaneous: The Power of the New Digital Disorder*. New York, NY: Macmillan, April 2008.

Wemple, Erik. “The ravages of BuzzFeed’s Benny Johnson.” *The Washington Post* (July 27, 2014). Accessed April 19, 2015. `http://www.washingtonpost.com/blogs/erik-wemple/wp/2014/07/27/the-ravages-of-buzzfeeds-benny-johnson/`.

Wolf, Gary. “The Curse of Xanadu.” *Wired* 3, no. 6 (June 1995).

Yates, Frances Amelia. *The Art of Memory.* London: Routledge, 1966.

Yeo, Richard. “A Solution to the Multitude of Books: Ephraim Chambers’s "Cyclopaedia" (1728) as "The Best Book in the Universe".” *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 61–72.

Yglesias, Matthew. “Refreshing the evergreen.” Vox. January 15, 2015. Accessed February 4, 2015. `http://www.vox.com/2015/1/15/7546877/evergreen-experiment`.

Zimmer, Michael. “Renvois of the Past, Present and Future: Hyperlinks and the Structuring of Knowledge from the Encyclopédie to Web 2.0.” *New Media & Society* 11, no. 1 (February 1, 2009): 95–113.

Zittrain, Jonathan. *The Fourth Quadrant.* SSRN Scholarly Paper ID 1618042. Rochester, NY: Social Science Research Network, May 30, 2010.

Zittrain, Jonathan, Kendra Albert, and Lawrence Lessig. *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*. SSRN Scholarly Paper ID 2329161. Rochester, NY: Social Science Research Network, October 1, 2013.

Zuckerman, Ethan. “Helen Nissenbaum on Ad Nauseum, resistance through obfuscation, and weapons of the weak.” My Heart’s in Accra. October 6, 2014. Accessed April 20, 2015. `http://www.ethanzuckerman.com/blog/2014/10/06/helen-nissenbaum-on-ad-nauseum-resistance-through-obfuscation-and-weapons-of-the-weak/`.

———. “International reporting in the age of participatory media.” *Daedalus* 139, no. 2 (April 1, 2010): 66–75.