The Missing Links: Archiving News and Filtering History

by

Liam Phalen Andrew

B.A., Yale University (2008)

Submitted to the Department of Comparative Media Studies
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
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: Archiving News and Filtering History

by

Liam Phalen Andrew

Submitted to the Department of Comparative Media Studies
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 increases on the web, citizen journalism has threatened the traditional, 20th-century role of journalism and institutional newsmaking. Journalists and publishers are beginning to adapt and evolve to fit into the new news landscape, but the role of legacy media institutions, and even newer, digital-native outlets, is still in flux and under debate. 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, one that places historical context on the same footing as new content, and emphasizes the journalist's role as explainer and verifier. Informed by a theoretical understanding of the web's structural affordances and limitations, and especially by the untapped networking power of the hyperlink, publishers can offer an archive-oriented model of sustainable and scalable journalism. Drawing from theories in library and computer science, an archivally focused journalistic model can save time for reporters and improve the research and reading process for journalists and audiences alike, treating news items as part of a conversation rather than a static box or endless feed, and putting the past in fuller and richer dialogue with the present.

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

Acknowledgments

My acknowledgements will go here.

Contents

0.1 Preface

For the past two years I've been doing research online about the perils of doing research online. This has made my head spin more than once. It is a slippery subject; I keep running into the very problems I want to address. I've encountered so many tantalizing online resources, only to discover 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" for organizing our resources and ideas. And I've spent sleepless nights trying to organize my thoughts, and make connections between everything I read, process, and understand. I want my notes and citations to reflect, enhance, and expand my own memories and ideas; too often they obfuscate and distract from them instead. Computers are very good at storing and remembering information, but they are less adept at making connections between the bits of data that they remember.

This is a problem on a collective as well as personal level; it affects not only personal memory, but collective history. We are overloaded with information on a massive, unprecedented scale, and new material arrives faster than we can contain it. 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 and less tried methods for determining what something is about, whether it's worth saving, and whether it has any meaningful impact on the world. And the librarians and archivists aren't the main ones doing the sorting.

The web has enabled an unprecedented explosion of data, and an unprecedented amount of access to it. Another way to frame this is that the past – the archive – is bigger and more accessible than ever before. It's an exciting prospect, but the archive has exploded and networked to an unmanageable scale. Machines help to sort out the results, but the best machines don't treat the web as an archive; they treat it as a network. Instead of using categories, they rely on links.

The link is an ideally situated entity for the post-deconstruction, networked age.

There is no hierarchy in a network, only a collection of nodes and links.[1] Unlike in a library, bookstore, department store, or anywhere that contains physical *things* there is no traditional category; no singular, fixed decision made about what something *is*, what it means, or where it belongs. Instead machines look for what and where something *points to*, and let the links sort everything out. Links serve a double function: you not only see who is linking, but how many links there are. Links not only categorize, they measure importance and impact.

I especially notice the powers of links in my work as a software engineer and backend web developer. I've built a variety of news and event curation and monitoring applications, using many different programming languages and frameworks.[2] I am essentially a Link Wrangler. I corral articles, emails, events, tweets, and the like in order to classify and ultimately rank them for users. But I have grown frustrated by the link. Links tend to be the unique identifier for a resource, and an atomic unit of information. But links are more elusive and complicated than that; they contain multitudes, and are aggregations themselves. I'm often frustrated by the link's limitations in defining, classifying, and measuring online content.

In this thesis I aim to unpack what links do – and what they fail to do – for developers, creators, publishers, aggregators, and everyday users of the web. In doing so, I hope to elucidate the ways that information becomes knowledge, news becomes history, and the archive unfolds in a hyperlinked environment. I want to bridge the ways links affect public discourse and cultural memory. My goal is to speak on one hand to news and media organizations, to enact and enliven their own archives and research tools, and on the other hand to libraries and archives, to inject historical resources and context into current events and social issues.

In the process, I hope to help define the borders and the limitations of the web in particular, and networks in general; what's exciting and new about big data, and what's risky. The link is a smaller and more manageable entity than the network, so it deserves more exploration to see what the smaller unit can teach us about the

1. However, there are centers in a network.

2. e.g. for MIT HyperStudio (*Artbot*, Ruby on Rails), Nieman Journalism Lab (*Fuego*, Flask), and Wiser (Django).

bigger picture. Some other writings aim to teach the reader how to "think networks". I wonder if it might be easier and healthier to "think links".

The link might seem like an abstract, academic or even trivial thing, one that is too academic or tangential to real industries like news and libraries. But "links give power." They are the foundation of the web, and they serve as the battleground for much of its political economy. The goal is to keep my focus not on any one medium or industry, but instead on the nature, the identity, and the mechanics of comparison, difference, and connection.[3]

Whenever anyone blogs about, emails, tweets, likes, or searches for a resource, that resource is recalibrated, recategorized and re-measured. So in one sense, we're all archivists: we constantly save, edit, and delete our traces on emails, files, and social media–and this in turn affects what others will see. We make links, and links make history. But the web is no traditional archive; it's a cloud, not a vault. And the archive is in new hands, where we can't determine or even know the rules under which information has influence, and we can't opt out.

3. One could say that I've taken the "comparative" part of Comparative Media Studies too seriously.

Chapter 1

Introduction

1.1 The Stakes

In the summer of 2014, I was closely following news about the news. Working for the Nieman Journalism Lab as a Google Journalism Fellow, I split my time between writing articles about innovation in journalism, monitoring social media for stories worth sharing, and building an app that tracked link-sharing and conversations on Twitter (so even when writing code, I was following the news). During this summer, several news events coalesced to form the backbone of the exploration of this thesis topic. While these incidents may seem unrelated, my goal will be to showcase what these news events have in common, and set the stakes for the exploration of the changing nature of online research and cultural production on the web.

1.1.1 BuzzFeed plagiarism incident

In the summer of 2014, Benny Johnson, a BuzzFeed editor, was accused of plagiarism by two enterprising web-divers known only as @blippoblappo and @crushingbort. Publishing the article on a blog created just for the occasion called Our Bad Media, it was initiated when Johnson attempted to call out the Independent Journal Review for plagiarizing his own work. @blippoblappo and @crushingbort noticed the irony of a BuzzFeed writer accusing another publisher of stealing. BuzzFeed has long

been accused of aggregating, remixing, and appropriating other outlets' work without payment. Perhaps because of this, they turned to web searches for examples of Johnson's own lifting.

The pair of detectives were likely not aware of how deep the copying went, though; they found three instances of unattributed sentences taken from everywhere from the Guardian to Wikipedia to Yahoo! Answers. When BuzzFeed editor Ben Smith replied by calling Johnson "one of the web's deeply original writers," @blippoblappo and @crushingbort responded with six more offenses, here from the National Review, About.com, and the New York Times.

This set forced BuzzFeed to investigate, and a day later they fired Johnson and apologized to their readers; they had found a whopping 41 plagiarized phrases in 500 Johnson pieces. The rate and ease at which these seem to have been found is startling. If two researchers found so much bad-faith plagiarism in one day, and BuzzFeed's internal investigation had turned up dozens more, how could they – how could *anyone* not have not discovered this during any of Johnson's [HOW MANY?] 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. 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. BuzzFeed's problem is still fairly new, in that it is trying to ethically aggregate and reappropriate from other online sources. While it's clear that Johnson stepped across this ethical line, it's still unclear where this line is. Smith's first reaction suggested that three offenses was not enough; he also implied that plagiarism on older articles or trite listicles would be more acceptable than newer, investigative pieces. But it seems that Johnson's attitude towards online aggregation bled into even more "original" investigative works.

While the legal and ethical implications of aggregating is a crucial topic for jour-

nalism and e-research in the 21st century, this is not so much my focus as the way in which the aggregational mentality changes the *practice* of journalism and e-research. This is the case for both what can actually be found online, and what we perceive to be findable online. It is amazing that Johnson did not see himself as vulnerable; despite his obvious offenses, he assumed that no one would ever find them, and quickly accused others of plagiarism instead.

Moreover, the incident reflects a new paradigm of attribution and authorship. 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 as if tax dollars allowed him to. His liberal use of Yahoo! Answers and About.com also points to interesting questions; did he somehow feel that it was more ethical 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 treat them as original?

Johnson's safest option would have been to simply *link to* the sources, and one wonders whether he now wishes he had. Linking would be safe; but it would also be tedious. It would 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 bald pilfering of often 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 the anchor text. 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. In more extreme cases, BuzzFeed has deleted older content outright that did not adhere to their journalistic

standards.

In short, this controversy and BuzzFeed's reaction to it encompass many of the problems with assigning attribution and measuring impact on the web. It also points to the difficulty of online research, and the lack of standards and technologies for ethical, creative, original remix and reuse. This is as true for a tweet from today as it is a photo from decades ago. As newsrooms increasingly play the role of aggregator and context provider, they have a newfound ability *and* responsibility to leverage archives – whether their own proprietary archives or the web-as-archive – to create and appropriate old content into new stories, merging news and history, placing sensational events in the longer phenomena that surround them, and centering the daily news in broader contexts.

1.1.2 New York Times Innovation Report

A couple months before BuzzFeed's plagiarism incident, a staffer at the New York Times leaked the company's internal Innovation Report, which my colleagues at the Nieman Lab called "one of the key documents of this media age." The report looks closely and especially at the revitalization of its archives.

Not only do the archives have the power to historicize current pieces, trends, and events, they can also have amazing financial value, giving new life to old content that is repurposed, repackaged, and recontextualized.

The problem goes both ways; while not enough tools exist for Times staffers to resurface the past, it's also true that their new content is not properly prepared for the future. The Innovation Report likewise cites many problems that the company has with structured data and categorization.

Journalists have traditionally called the archive "the morgue," and the Times Innovation Report both explains why this is the case and challenges its issues.

Finally, the Innovation Report confirmed that the role of repackaging and reappropriating old content was not just a problem for the BuzzFeeds and Huffington Posts of the world; old stalwarts with canonical archives are in the same business. This is, in effect, the new business of journalism: while citizens and activists increasingly

serve as the newsbreakers, the journalists must take a step away from the epicenter of the event and report on everything that surrounds it instead. The web provides many new tools and affordances to do this creatively and engagingly; but the news industry has a long way to go and a lot to learn.

1.1.3 Project Xanadu and Newslynx

In this same summer, Theodor Nelson's Project Xanadu was finally released on the web. First conceived 45 years prior, Project Xanadu was famously the first hypertext project, under development for decades. Xanadu was the realization of an alternate hypertext system, one in which many of the pitfalls of the web – the problems of attribution, measurement, and research that I aim to highlight – are laid bare to be scrutinized and reimagined. On one hand, the fact that the project was finally released on the web seems like a sort of admisison of defeat. On the other hand, the project's persistence and rebirth has potential to help researchers think of online archives and repositories in a new way. Indeed, Nelson is setting his sights on overtaking PDFs.

As the coiner of the term "hypertext" and one of its pioneers, Nelson has a wide set of acolytes and followers. Among them are the founders of NewsLynx, a research project and platform under development at Columbia University's Tow Center for Digital Journalism. In August 2014, they 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. They pointed out the "hole at the center of the web" that let Google "step in and play librarian." Here they recognized how intensely the structure of the web has affected its content, whether by allowing for transgressions like Benny Johnson's, obscuring archives like the New York Times', and leaving Google to determine how to sort everything out.

So in the summer of 2014, not only did Xanadu come to life, but its concept was validated. But in both cases (from Xanadu itself and its NewsLynx acolytes), the solutions were grafted onto the web, rather than proposed as a radical alternative. The web is only to be added and appended to, not replaced. In later sections, I will be looking closely at these two appendages to analyze their histories, strengths, and

failures, and to suggest what they can teach us about structure of the web itself, and the ways that our thinking might have and might need to adapt to it.

1.1.4 Semantic Web

1.2 Outline of chapters and methods

Chapter 2

The Size and Shape of Archives

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 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. Such terms can expand and contract in meaning, and in the process they skirt debate and risk glossing over embedded biases and controversies. They are at the top of a linguistic hierarchy, and threaten to subsume the nuances and contingencies within the subcategories. At the risk of sounding trite, everything is a thing, which is logically impossible to argue (and in fact, the ontology language that underlies the Semantic Web uses "thing" as the base layer under which all other words go). But what is a document, or data? How does our use of these words carry contextual weight?

Tech 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 18$^{\text{th}}$ 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 counter-

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), accessed May 1, 2014, `http://firstmonday.org/ojs/index.php/fm/article/view/4868`.

part "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 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 co-opted "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 much controversy. Ironically, many of the web's detractors pointed precisely to the web's lack of an "original" document copy as its primary shortcoming, a critique that informs my own inquiry into its infrastructure.[7]

---

3. Johanna Drucker, "Humanities Approaches to Graphical Display," *Digital Humanities Quarterly* 5, no. 1 (2011), accessed October 24, 2013, `http://www.johannadrucker.com/pdf/hum_app.pdf`.

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, accessed April 22, 2014, `http://nms.sagepub.com/content/12/3/347`.

6. Tim Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (HarperBusiness, November 7, 2000).

7. Theodor H. Nelson, "Ted Nelson's Computer Paradigm, Expressed as One-Liners," 1999, ch. 18, accessed December 15, 2013, `http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html`; Jaron Lanier, *Who Owns the Future?* (New York, NY: Simon & Schuster,

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 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

May 7, 2013).

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

9. Jacques Derrida, "Archive Fever: A Freudian Impression," trans. Eric Prenowitz, ArticleType: research-article / Full publication date: Summer, 1995 / Copyright © 1995 The Johns Hopkins University Press, *Diacritics* 25, no. 2 (July 1, 1995): 9, accessed December 2, 2013, `http://www.jstor.org/stable/465144`.

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 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

10. Jorge Luis Borges, *Collected Fictions*, trans. Andrew Hurley (New York, N.Y., U.S.A.: Penguin Books, September 1, 1999), 325.

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, accessed May 3, 2014, `http://sth.sagepub.com.libproxy.mit.edu/content/35/5/645`.

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. 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. 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

13. Lev Manovich, "Database as Symbolic Form," *Convergence: The International Journal of Research into New Media Technologies* 5, no. 2 (June 1, 1999): 84, accessed December 2, 2013, `http://con.sagepub.com/content/5/2/80`.

14. Marlene Manoff, "Archive and Database as Metaphor: Theorizing the Historical Record," *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398, accessed December 2, 2013, `http://muse.jhu.edu.libproxy.mit.edu/login?auth=0&type=summary&url=/journals/portal_libraries_and_the_academy/v010/10.4.manoff.html`; 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, accessed December 2, 2013, `http://www.mlajournals.org/doi/abs/10.1632/pmla.2007.122.5.1580`; 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): 217–231, accessed December 10, 2013, `http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=4645639&site=ehost-live`.

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

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. As Anne Helmond emphasizes, archives are also deeply embedded within and linked to one another through 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 information society an "information management society."[18] Belinda Barnet considers it a "pack-rat" mentality, and Derrida succinctly titles the phenomenon "archive fever."[19] Viktor Schoenberger writes that by default, the web saves rather than forgets; this may be the effect, but it misplaces the agency at hand.[20] The web stores nothing,

16. Freedman et al., "Responses to Ed Folsom?," 1588.

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

18. Jussi Parikka, "Archival Media Theory: An Introduction to Wolfgang Ernst′s Media Archaeology," in *Digital Memory and the Archive*, by Wolfgang Ernst, Explores how media infrastructure, not content, shapes contemporary digital culture (Minneapolis: University of Minnesota Press, 2012), 2, accessed December 19, 2014, `https://www.upress.umn.edu/book-division/books/digital-memory-and-the-archive`.

19. Barnet, "Pack-rat or Amnesiac?"; Derrida, "Archive Fever."

20. Viktor Schoenberger, *Useful Void: The Art of Forgetting in the Age of Ubiquitous Comput-*

while *we* save, store, and sort with abandaon.

While the following chapters will increasingly hone in on the archives of legacy and digital media publishers, my use of the term here may be broader and slipperier, in order to place the news publishing archive in the context of other archives, and glean lessons from its usual shape and scope. An archive is a collection that is organized for future retrieval and reuse; it is any collection that is not a dump. This encompasses traditional archives, modern databases, and the algorithms and interfaces in between.

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."[21] 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.[22]

The archive is the container—content's resting place—but it is not the only containing force on the web. 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

ing Working Paper RWP07-022 (Cambridge, MA: John F. Kennedy School of Government, Harvard University, April 2007), accessed December 8, 2013, `http://ksgnotes1.harvard.edu/Research/wpaper.nsf/rwp/RWP07-022`.

21. Andrew Whitacre, *Henry Jenkins Returns*, accessed March 10, 2014, `http://cmsw.mit.edu/henry-jenkins-returns/`.

22. 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/`.

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 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 wrappers around the original source. [23] 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 moments that mark events in a thing's history. Kopytoff complicates the idea of any single, indelible act of

23. (add figures here; one abstract concentric-circle view, one concrete view with a sample webpage)

categorization on an object—instead, an object is "classified and reclassified into culturally constituted categories."[24] This especially lends itself to digital content as well due to its 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 "https://www.flickr.com/photos/paulaloe/621178137/" is in over a dozen Flickr groups, albums, and galleries, such as "Inspiration," "Hempstead Harbor," and "On the Water."[25] One user might find it when searching for the Inspiration series jet boat (part of the photo's content), while another could be searching for inspirational photos (part of the photo's effect). Still another user could be looking for photos taken on a Nikon D200, or photos with more than 10,000 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 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.

As such, 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 homo-

24. Igor Kopytoff, "The Cultural Biography of Things: Commoditization as Process," in *The Social Life of Things: Commodities in Cultural Perspective*, ed. Arjun Appadurai (Cambridge University Press, 1986), 64–91.

25. (add image or screenshot here)

geneity of too much lumping."[26] 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 jet boat photo could turn up in a news article about the Inspiration boat, or about Hempstead Harbor, or about the Nikon D200. 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 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 more than 200 standards imposed and enacted when a person sends an email; these standards interact and depend on one another in important ways.[27] *Sorting Things Out*, along with Star's companion article "The Ethnography of Infrastructure," point to the large-scale effects of small-scale sorting.[28] They highlight the problems and limits with traditional classification, and suggest ways to render it more dynamic and responsive.

26. Kopytoff, "The Cultural Biography of Things: Commoditization as Process."

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

28. Susan Leigh Star, "The Ethnography of Infrastructure," *American Behavioral Scientist* 43, no. 3 (November 1, 1999): 377–391, accessed May 3, 2014, `http://abs.sagepub.com.libproxy.mit.edu/content/43/3/377`.

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.

2.2.1 The URL

As Tim Berners-Lee tells it, the Uniform Resource Locator, or URL, was one of the most difficult concepts to develop and understand as he began to weave the web.[29] To this day, he sees it as the web's most foundational element, and its importance is amplified even further in the Semantic Web. 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. As I will discuss in the following chapter, 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." 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 specific document (or piece of content), but one that ends with "/users?friendrequest=true" is more likely to be telling a social

29. Berners-Lee, *Weaving the Web*, FIND PAGE.

media site to request friendship with another user. Indeed, at the current stage of the web's evolution, a URL is not by definition a resource; it 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.[30]

Eszter Hargittai points to the need for technical expertise in order to properly understand a URL and the implications behind it.[31] It is easier for a user with less 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. For instance, some users do not fully understand the difference between the web and Google, or whether a link in an article or feed will take them to the same source (the same domain) or a different one entirely. The URL thus serves as a barrier to understanding and retrieving information from the web for those who have less familiarity; technical knowledge enables information retrieval, and a lack thereof leaves users in the dark and vulnerable.

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 look 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 often easily be "broken" through a simple reordering of the URL, allowing information access through technical knowledge.[32] Again this stratifies those who can navigate the web and those who are further 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

30. (figure here with detailed annotations of a single URL)
31. 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.
32. **find cite for this?**.

analyses of links (discussed at greater length in Chapter 4) rely on URLs to gain insight about the content within, because it is often too technically taxing to gather richer metadata. They rely on domains to determine whether a link is an inlink or outlink—wheter it is altruistic or "nepotistic."[33] This is despite the fact that the URL is only a pointer or reference, and its address can hide or gloss over complex contingencies and nuances. For instance, an organization might be linking to its parent or sister company at a different domain, or it may register a ".org" website despite being a for-profit operation. Some studies have noted that users tend to trust ".org" websites, and many studies lean on these distinctions in quantitative link analysis; however, there is no technical or legal border between these particular domains.[34]

Country-based TLDs sometimes impose certain restrictions, so suffixes like ".es", ".uk" or ".tv" belong to Spain, the United Kingdom and Tuvalu, respectively. These countries are generally free to do what they like with their TLD, which leads to a heterogeneous set of practices that are not accounted for in higher-level quantitative analysis. Some studies use these TLDs to map international communication networks by analyzing the link flows between these TLDs; for instance, between Brazil and Germany, or between developing countries and Western hubs.[35] 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, as it contains social and conventional complexities. The ".tv" domain, for instance, was sold by the small Polynesian nation Tuvalu to a Verisign company; while Tuvalu maintains a 20 percent stake and receives 4 million dollars a year to lease it out, it means that

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

34. **find cite for this?**.

35. 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, accessed February 20, 2015, `http://link.springer.com/article/10.1007/s11135-013-9847-z`; 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, accessed December 17, 2013, `http://www.tandfonline.com/doi/abs/10.1080/13614568.2011.587030`; 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, accessed February 8, 2015, `http://dx.doi.org/10.1080/08838151.2010.499050`.

such quantitative analyses are glossing over the complex commercial and financial transactions that occur within these seemingly simple and purely technical URLs.[36]

URL "shorteners" such as those employed by the New York-based company bit.ly (whose top-level domain 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"). Shortened URLs have the same final reference point, but they no longer 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.[37]

Perhaps the explosion of the URL was an inevitable byproduct of the web's very structure. The web is widely distributed and highly networked; it shuns hierarchical organization schemes like the "domains" and "paths" of the URL itself. Indeed, both Berners-Lee and Theodor Nelson (the original coiner of the term "hypertext" and its first champion) explicitly highlighted the power of the link to cut across tree structures and find new, unexpected associations.[38] Where knowledge was once shaped like a tree, on the web it looks more like Deleuze and Guattari's rhizome.[39][40] As such, one cannot make sense of it using URLs alone. However, links offer a start.

36. **(find cite for this)**.

37. Anne Helmond, "The Algorithmization of the Hyperlink," *Computational Culture* 3 (November 2013), `http://computationalculture.net/article/the-algorithmization-of-the-hyperlink`.

38. 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): 104, accessed December 12, 2013, `http://nms.sagepub.com/content/11/1-2/95`.

39. Gilles Deleuze and Félix Guattari, *A thousand plateaus: capitalism and schizophrenia* (Minneapolis: University of Minnesota Press, 1987).

40. figure here too?

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.[41]

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."[42] 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*."[43] 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>..." 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

41. (figure here of nodes vs. network; also of tree vs. rhizome)
42. Bowker and Star, *Sorting Things Out*, 7.
43. Jill Walker, "Links and Power: The Political Economy of Linking on the Web" (Baltimore, MD, June 2002).

permitting "footnotes on footnotes on footnotes,"[44] 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 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 may follow something like Charles Sanders Peirce's semiotic triad; when a link says "click here" as opposed simply linking the text as so, it may be forming an indexical rather than symbolic relationship to the target.[45] 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."[46]

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 later. On the other hand, the lack of relation between text and target seems to point to the 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. One exception, a study of "scholarly hyperwriting" in academic weblogs, singles out patterns in types of hyperlink use but

44. Theodor H. Nelson, *Computer Lib / Dream Machines* (Self-published, 1974), DM19.
45. (Background on Peirce here?)
46. (explain rickrolling? Mostly because I want to...)

forgoes a systematic, quantitative, or aggregative approach.[47]

The link and its anchor text are increasingly used as a creative textual device, and writers, reporters, and bloggers are continuing to discover its uses. This may be best exemplified by aggregators and email newsletters, which 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 not to treat a link as an endorsement), 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.

Link analysis and page ranking that incorporates anchor text is difficult to do at scale because it tends to require manual coding; the many cultural and contextual nuances behind a link 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,

47. Mj Luzon, "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, accessed February 20, 2015, `http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000262424900009&site=eds-live`.

we could search for "users who disagreed with this article" and not just "users who linked to this article."[48] 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 is progressing 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 imposes its own set of ontologies, hierarchies and categorization schemes.

Another alternative to the web's form of linkage comes from Ted Nelson, a long-time critic of the web's architecture, whom I will discuss at greater length in the following chapter. As the original hypertext visionary, 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 outward only, quotes you can't follow to their origins, no version management, no rights management."[49] 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

48. Berners-Lee, *Weaving the Web*, FIND PAGE.
49. Nelson, "Ted Nelson's Computer Paradigm, Expressed as One-Liners."

this simple omission has profoundly affected culture and economics, which forms a cornerstone of his 2013 book *Who Owns the Future?*[50]

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 traditional headlines 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 Superheroes."[51] The headline is colloquially known as "click bait," playing to a user's innate curiosity (Atlantic writer Derek Thompson calls it the "curiosity gap")[52] 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

50. 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 August 6, 2014, `http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/`; Lanier, *Who Owns the Future?*

51. **find upworthy ref**.

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

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. For instance, while an article might embed an image, the article itself is then embedded and contained as a search result or single item in a table—in other words, in a single blue link, and sometimes an accompanying image and headline. This 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 immense traffic to half of its previous size.[53] Another major content distributor, the lyrics annotation website Rap 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.[54] 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

53. Nicholas Carlson, "Upworthy Traffic Gets Crushed," Business Insider, February 10, 2014, accessed February 12, 2014, `http://www.businessinsider.com/facebook-changed-how-the-news-feed-works--and-huge-website-upworthy-suddenly-shrank-in-half-2014-2`.

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

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 what made them the dominant information provider that they are today. Prior to PageRank, web crawlers and indexers like Yahoo, HotBot, and AltaVista provided a plethora of options for Internet search (even these, in all their heterogeneity, were seen at the time as a major threat to the open web). But each was based on a traditional, hierarchical classification scheme. In PageRank, Google found a way to embrace the web's disorder; where Yahoo insisted on keeping an organized system, Google relied on links to sort everything out. Clay Shirky argues that this is what allowed Google to surpass Yahoo and become the first truly "Web 2.0" company, asserting that on the web, "ontology is overrated."[55]

Google famously published their initial PageRank algorithm, and once the cat was out of the bag, advertisers and spammers began to exploit it, inserting links not for their usefulness or relation to the text, but to improve their pages' search rankings. 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

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

responsible for a substantial amount of the links created—as well 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 the two:

> 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.[56]

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 turn, 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. If two bloggers like one another's content, they will agree to link to each other on their respective blogs. This happens on the "blogroll," a list of other blogs that a blogger might recommend, usually presented as links in the blog's sidebar. Here the link functions as an act of exchange under the guise of information exchange and mutual social gain. Social media and blogging sites in particular are places of transactional link exchange, with implicit conventions and expectations beneath each link or like.

Moreover, these link exchanges solidify existing networks of bloggers and content creators, perhaps galvanizing the network but at the risk of collapsing into "filter bub-

56. 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, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998), 110, accessed December 12, 2013, `http://dl.acm.org/citation.cfm?id=297805.297827`.

bles." 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 inward, allowing searchers to see only blogs that have overtly linked to one another (blogs which, presumably, have similar views and opinions).[57] Many of these studies lead with the stance or assumption that "linking out" to other publishers represents healthy information flow, while "linking in" to one's own content is nepotistic; but the picture is more complex than this when considering the closely linked networks of blogs that may link out, but only to one another. While the Internet 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.[58]

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.

57. Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia, "Data Portraits: Connecting People of Opposing Views," *arXiv:1311.4658 [cs]* (November 19, 2013), accessed December 3, 2013, `http://arxiv.org/abs/1311.4658`.

58. 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), `http://hdl.handle.net/2027/spo.5680986.0001.001`.

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 (for instance, with a photo that is uploaded to Flickr or Instagram).

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, interacts with it, and works with it. While users increasingly treat the web as a database, there is no single database, but rather 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.

As a result, the archive must 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.",[59] 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. Given the highly networked, context-focused organization of the web, it is an impossible task.

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.

59. Derrida, "Archive Fever," 18.

But 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.[60] 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; as I will discuss in the following chapter, it is this very impulse that has often set back the work of retaining knowledge and history. 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."[61] 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. While their approach is more of a framework than a solution, it rightly discourages 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 centralized and distributed; it aims to bridge traditional taxon-

60. Lisa Gitelman, "Response to "Algorithms, Performativity and Governability"" (New York, NY, May 5, 2013).

61. Bowker and Star, *Sorting Things Out*, 297.

omy 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 comes in complaining of shortness of breath and a fever, a doctor can ask a linked database for all diagnoses of similar patients automatically. 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 multiple ontologies can coexist, they are still trying to compromise the web's inherent disorder.

Archive fever is both a personal and an institutional drive. Google and Facebook store user data (including user-created content) with abandon, inventing new contexts at each turn. Users bookmark, download, pin, and clip online resources, sometimes all at once. Built-in browser solutions like bookmarks and history haven't changed their structure in years, and it shows—they store nothing but the URL. "Bookmark" is a misnomer of a remediated word, as books can't change or disappear overnight, while "history" implies a time machine that the web doesn't have. Personal note-taking and online "snapshot" tools aim to create a sort of personal, annotatable intranet for users that want to filter signal from the noise (see applications like Evernote, Pinterest and Zotero). However, aside from folders and tags, none of these systems provide a useful way to store meaningful associations between these documents.

The associations, trails, and lists sparked by the web add to the possible avenues for research; the myriad interconnections between documents may be more responsible than anything else for the seemingly unprecedented amount of information in the information society. In response to this, as well as the web's innate ephemerality, users store everything.

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.[62] 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, pointing to a need to think deeply about the structure and format of knowledge online, and the best ways to track and distribute it.

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).[63] 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

62. James Trevino and Doerte Doemeland, *Which World Bank reports are widely read?* WPS6851 (The World Bank, May 1, 2014), 1–34, accessed May 10, 2014, `http://documents.worldbank.org/curated/en/2014/05/19456376/world-bank-reports-widely-read-world-bank-reports-widely-read`.

63. Samuel Arbesman, *The Half-life of Facts* (Current Hardcover, September 27, 2012), accessed April 22, 2014, `http://www.goodreads.com/work/best_book/19175842-the-half-life-of-facts-why-everything-we-know-has-an-expiration-date`.

with "channel surfing."[64] As users, we know the phenomenon as the dreaded "404 Not Found" page. 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 (this is "reference rot").[65] 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.[66] Similarly, the NewsDiffs project discovered that the average breaking news story on major news media websites, at a given URL, is edited and updated six times.[67]

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.[68] 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 computer browsing; 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. Google is the primary target of these debates, with controversies like Google Books' battle with libraries and the battle over the "right to be forgot-

64. Barnet, "Pack-rat or Amnesiac?," 217.

65. 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), accessed December 8, 2013, `http://papers.ssrn.com/abstract=2329161`.

66. **http://www.lanl.gov/discover/news-release-archive/2015/January/01.26-scholarly-articles-affected-b**

67. **NewsDiffs; check this figure**.

68. **https://cyber.law.harvard.edu/node/92313**.

ten."[69] Google Books is a particularly interesting case: undoubtedly, one of Google's motivations for digitizing and structuring books was for machine reading and learning. There is much debate surrounding Facebook's role in our privacy and filter bubble, but less in how it is serving as digital history preservation. In a 2014 talk at MIT, Tarleton Gillespie emphasized the responsibilities that companies like Google, Facebook, and Twitter have towards the contemporary public sphere, global information flow, and implicitly, history.[70]

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."[71] 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 wouldnâĂŹt scan every page, but she is happy to save and even re-host an entire database with just a few clicks.[72] This creates an arbitrary number of digital replicas online and explodes the notion of a "canonical" version at a specific URL.[73]

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.

69. (cites for these?)
70. **gillespie_algorithms_????**.
71. **http://www.lockss.org/**.
72. Manoff, "Archive and Database as Metaphor," 386.
73. (figure here with layers of containment/petrification of a document: scanned PDF, image, digital pdf, html doc)

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 the previous chapter's "Defining the archive," I confessed that my use of the term might expand and contract in scope, with my use of the word signifying a token of preservation and access 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 wide in scope, 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,

1. Ann Blair, "Note Taking as an Art of Transmission," ArticleType: research-article / Full publication date: Autumn 2004 / Copyright © 2004 The University of Chicago Press, *Critical Inquiry* 31, no. 1 (September 1, 2004): 85–107, accessed December 8, 2013, `http://www.jstor.org/stable/10.1086/427303`.

bibliographies and encyclopedias that first forayed into networked information.

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

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, accessed December 19, 2014, `http://www.sciencemag.org/content/333/6043/776`.

own internal memory.

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.[3] 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.[4] The third order allows items to be in multiple categories at once, as if in multiple sections of the library—a phenomenon that information architects call *polyhierarchy*.[5]

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

3. Belinda Barnet, *Memory Machines: The Evolution of Hypertext* (London: Anthem Press, July 15, 2013); John Willinsky, *Technologies of Knowing: A Proposal for the Human Sciences* (Boston: Beacon Press, January 1, 1999).

4. David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder*, ıPerfectly placed to tell us what′s really new about [the] second-generation Web.ȷ¯*Los Angeles Times* Business visionary and bestselling author David Weinberger charts how as business, politics, science, and media move online, the rules of the physical world¯in which everything has a place¯are upended. In the digital world, everything has its places, with transformative effects: ffl Information is now a social asset and should be made public, for anyone to link, organize, and make more valuable. ffl There′s no such thing as ıtoo muchȷ information. More information gives people the hooks to find what they need. ffl Messiness is a digital virtue, leading to new ideas, efficiency, and social knowledge. ffl Authorities are less important than buddies. Rather than relying on businesses or reviews for product information, customers trust people like themselves. With the shift to digital music standing as the model for the future in virtually every industry, *Everything Is Miscellaneous* shows how anyone can reap rewards from the rise of digital knowledge. ** (New York, NY: Macmillan, April 2008), 17-23.

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

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.[6] 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 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.[7] Such an exercise emphasizes the temporal

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

7. Peter Burke, *A Social History of Knowledge: From Gutenberg to Diderot*, In this book Peter Burke adopts a socio-cultural approach to examine the changes in the organization of knowledge in Europe from the invention of printing to the publication of the French Encyclopedie. The book opens with an assessment of different sociologies of knowledge from Mannheim to Foucault and beyond, and goes on to discuss intellectuals as a social group and the social institutions (especially universities and academies) which encouraged or discouraged intellectual innovation. Then, in a series of separate chapters, Burke explores the geography, anthropology, politics and economics of knowledge, focusing on the role of cities, academies, states and markets in the process of gathering, classifying, spreading and sometimes concealing information. The final chapters deal with knowledge from the point of view of the individual reader, listener, viewer or consumer, including the problem of the reliability of knowledge discussed so vigorously in the seventeenth century. One of the most original features of this book is its discussion of knowledges in the plural. It centres on printed knowledge, especially academic knowledge, but it treats the history of the knowledge 'explosion' which followed the invention of printing and the discovery of the world beyond Europe as a process of exchange or negotiation between different knowledges, such as male and female, theoretical and practical, high-status and low-status, and European and non-European. Although written primarily as a contribution to social or socio-cultural history, this book will also be of interest to historians of science, sociologists, anthropologists, geographers and others in another age of information explosion. ### Review 'In Peter Burke's scholarly hands the notion of a social history of knowledge sheds its philosophical provocation and becomes judicious, prudent and historically rich. A beautifully written and accessible exercise in historical synthesis.' *Steven Shapin, author of "A Social History of Truth: Civility and Science in Seventeenth-Century England" (1994) and Professor of Sociology, University of California, San Diego* 'Peter Burke is an exceptional historian: a polyglot, at home in a dozen languages; an intellectual, who is well versed in theoretical developments adjacent to history; a superb expositor, with the capacity to distil his findings in unpretentious and limpidly accessible prose; and an author of unflagging vitality, whose prolific studies in the cultural history of early modern Europe and in modern historiography constitute a formidable *oeuvre *... He has succeeded in producing a balanced, judicious and highly stimulating work of synthesis. His book will be an indispensable starting point for years to come.' *Keith Thomas, History Today* 'Burke has made a significant contribution to cultural history ... [He] shows how knowledge was a form of exchange and how it became what we would recognize it as today. Burke's achievement in A Social

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,"[8] 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

History of Knowledge is to remind us that people in the past did not view knowledge in the same way as we do today.' *History* ### From the Back Cover In this book Peter Burke adopts a socio-cultural approach to examine the changes in the organization of knowledge in Europe from the invention of printing to the publication of the French *Encyclopédie*. The book opens with an assessment of different sociologies of knowledge from Mannheim to Foucault and beyond, and goes on to discuss intellectuals as a social group and the social institutions (especially universities and academies) which encouraged or discouraged intellectual innovation. Then, in a series of separate chapters, Burke explores the geography, anthropology, politics and economics of knowledge, focusing on the role of cities, academies, states and markets in the process of gathering, classifying, spreading and sometimes concealing information. The final chapters deal with knowledge from the point of view of the individual reader, listener, viewer or consumer, including the problem of the reliability of knowledge discussed so vigorously in the seventeenth century. One of the most original features of this book is its discussion of knowledges in the plural. It centres on printed knowledge, especially academic knowledge, but it treats the history of the knowledge 'explosion' which followed the invention of printing and the discovery of the world beyond Europe as a process of exchange or negotiation between different knowledges, such as male and female, theoretical and practical, high-status and low-status, and European and non-European. Although written primarily as a contribution to social or socio-cultural history, this book will also be of interest to historians of science, sociologists, anthropologists, geographers and others in another age of information explosion. (Cambridge: Polity, December 2000), 90.

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

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.[9] On the other hand, an alphabetical scheme offers a linear view, one that "flattens" the brittle hierarchy of taxonomy, but dulls its rich network of links, trails, and associations. The linked hypertext view might be seen as a multi-dimensional graph, more nuanced and flexible but more 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."[10] 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 "intellec-

9. (Figure here with historical tree views)

10. Ephraim Chambers, *Cyclopædia, or an Universal Dictionary of Arts and Sciences* (1728), 64, accessed December 18, 2014, `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"," ArticleType: research-article / Full publication date: Jan., 2003 / Copyright © 2003 University of Pennsylvania Press, *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 61–72, accessed December 10, 2013, `http://www.jstor.org/stable/3654296`.

tual entropy" and "epistemological urgency" of the time.[11] 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.[12] 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 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.[13] Chambers called this complaint "as old as Solomon."[14]

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."[15] In order to truly demonstrate a "continued discourse," Chambers needed a graph, a map. Each of the Heads in the diagram contains a footnote that lists that heads' terms (known as Common Places).[16]

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).[17] *Loci* ("places") bring to mind memory palaces, but also the "commonplace book"—to which

11. Daniel Rosenberg, "Early Modern Information Overload," ArticleType: misc / Full publication date: Jan., 2003 / Copyright © 2003 University of Pennsylvania Press, *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 5, accessed December 10, 2013, `http://www.jstor.org/stable/3654292`.

12. 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, accessed December 12, 2013, `http://nms.sagepub.com/content/11/1-2/95`.

13. Ann Blair, "Reading Strategies for Coping With Information Overload ca.1550-1700," *Journal of the History of Ideas* 64, no. 1 (2003): 11–28, accessed December 11, 2013, `http://muse.jhu.edu/content/crossref/journals/journal_of_the_history_of_ideas/v064/64.1blair.html`.

14. Yeo, "A Solution to the Multitude of Books," 11.

15. Ibid., 67.

16. (add more historical diagrams of graphs, trees, maps)

17. Burke, *A Social History of Knowledge: From Gutenberg to Diderot*, 95.

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.[18] 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."[19] 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.[20] 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*."[21]

It is clear that Diderot fully recognized and built upon Chambers' developments in 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."[22] 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

18. Yeo, "A Solution to the Multitude of Books," 65-66.
19. 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".
20. Zimmer, "Renvois of the past, present and future," 103.
21. Ibid., 104.
22. Ibid., 99.

imperfect or incomplete.[23] Leibniz also praised his famous Note Closet, developed by Thomas Harrison, for this same ability: "A single truth can usually be put in different places, according to the various terms it contains... and different matters to which it is relevant."[24]

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.[25] 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."[26] 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, his ideas and innovations

23. **lost this cite**; **who is it? burke?**.
24. Blair, *Too Much to Know*, "Managing Abundant Notes".
25. Burke, *A Social History of Knowledge: From Gutenberg to Diderot*, 94.
26. "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`.

can be broken into three themes. First, he envisioned (and even began to amass) a universal library to serve as the heart and central authority of the world's information. Second, following his belief that books were redundant and arbitrary agglomerations 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.[27] Finally, 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."[28] Otlet's "radiated library" and "televised book" combine to suggest the networked multimedia of the web, more than 50 years before its creation.

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."[29] 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 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."[30] In his discussion of the architecting of knowledge, Otlet also crucially recognized that ideas are never so fixed as physical structures; as Charles van den Heuvel puts it, "For Otlet it was important to leave space for transformation and modification in response to

27. (add figure here with an example?)

28. .

29. 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., February 15, 2008), 129.

30. Ibid., 130.

the unforeseen and unpredictable."[31] 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 thus 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.[32] 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 become 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. I then hone in on Vannevar Bush's memex machine, which wavered between personal study aid and collective knowledge generator.

31. Ibid., 131.

32. "A Limited Company of Useful Knowledge : Paul Otlet, the International Institute of Bibliography and the Limits of Documentalism."

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."[33] 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."[34] 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.

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

33. Blair, "Note Taking as an Art of Transmission."
34. **Borges "The Analytical Language of John Wilkins"**.

difference between the collective search engine and the learned box of paper slips lies in its contingency."[35] Whenever we add a tag or make a connection in an archive, we are attempting to predict what will be searched for later; this is why Derrida calls the archive "a *pledge*, a token of the future."[36] 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 notes in wills, which suggests that any time a commonplace book is begun, it has some kind of common value.[37] 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. Is it a book, an archive, or a database? Who is it for? What happens to individual 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 systems "algorithmically rearticulate the relationship between individual and aggregate traits."[38] The communities and demographics that form around individuals can in turn be aggregated and intersected into a single, massive whole. At each stage, mem-

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

36. Jacques Derrida, "Archive Fever: A Freudian Impression," trans. Eric Prenowitz, ArticleType: research-article / Full publication date: Summer, 1995 / Copyright © 1995 The Johns Hopkins University Press, *Diacritics* 25, no. 2 (July 1, 1995): 18, accessed December 2, 2013, `http://www.jstor.org/stable/465144`.

37. Blair, "Note Taking as an Art of Transmission," 104.

38. Nick Seaver, "Algorithmic Recommendations and Synaptic Functions," *Limn*, no. 2 (2012): 44–47, accessed December 14, 2014, `http://limn.it/algorithmic-recommendations-and-synaptic-functions/`.

ory 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.[39] 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, individual memory, collective history

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.[40] Perhaps most crucially, Bush conceived of new ways to 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."[41]

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

40. Vannevar Bush, "As We May Think," *The Atlantic* (July 1945), accessed December 11, 2013, `http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/`.

41. Belinda Barnet, "The Technical Evolution of Vannevar Bush′s Memex," *Digital Humanities*

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."[42] 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."[43]

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).[44] 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."[45]

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."[46] This now seems prophetic, a prediction of contemporary culture's emphasis 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."[47] Bush does not dwell on this to consider where this "common record"

Quarterly 2, no. 1 (2008), accessed December 16, 2013, `http://www.digitalhumanities.org/dhq/vol/2/1/000015/000015.html`.

42. Bush, "As We May Think."
43. Ibid.
44. Barnet, "The Technical Evolution of Vannevar Bush′s Memex."
45. Bush, "As We May Think."
46. Ibid.
47. Joseph Reagle, *Good Faith Collaboration: The Culture of Wikipedia* (Cambridge, Mass.: MIT

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.[48] 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 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."[49] 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.[50] It appears to me that the

Press, 2010), 23, `http://mitpress-ebooks.mit.edu/product/good-faith-collaboration`.

48. 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, accessed December 15, 2013, `http://doi.acm.org/10.1145/800197.806036`.

49. 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, accessed December 19, 2014, `http://dl.acm.org/citation.cfm?id=132180.132193`.

50. James M. Nyce and Paul Katin, "Innovation, Pragmaticism, and Technological Continuity:

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.

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."[51] 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.[52] While later hypertext systems and the Semantic Web come closer to such a 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

Vannevar Bush's Memex," *Journal of the American Society for Information Science* 40, no. 3 (May 1989): 217, accessed December 16, 2014, `http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16801186&site=eds-live`.

51. Barnet, "The Technical Evolution of Vannevar Bush′s Memex."

52. Bush, "Memex Revisited," 209.

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 all of 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."[53] 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 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.[54]

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."[55] 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.[56] 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

53. Reagle, *Good Faith Collaboration: The Culture of Wikipedia*, 20.
54. Jorge Luis Borges, *Collected Fictions*, trans. Andrew Hurley (New York, N.Y., U.S.A.: Penguin Books, September 1, 1999), 112-118.
55. Blair, *Too Much to Know*, "Information Management in Comparative Perspective".
56. Burke, *A Social History of Knowledge: From Gutenberg to Diderot*, 93.

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."[57] 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,"[58] which is reminiscent of Google's: "to organize the world's information and make it universally accessible and useful."[59] 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 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."[60]

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, Amazon—may have come closer than any previous efforts to capturing the world's knowledge, but they does 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

57. Blair, *Too Much to Know*, "Bibliographies".
58. Reagle, *Good Faith Collaboration: The Culture of Wikipedia*, 18.
59. **"Company Overview**; **" Google**; **http://google.com/about/company**.
60. Reagle, *Good Faith Collaboration: The Culture of Wikipedia*, 17.

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."[61] 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.

The result would be, as Barnet says, "like the web, but much better."[62] 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 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.[63] 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.[64]

61. Theodor H. Nelson, *Computer Lib / Dream Machines* (Self-published, 1974), 313.
62. Barnet, *Memory Machines*, "The Magical Place of Literary Memory: Xanadu".
63. Reagle, *Good Faith Collaboration: The Culture of Wikipedia*, 11-12.
64. Jaron Lanier, *Who Owns the Future?* (New York, NY: Simon & Schuster, May 7, 2013), 221-232.

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. In the end, we cannot easily filter or measure content on the web, and we need to rely on search and indexing services like Google to do it for us. One blog post by the Tow Center's NewsLynx project laments, "the inefficiency of one-way links left a hole at the center of the web for a powerful player to step in and play librarian."[65]

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.[66] 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.[67]

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

65. 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 August 6, 2014, `http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/`.

66. Gary Wolf, "The Curse of Xanadu," *Wired* 3, no. 6 (June 1995), accessed December 15, 2013, `http://www.wired.com/wired/archive/3.06/xanadu_pr.html`.

67. Jeff Atwood, "The Xanadu Dream," Coding Horror, October 12, 2009, 221-232, accessed June 6, 2014, `http://blog.codinghorror.com/the-xanadu-dream/`; Lanier, *Who Owns the Future?*

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 on the web, after more than 50 years of development, in summer 2014.[68] 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).[69] 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

68. "Pioneering hypertext project Xanadu released after 54 years," kottke.org, accessed September 8, 2014, `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 September 8, 2014, `http://www.theguardian.com/technology/2014/jun/06/vapourware-software-54-years-xanadu-ted-nelson-chapman`.

69. Frank Halasz, "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems," *Commun. ACM* 31, no. 7 (July 1988): 836–852, accessed December 8, 2013, `http://doi.acm.org/10.1145/48511.48514`.

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 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."[70] 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

70. Kate Crawford, "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society* 15, no. 5 (June 2012): 662–679, accessed September 5, 2013, `http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878`.

respective archive's size, shape, and scope.

Chapter 4

Networking the News

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

In the publishing world, 2014 saw 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. In the digital-publishing sphere, some new publishers began to champion an "explainer" model of journalism; with headlines like "Everything you need to know about the government shutdown" or "40 charts that explain money in politics," the explainer model suggests that newsrooms can be "as good at explaining the world as it is at reporting on it."[2] Meanwhile, legacy publishers have led a new focus on renewing and reanimating their historical archives; whether they're cherry-picking old curiosities and republishing them, provid-

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 December 23, 2014, `http://www.theverge.com/2014/1/26/5348212/ezra-klein-vox-is-our-next`.

ing subscribers with an interface to dive in and leaf through, or leading 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 rich sense of their brand's history, and a new take providing context and resurfacing the past.

These two trends reflect a seismic shift in the 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; as journalism on the web serves a different function than its pre-online counterpart, publishers need to recognize that their commercial product and public service has fundamentally changed. Publishers have a newfound opportunity to thrive in the information industry as much as the news or publishing industry, but in order to compete in an information landscape 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.

As such, this chapter is divided into two sections, one cultural and the other technical. The cultural section aims to outline the telling origins of archive-oriented and explainer journalism, emphasizing that the rapid proliferation of new content and connections have fundamentally changed journalism's roles and practices, as they downplay their role as news breakers and scoopers, instead amplifying their value in archiving, explaining, and contextualizing. The second section will begin by analyzing in detail the architecture of a digital news story, then I will suggest practical solutions, tools, and technologies 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."[3] 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 the National Geographic are held up as examples of combining rich content with historical context, improving credibility and brand recognition in the process.

The Times closely examined its own archives in their celebrated *Innovation* report of 2014, suggesting that a clever use of archives could revitalize new content by seamlessly integrating with historical context. "Our rich archive offers one of our clearest advantages over new competitors. . . [b]ut we rarely think to mine our archive, largely because we are so focused on news and new features," arguing that "we can be both a daily newsletter and a library."[4] The report suggests that arts and culture content, more likely to be evergreen, could be organized "more by relevance than by publication date," and that topic homepages should be more like guides than wires.[5] The report goes on to enumerate successful experiments with repackaging old content in collections, organized by categories and themes. They suggest allowing users to create their collections of stories—something that readers could also do without risk to the Times brand. By creating "no new articles, only new packaging," the Times can easily give new life to old stories.[6]

In 2014 we also saw another trend towards "explainer journalism," and an intense focus on context provision for readers. Vox.com, the poster child for the explainer movement, wants "to create a site that's as good at explaining the world as it is at reporting on it."[7] Explainer journalism aims to take a step back from the immediate news event and place it in a larger phenomenon, and it reflects a deep shift in the

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

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

5. Ibid., 29-30.

6. Ibid., 34.

7. Klein, "Vox is our next."

roles and practices of online journalists; as news is increasingly broken and scooped on social media, journalists are increasingly becoming summarizers, filterers, and context providers. News has traditionally been delivered in a stream format, full of boilerplate text that is repeated across every story related to a given theme. In the archive and explainer movements, we see a pattern among some news outlets attempting to evade and reconsider the news cycle's obsession with speed and feeds, instead experimenting with new forms of what a news story can be, and how it can connect to other stories within the archive and around the web.

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

4.2 The structure of stories

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

endless potential for insightful metadata; newspapers might seem to naturally lend themselves to the digital world.

The pre-hypertextual newspaper started as a response to a sort of historical information overload; the newspaper frontpage and summary lead paragraph, both solidified in 1870, were part of a broader trend towards "helping readers to economize their scarce time in scanning a paper."[8] Larger type, illustrations, and bolder headlines drew criticism for trying to grab attention, but they also directed and focused attention to the major stories of the day, allowing for nonlinear readings of a newspaper as a fragmented collection of stories, headlines, or leads. A newspaper's layout and seriality therefore scaffold a pseudo-hypertextual structure, one that can be computationally mined for insights. Some libraries and cultural heritage institutions are leading these endeavors, such as the Library of Congress' National Digital Newspaper Program, Europeana, and Trove.[9]

But traditional newspapers have a major limitation: they cannot *explicitly* link to other work in a structured and idiomatic way. Scholars have long relied on the footnote and bibliography to systematically track influence and dialogue, and networks of citations can be created out of them. This forms the basis for citation analysis, or bibliometry, a practice with a long history and strong conventions that I will dive into more closely in the following chapter. Its essential principle is that the more an item is cited, the more influential and credible it is. The online version is known as "webometrics," and it applies certain new standards which online newspapers can take advantage of, both in measuring impact on the web and inside their own archives. But citation is "as old as written language itself," and it is *itself* a language, with its own idioms, syntaxes and exceptions.[10] The footnote has its limitations, only linking back to the past—but newspapers don't even get footnotes.

The journalistic affordances that the web brings can be conceptually divided into a few core features, which Mark Deuze outlines as hypertextuality, multimediality,

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

9. **europeana**; **trove**.

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

and interactivity.[11] Peter Dahlgren adds a fourth and fifth: for him, media logic is also *figurational* and *archival*. Discussing archivality, he asserts that "users of cyberspace for journalism are in principle no longer so bound to the present," and points to hypertextuality as enabling new, more usable archives.[12] While many projects have closely mapped the role of hypertextuality in online journalism—examining newsrooms' approaches towards hyperlinking through network analysis, surveys, interviews, and newsroom ethnography—there has been less research considering hypertextuality's influence on newsrooms' archival practices.

Links—and their siblings, linked tags—allow for a new standard of citation, reference, and context provision for news. The link can even go beyond the footnote by linking in both directions, allowing readers to see who referenced the story; an old article in *The New York Times*, for instance, can link out to more recent related Times articles, other publishers or blogs that picked up on the story, or conversations in the Times forum or on Twitter. Linking offers great potential, not only for enlivening the reading experience, but for creating a traceable dialogue that can improve a story's discoverability in the future. A number of search algorithms, such as Google's PageRank and Jon Kleinberg's HITS system, create "hyperlink-induced communities" between websites, and the same principles can be adopted and expanded within news websites.[13]

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

11. 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`.

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

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

tagging *the story* instead of its constituent parts, we can turn to entities and links as supplements to categories and tags in the newsroom archive. These play to a journalist's strengths, and augment rather than replace the human touch that comes with inline links and curated collections.

4.2.1 Atoms of news

In technical terms, stories are usually objects in a database that have associated text, images and tags. Stories contain multitudes, and a typical story might have a variety of metadata attached to it; authors, dates, versions, categories, images, events and collections it's a part of, tags, and so on.[14] While more metadata and structure requires more investment at the outset, smart use of such metadata prepares a story for archival reuse. Some of it will be useful for linking or embedding on the website, others for use in an API or application. Stories can include other stories as part of their metadata too, either related manually (by a hyperlink or a human editor) or automatically (via similarity algorithms that analyze the words or topics in the article, the communities it reaches, and so on).

The story has long been the basic unit of news, and so it tends to have a one-to-one relationship with the URL, the basic unit of the web. One section of the *Times' Innovation* report announces that they produce "more than 300 URLs a day," using URL as a sort of "thing" word, their default unit of work.[15] 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.[16] You can be sure to find the most canonical version at the canonical URL, but the article lives in various forms across the web.

But if stories contain multitudes, then why is the article the basic unit of infor-

14. Figure here with typical story objects; a sample database schema
15. *Innovation*, 27.
16. 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.

mation 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, 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.

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.[17] 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."[18]

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

Shirky and David Weinberger champion the tag 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.[19] These systems, sometimes called "folksonomies," emerge as manifestations of the "boundary infrastructures" proposed by Bowker and Star.[20]

Tags have their limitations; if another user tags an item "World War 2," the system needs to recognize that it means the same thing as "World War II," and publishers employ controlled vocabularies to avoid such ambiguities. Some research has shown that the first tags on an item are likely to influence future tags in turn, resulting in a sort of ontological groupthink.[21] Still, whether a Flickr tag, a Delicious bookmark, or a Twitter hashtag, these crowdsourced approaches to tagging function as links between content; it is not about the tag itself, but the *connection* being made to other content. Shirky even considers the possibility that syonymous terms aren't desirable; perhaps the people who are searching for "films" would be better served by just seeing results tagged as "films," and not "movies" or "cinema."[22] 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 New York Times sees tagging as core to its business, and the reason it has remained the "paper of record" for so long.[23] This title was bestowed to them largely on the back of the legacy Times Index, which has offered an annual reference version of Times stories since 1913, still published in hard copy. There is no doubt that the

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

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

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

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

23. *Innovation*, 41.

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; it took seven years for the Times to start tagging stories "September 11." Their team of librarians is required to shepherd about 300 new articles a day into the archive, making sure to keep them discoverable under any context. Tags can concern more than just the contents or event of a story—the Report suggests tagging stories by timeliness, story tone, and larger "story threads"—but they admit that many of the more exciting tagging potentials would require them to "better organize our archives."[24]

So the linking of the archive can occur not only through explicit hyperlinks, but implicit tags and entities that reside within the stories themselves. Most news articles rely on tagging to be connected to other media; if *The Boston Globe* writes a story about New England Patriots coach Bill Belichick, an editor might tag the story "football" in order to place it in dialogue with other football stories, landing on the *Globe*'s football topic page, and so on. But this belies a more nuanced dialogue between the words, images, and hyperlinks used within the story itself; a short story that has "Bill Belichick" and "Cincinnati Bengals" in the text is likely to be referencing a recent game or a trade, while a longer story that brings up his family members or his hometown is likely to be a biographical piece about his life and upbringing. By combining natural language processing and entity linking tools, a story can be automatically and dynamically tagged according to the myriad possible contexts that a user might search for, whether she is looking for stories about football, the Patriots, or Bill Belichick himself.[25]

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

4.2.2 From tags to links

While tagging practices can be enhanced in a variety of ways, Stijn Debrouwere thinks that even "tags don't cut it."[26] 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 opnions that carry value in their own right.[27]

Debrouwere cites a blog post by Adrian Holovaty, co-creator of the Django web framework and its now-retired "benevolent dictator for life." Holovaty's essay revolves around one idea: "newspapers need to stop the story-centric worldview."[28] Each story, he notes, contains a vast amount of structure that is being thrown away with every click of the "publish" button. He leads through several examples, such as:

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

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

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

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

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

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

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

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

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

Journalists as well as readers would gain improved access to their organization's history, improving the research, context, and tagging of future stories. Debrouwere is suggesting a return of structure to the open web; he envisions a tagging system where a tag can double as a card or widget, linked in turn to other cards and widgets in a network of knowledge. This could be extended to events and phenomena as well as proper names and entities; some emerging systems can recognize phrases like "Barack Obama announced his candidacy for president" and ground it as as a unique, unambiguous newsworthy event.[29] While this research has a long way to go, it is a promising start towards extending the "ankle-deep semantics" that Chakrabarti advocates.[30]

This return of structure does not have to be a step backwards to broad ontologies and topics; instead of manually and unilaterally structuring from on high, we can focus on the structure built into the stories already. Rather than replacing stories entirely, or requiring editors to exhaustively tag every component of a piece, a system could automatically supplement a story with new metadata that gives its inherent information an afterlife. Every story—whether a blog post, a map, a listicle, or an interview—has its own structures and patterns.

One example comes from the Boston Globe's March 2015 coverage of the Boston Marathon bombing trial. Data editor Laura Amico knew well that trials are far from linear stories; since they are told by lawyers, they 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 court witnesses, documents, and so on. The key players—witnesses, evidence, victims and lawyers—would have a long history of Globe coverage from the bombing itself. Amico and her team knew there was more than one way to tell this story, so they decided to focus on the facts; the witnesses, exhibits, and arguments are all entered into a spreadsheet, which can generate snippets of facts and entities—

29. Joel Nothman, "Grounding event references in news" (August 31, 2013), accessed February 8, 2015, `http://ses.library.usyd.edu.au:80/handle/2123/10609`.
30. Chakrabarti, *Mining the Web*, 289.

designed as cards—for later review.[31] Each of these cards can embed and contain other cards, or be combined to form a story. Such a framework recognizes the interlinked nature events, and the stories that summarize and filter them. This project also highlights that "structured journalism" does not need to be adopted wholesale; it can work for one-off stories and experiments as well.

One of the most obvious and underexplored structures is the hyperlink. A year after publishing his "Information architecture" series, Debrouwere followed up by questioning many of his own allegiances; tags still don't cut it, but maybe taxonomies don't either. He realized: "The best content recommendations on news websites are inside of the body copy: inline links. With recommendations, you never know what you're getting. It's mystery meat. With links, a writer tells you why she's pointing at something the moment she's pointing at it."[32] It is better to draw on the connections from these links than to rely on automated recommendation engines to organize content. Journalists are better at explaining and contextualizing than they are at tagging and structuring, which are a librarian's craft. Debrouwere knows that newsroom developers are building for journalists, and he ends by asserting that he wants to build "prosthetics, not machines."

Another under-mined source of insight lies in the plethora of "human-generated lists" (as Google calls them in one patent) around the web.[33] 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

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

32. **debrouwere_taxonomies_2011**.

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

is linked by country and location, while a list of my favorite hip-hop albums of 2014 is linked by year, quality, and musical genre. The Times' *Innovation* report suggests allowing users to create lists, since it could allow for deep interactivity without risk to their brand; such a system could leverage readers' collective wisdom by asking users to specify the context behind their lists.

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.[34] This could have described a pre-web newsroom, but no longer; the publisher's corpus has expanded and linked beyond measure, and its ranks of expert catalogers are rapidly dwindling. This suggests a need to adopt new schemes, which leverage automatic and dynamic tagging, linked entites, image recognition, and the knowledge of experts and crowds.

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

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

4.3 Stages of digital history

In 1997, John Pavlik suggested that there were three stages of development in online versions of newspapers. Keeping in mind this early date, Pavlik observed that newspapers' first stage online was to copy the print edition to a given website (a stage known and maligned as "shovelware"), followed by supplementing the copy with interactive features (like hyperlinks or comments), then in the third and final stage, writing copy specifically for the online version of the story.[35] Twenty years later, it is safe to say that publishers have all reached the final stage. I aim to suggest a three-stage development of my own in the state of the digital legacy news archive, referring to each stage respectively as *digitizing*, *atomizing*, and *linking*.

The first stage for any legacy publisher is to ***digitize*** the archive. This tends to consist of scanning the pages of old publications, running OCR (optical character recognition) on each page, and exposing the results to a search interface for researchers and, perhaps, interested readers. It is a crucial first step for enlivening the archive, but a physical record can often limit the digital equivalent's potentials. Digital versions of physical articles often do not leverage links, images, and mixed media to the same effect. While a digital-native version of a print article might directly cite more sources or feature an intriguing interactive, these elements remain second-class citizens to the print article, which digital versions must remain faithful to. The Times' *Innovation* Report argues that by modeling their website and apps on their print structure, the Times "ask[s] too much of readers."[36] So it is crucial to remember at this stage that the digital archive has different potential from its physical counterpart. As many theorists and historians remind us, too, a paper's physical appearance and content are closely linked together, so simply "digitizing" and newspaper changes it massively, reshaping a great deal of context.[37] Richard Abel breaks down the promises and

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

36. *Innovation*, 26.

37. James Mussell, "Elemental Forms," *Media History* 20, no. 1 (January 2014): 388-389; Marlene Manoff, "Archive and Database as Metaphor: Theorizing the Historical Record," *portal: Libraries and the Academy* 10, no. 4 (2010): 385–398, accessed December 2, 2013, `http://muse.jhu.edu`.

challenges in generating archival "big data" in research on 1910 US cinema. Using digital newspapers led to "a wealth of unexpected documents," but he notes the unreliability of completeness and searchability, and the collapse of community.[38]

Given the print newspaper's proto-hypertextual status, it presents a unique metadata challenge for archivists. Paul Gooding, a researcher at University College London, sees digitized newspapers as ripe for analysis due to their irregular size and their seriality.[39] In order to learn more about how people use digitized newspaper archives, Gooding analyzed user web logs from Welsh Newspapers Online, a newspaper portal maintained by the National Library of Wales, hoping to gain insight from users' behavior. He found that most researchers were not closely reading the newspapers page by page, but instead searching and browsing at a high level before diving into particular pages. He sees this behavior as an accelerated version of the way people browse through physical archives—when faced with boxes of archived newspapers, most researchers do not flip through pages, but instead skip through reams of them before delving in. So while digital newspapers do not replace the physical archive, they do mostly mimic the physical experience of diving into an archive. Still, something is lost when the physical copy becomes digital; the grain of history—the old rip, annotation, or coffee stain—is reduced to information.

It might go without saying, but it is also crucial to back up the archive, in a variety of formats. More and more evidence suggests that digital content could have a shorter shelf-life than tapes, film, or other analog media.[40] The Missouri School of Journalism's *Columbia Missourian* lost 15 years of stories and seven years of images in

---

libproxy.mit.edu/login?auth=0&type=summary&url=/journals/portal_libraries_and_the_academy/v010/10.4.manoff.html.

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

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

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

a single server crash.[41] For Missouri's Reynolds Journalism Institute, "News archives serve as a form of institutional knowledge allowing newer staff members to understand and convey the historical context of their stories to the readership. It is difficult to calculate the full value of news archives given the countless hours of reporting and editing they distill, not to mention the treasure they represent in terms of their community's cultural heritage."[42]

In response to the server crash, the Reynolds Institute led a survey of 476 news websites, finding that found that 88–93 percent of them highly value their archives, but about a quarter of them had lost significant portions of their archive due to technical failure. Without these full archives, as Tom Warhover says, "You can't offer up a comprehensive product to sell—your archives—if they aren't complete. You can't be sure you've really vetted a candidate for school board or city council. You can't find those historical pieces from events that now are historic and worth reporting again on anniversaries."[43] Those publishers who have digitized their archives have already made the pledge to preserve and organize, and should be careful to protect their archive from obsolescence.

News archives form a symbiotic relationship with data-driven news apps, whether they exist for a specific event (think World Cup coverage or election results), or exist as standalone platforms (such as Homicide Watch, Syria Deeply, or Timelines). Such apps bypass the URL, but in the process they have their own challenges when saving and archiving. Scott Klein, luminary of news apps at the *Times*, brings up Adrian Holovaty's ChicagoCrime.org, which Holovaty described as "one of the original map mashups."[44] 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

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

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

43. Ibid.

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

work, process, and context behind it.

There's no doubt that software constantly races against obsolescence, and some apps must be retired when their event passes or their function is done. But the lost process and context is lost knowledge. In March 2014, a group of NICAR conference attendees gathered at the Newseum to brainstorm the challenges and potentials of preserving news apps, suggesting more collaboration with libraries, museums, and cultural heritage institutions.[45] Some such institutions are offering novel ways of preserving and maintaining digital work for the future. At the Cooper–Hewitt Museum, a team led by Seb Chan has been preserving the previously for-profit Planetary app as "a living object." For the Cooper–Hewitt, preserving an app is more like running a zoo than a museum: "open sourcing the code is akin to a panda breeding program."[46] 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.

Some publishers have thrown up their hands altogether, relying on third-party services to organize and provide access to their own archives.[47] Reporters might suddenly see old stories disappeared, locked away behind services like LexisNexis. Such services provide fast and effective text search at low cost; but at what cost to an organization's brand, legal rights, and sense of history? A digital story is not just text, and increasingly, an archive includes images, videos, charts, maps, interactives, facts, statistics, quotations, comments, and annotations. Newer forms of classification can take a more holistic view of media, allowing a researcher to browse through text, image, sound, and video alike, and minimize the language limitations of search.

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

46. 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/`.

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

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.[48]

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

Newspapers are rich archival documents, because they store both ephemera and history. Journalists sometimes divide these types of news into "stock" and "flow"; the constant stream of information built for *right now*, versus the durable, evergreen stuff, built to stand the test of time.[49] Newspapers also have 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 archivist's perspective, everything needs to be stored.[50] But historians might treat or navigate through ephemera differently, and each set of documents could have its own metadata or interface as a result. Casual browsers, rather than hardened researchers, could be after a still different history requiring a different interface.

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

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

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

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

The project of atomizing the archive should take advantage of the signals built into newspapers in the first place, accumulating metadata from its size, shape, and context. An atomized archive should also provide a solid interface for viewing the original in its context. When legacy news publishers refer to a "linked" record in a database, they are referring to this ability to click on it and see the original, scanned source page, usually as PDF. Some publishers do not even have linked records for their entire archive, which makes context difficult to grasp for interested researchers.

It is telling that 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. Nicole Maurantonio criticizes old newspapers for ignoring the visual in favor of text, "propelling scholars down a misguided path."[51] Keith Greenwood finds 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.[52]

Historical images are one of the greatest potential sources of engagement and revenue for news archives, and it would be relatively easy for some news archives to sell old photographs with historic value.[53] Some metadata projects in the publishing world are aiming specifically at images, like the New York Times' *Madison* project, which hopes to crowdsource insight about 1950s *Times* advertisements.[54] Outside the publishing sphere, Kalev Leetaru took an image-centric approach to the Internet Archive. The Internet Archive's OCR software threw out images, and Leetaru's would save whatever it threw out as an image file. He has since put 2.6 million of these

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

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

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

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

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."[55] 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.[56]

The final stage is to ***link*** the archive, which 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. But linking the archive requires learning from the explicit and implicit references that already reside in the stories. This combines manual and automatic means, supplementing dedicated search in the surfacing of archival content, and using a "push" rather than "pull" method for finding archival materials. A user doesn't always know exactly what he or she wants, and a linked archive can work with a user to surface it. If the archive doesn't have the resource a user needs, could it at least point the user in the right direction? Could it interface with other knowledge bases to retrieve the answer?

The linked archive borrows from, but is distinct from the notion 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."[57] Building on this idea, Charlie Beckett posits that linking between sources leads to editorial diversity, connectivity and interactivity, and relevance.[58] A linked archive turns the conversation inward—as

55. 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`.

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

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

58. 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/`.

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

It is unhelpful to have a massive, borderless archive, but linked archives can expand their borders strategically through clever use of APIs. As Anne Helmond's "Boundaries of a website" and the Open Knowledge Foundation's "The News Reads Us" project remind us, publishing websites rarely operate alone; they rely on a plethora of third-party platforms and services for analytics, sharing, commenting, and recommendations.[59] One could similarly integrate with APIs that offer archival resources from around the web. If a user is searching a publisher's website instead of Google's, it's because she wants more context than a mere list or index of items. She wants less containment and more connection. A user should be able to see response articles, comments, tweets, timelines, images and videos, from around the web (as long as these are visually separate from the main content to avoid confusion). Otherwise, users will continue to go to Google and Wikipedia for information.

A publisher's archival search interface could include results from Google, Wikipedia, Creative Commons images from Flickr, or resources from digital libraries like Europaeana and the Digital Public Library of America—not to mention partnering with other organizations to merge archives or indices. These could be different, and differently useful, for reporters, researchers, and readers alike. 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.

59. Anne Helmond, "Exploring the Boundaries of a Website. Using the Internet Archive to Study Historical Web Ecologies" (MIT8, Cambridge, MA, 2013), accessed December 15, 2013, `http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-using-the-internet-archive-to-study-historical-web-ecologies/`; Stefan Wehrmeyer, Annabel Church, and Friedrich Lindenberg, "The News Reads Us," Open Knowledge Foundation, accessed March 9, 2015, `https://github.com/okfde/the-news-reads-us`.

4.4 Context in context

Structuring and linking the news cannot happen through technology alone. One reason that newspapers lacked a citation convention before the web is because journalistic practice didn't require it; while scholars read and cite articles, journalists traditionally go into the field and gather quotes. These are not so easily cited; a well-organized journalist could keep recordings of sources and, for instance, link to a snippet of audio when adding a quote, but this isn't always possible. Journalists may also be reticent to link because they feel they don't need to; sometimes journalists need sources to remain anonymous, obfuscated, or off the record. Most journalists agree that linking manifests many core tenets of journalism, but forcing a journalist to obsessively link to every source would stifle journalistic practice.

Still, journalists are increasingly citing sources that are already published, whether documents on DocumentCloud, data points from an open government database, or aggregated information from other news articles. New journalistic practices thus lend themselves well to structuring and linking to sources to position a story; the challenge is, in part, in shifting journalistic focus. A journalist might be more apt to say that a court hearing occurred "last week," but it is more archivally valuable to say it happened on "December 12, 2014." This doesn't necessarily require changing the prose in the story itself, but a semi-automated process could occur when publishing a story that suggests these structures behind the scenes.

Named Entity Recognition and topic modeling tools can't do all of the work for us; there are too many missed signals and false hits. But a quick widget or plugin that works *with* a reporter or editor to structure a story could prove fruitful. The challenge remains to convince editors of the value of this, and spend the extra minute verifying a story's structure.

Chapter 5

Tracing the Links

5.1 History/theory of link analysis

5.2 Link breakdown by category

5.3 Link breakdown by graph/network

The last chapter dealt with how publishers think about their archives. This section focuses instead on how they're *linking*, through research and quantitative link analysis. It will combine historic news network analyses and a study that I performed with the help of Media Cloud.

Given these potentials, it might be surprising to find that many publishers are very reticent to link. Those who do have linking policies are often quite conservative.

Some of this is due to SEO; while no one knows exactly what Google's mercurial PageRank algorithm is doing, it's clear that links form a fundamental component (and as critics such as Clay Shirky have argued, relying on links rather than traditional categories and tags has been the crux of their success over competitors).[1] Publishers are also wary of taking a reader away from their own website. But I'm also sure that much of the fear of links comes from inertia and tradition; since journalists never used to have a way to link, some don't see a need to start.

1. .

While many have debated the potentials and pitfalls of hyperlinking the news, I am proposing an additional wrinkle to the conversation; a smart use of linking can be, to borrow Derrida's term, a *pledge* to better structure the news and keep archives continuously animated and relevant.

Chapter 6

Conclusion

6.1 News apps

6.2 HTML5, multimedia, annotation

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. Accessed March 9, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=88234573&site=ehost-live.

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 August 6, 2014. http://towcenter.org/blog/hyper-compensation-ted-nelson-and-the-impact-of-journalism/.

Arbesman, Samuel. *The Half-life of Facts*. Current Hardcover, September 27, 2012. Accessed April 22, 2014. http://www.goodreads.com/work/best_book/19175842-the-half-life-of-facts-why-everything-we-know-has-an-expiration-date.

Atwood, Jeff. "The Xanadu Dream." Coding Horror. October 12, 2009. Accessed June 6, 2014. http://blog.codinghorror.com/the-xanadu-dream/.

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. Accessed December 10, 2013. http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=4645639&site=ehost-live.

———. "The Technical Evolution of Vannevar Bush's Memex." *Digital Humanities Quarterly* 2, no. 1 (2008). Accessed December 16, 2013. http://www.digitalhumanities.org/dhq/vol/2/1/000015/000015.html.

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

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

Berners-Lee, Tim. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. HarperBusiness, November 7, 2000.

Blair, Ann. "Note Taking as an Art of Transmission." ARTICLETYPE: research-article / Full publication date: Autumn 2004 / Copyright © 2004 The University of Chicago Press, *Critical Inquiry* 31, no. 1 (September 1, 2004): 85–107. Accessed December 8, 2013. http://www.jstor.org/stable/10.1086/427303.

———. "Reading Strategies for Coping With Information Overload ca.1550-1700." *Journal of the History of Ideas* 64, no. 1 (2003): 11–28. Accessed December 11, 2013. http://muse.jhu.edu/content/crossref/journals/journal_of_the_history_of_ideas/v064/64.1blair.html.

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

Borges, Jorge Luis. *Collected Fictions*. Translated by Andrew Hurley. New York, N.Y., U.S.A.: Penguin Books, September 1, 1999.

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

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, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998. Accessed December 12, 2013. http://dl.acm.org/citation.cfm?id=297805.297827.

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.* In this book Peter Burke adopts a socio-cultural approach to examine the changes in the organization of knowledge in Europe from the invention of printing to the publication of the French Encyclopedie. The book opens with an assessment of different sociologies of knowledge from Mannheim to Foucault and beyond, and goes on to discuss intellectuals as a social group and the social institutions (especially universities and academies) which encouraged or discouraged intellectual innovation. Then, in a series of separate chapters, Burke explores the geography, anthropology, politics and economics of knowledge, focusing on the role of cities, academies, states and markets in the process of gathering, classifying, spreading and sometimes concealing information. The final chapters deal with knowledge from the point of view of the individual reader, listener, viewer or consumer, including the problem of the reliability of knowledge discussed so vigorously in the seventeenth century. One of the most original features of this book is its discussion of knowledges in the plural. It centres on printed knowledge, especially academic knowledge, but it treats the history of the knowledge 'explosion' which followed the invention of printing and the discovery of the world beyond Europe as a process of exchange or negotiation between different knowledges, such as male and female, theoretical and practical, high-status and low-status, and European and non-European. Although written primarily as a contribution to social or socio-cultural history, this book will also be of interest to historians of science, sociologists, anthropologists, geographers and others in another age of information explosion. ### Review 'In Peter Burke's scholarly hands the notion of a social history of knowledge sheds its philosophical provocation and becomes judicious, prudent and historically rich. A beautifully written and accessible exercise in historical synthesis.' *Steven Shapin, author of "A Social History of Truth: Civility and Science in Seventeenth-Century England" (1994) and Professor of Sociology, University of California, San Diego* 'Peter Burke is an exceptional historian: a polyglot, at home in a dozen languages; an intellectual, who is well versed in theoretical developments adjacent to history; a superb expositor, with the capacity to distil his findings in unpretentious and limpidly accessible prose; and an author of unflagging vitality, whose prolific studies in the cultural history of early modern Europe and in modern historiography constitute a formidable *oeuvre *... He has succeeded in producing a balanced, judicious and highly stimulating work of synthesis. His book will be an indispensable starting point for years to come.' *Keith Thomas, History Today* 'Burke has made a significant contribution to cultural history ... [He] shows how knowledge was a form of exchange and how it became what we would recognize it as today. Burke's achievement in A Social History of Knowledge is to remind us that people in the past did not view knowledge in the same way as we do today.' *History* ### From the Back Cover In this book Peter Burke adopts a socio-cultural approach to examine the changes in the organization of knowledge in Europe from the invention of printing to the publication of the French *Encyclopédie*. The book opens with an assessment of different sociologies of knowledge from Mannheim to Foucault and beyond, and goes on to discuss intellectuals as a social group and

the social institutions (especially universities and academies) which encouraged or discouraged intellectual innovation. Then, in a series of separate chapters, Burke explores the geography, anthropology, politics and economics of knowledge, focusing on the role of cities, academies, states and markets in the process of gathering, classifying, spreading and sometimes concealing information. The final chapters deal with knowledge from the point of view of the individual reader, listener, viewer or consumer, including the problem of the reliability of knowledge discussed so vigorously in the seventeenth century. One of the most original features of this book is its discussion of knowledges in the plural. It centres on printed knowledge, especially academic knowledge, but it treats the history of the knowledge 'explosion' which followed the invention of printing and the discovery of the world beyond Europe as a process of exchange or negotiation between different knowledges, such as male and female, theoretical and practical, high-status and low-status, and European and non-European. Although written primarily as a contribution to social or socio-cultural history, this book will also be of interest to historians of science, sociologists, anthropologists, geographers and others in another age of information explosion. Cambridge: Polity, December 2000.

Bush, Vannevar. "As We May Think." *The Atlantic* (July 1945). Accessed December 11, 2013. 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. Accessed December 19, 2014. http://dl.acm.org/citation.cfm?id=132180.132193.

Carlson, Nicholas. "Upworthy Traffic Gets Crushed." Business Insider. February 10, 2014. Accessed February 12, 2014. http://www.businessinsider.com/facebook-changed-how-the-news-feed-works--and-huge-website-upworthy-suddenly-shrank-in-half-2014-2.

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

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

Chambers, Ephraim. *Cyclopædia, or an Universal Dictionary of Arts and Sciences*. 1728. Accessed December 18, 2014. 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/.

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. Accessed February 20, 2015. http://link.springer.com/article/10.1007/s11135-013-9847-z.

Constine, Josh. "Google Destroys Rap Genius' Search Rankings As Punishment for SEO Spam, but Resolution in Progress." TECHCRUNCH. December 25, 2013. http://techcrunch.com/2013/12/25/google-rap-genius/.

Crawford, Kate. "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society* 15, no. 5 (June 2012): 662–679. Accessed September 5, 2013. http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878.

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

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

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

Deleuze, Gilles, and Félix Guattari. *A thousand plateaus: capitalism and schizophrenia*. Minneapolis: University of Minnesota Press, 1987.

Derrida, Jacques. "Archive Fever: A Freudian Impression." Translated by Eric Prenowitz. ARTICLETYPE: research-article / Full publication date: Summer, 1995 / Copyright © 1995 The Johns Hopkins University Press, *Diacritics* 25, no. 2 (July 1, 1995): 9–63. Accessed December 2, 2013. http://www.jstor.org/stable/465144.

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

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

Drucker, Johanna. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5, no. 1 (2011). Accessed October 24, 2013. http://www.johannadrucker.com/pdf/hum_app.pdf.

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

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. Accessed December 17, 2013. http://www.tandfonline.com/doi/abs/10.1080/13614568.2011.587030.

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. Accessed December 2, 2013. http://www.mlajournals.org/doi/abs/10.1632/pmla.2007.122.5.1580.

Gillespie, Tarleton. "The Politics of 'Platforms'." *New Media & Society* 12, no. 3 (May 1, 2010): 347–364. Accessed April 22, 2014. http://nms.sagepub.com/content/12/3/347.

Gitelman, Lisa. "Response to "Algorithms, Performativity and Governability"." New York, NY, May 5, 2013.

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

Graells-Garrido, Eduardo, Mounia Lalmas, and Daniele Quercia. "Data Portraits: Connecting People of Opposing Views." *arXiv:1311.4658 [cs]* (November 19, 2013). Accessed December 3, 2013. http://arxiv.org/abs/1311.4658.

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

Halasz, Frank. "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems." *Commun. ACM* 31, no. 7 (July 1988): 836–852. Accessed December 8, 2013. http://doi.acm.org/10.1145/48511.48514.

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 December 15, 2013. http://www.annehelmond.nl/2013/05/07/mit8-talk-exploring-the-boundaries-of-a-website-using-the-internet-archive-to-study-historical-web-ecologies/.

———. "The Algorithmization of the Hyperlink." *Computational Culture* 3 (November 2013). http://computationalculture.net/article/the-algorithmization-of-the-hyperlink.

Hern, Alex. "World's most delayed software released after 54 years of development." *The Guardian* (June 6, 2014). Accessed September 8, 2014. http://www.theguardian.com/technology/2014/jun/06/vapourware-software-54-years-xanadu-ted-nelson-chapman.

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., February 15, 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. Accessed February 8, 2015. http://dx.doi.org/10.1080/08838151.2010.499050.

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

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

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

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

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

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

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

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.

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

Lanier, Jaron. *Who Owns the Future?* New York, NY: Simon & Schuster, May 7, 2013.

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

Luzon, Mj. "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. Accessed February 20, 2015. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000262424900009&site=eds-live.

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

Manovich, Lev. "Database as Symbolic Form." *Convergence: The International Journal of Research into New Media Technologies* 5, no. 2 (June 1, 1999): 80–99. Accessed December 2, 2013. http://con.sagepub.com/content/5/2/80.

Markham, Annette N. "Undermining 'data′: A critical examination of a core term in scientific inquiry." *First Monday* 18, no. 10 (September 21, 2013). Accessed May 1, 2014. 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 | RJI." Reynolds Journalism Institute. July 16, 2014. Accessed March 9, 2015. http://www.rjionline.org/blog/saving-news-when-your-server-crashes-you-could-lose-decades-digital-news-content-forever.

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." March 8, 2015. Accessed March 9, 2015. http://www.poynter.org/news/mediawire/325301/how-the-boston-globe-is-covering-the-boston-marathon-bombing-trial/.

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

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

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. http://hdl.handle.net/2027/spo.5680986.0001.001.

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. Accessed December 15, 2013. http://doi.acm.org/10.1145/800197.806036.

———. *Computer Lib / Dream Machines*. Self-published, 1974.

———. "Ted Nelson's Computer Paradigm, Expressed as One-Liners." 1999. Accessed December 15, 2013. http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html.

Nothman, Joel. "Grounding event references in news" (August 31, 2013). Accessed February 8, 2015. http://ses.library.usyd.edu.au:80/handle/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. Accessed December 16, 2014. http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16801186&site=eds-live.

"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. Explores how media infrastructure, not content, shapes contemporary digital culture. Minneapolis: University of Minnesota Press, 2012. Accessed December 19, 2014. https://www.upress.umn.edu/book-division/books/digital-memory-and-the-archive.

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

"Pioneering hypertext project Xanadu released after 54 years." Kottke.org. Accessed September 8, 2014. http://kottke.org/14/06/pioneering-hypertext-project-xanadu-released-after-54-years.

Reagle, Joseph. *Good Faith Collaboration: The Culture of Wikipedia*. Cambridge, Mass.: MIT Press, 2010. http://mitpress-ebooks.mit.edu/product/good-faith-collaboration.

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

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." ARTICLETYPE: misc / Full publication date: Jan., 2003 / Copyright © 2003 University of Pennsylvania Press, *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 1–9. Accessed December 10, 2013. http://www.jstor.org/stable/3654292.

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

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

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 December 8, 2013. http://ksgnotes1.harvard.edu/Research/wpaper.nsf/rwp/RWP07-022.

Searls, Doc. "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/.

Seaver, Nick. "Algorithmic Recommendations and Synaptic Functions." *Limn*, no. 2 (2012): 44–47. Accessed December 14, 2014. http://limn.it/algorithmic-recommendations-and-synaptic-functions/.

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

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

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

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

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. Accessed December 19, 2014. http://www.sciencemag.org/content/333/6043/776.

Star, Susan Leigh. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43, no. 3 (November 1, 1999): 377–391. Accessed May 3, 2014. http://abs.sagepub.com.libproxy.mit.edu/content/43/3/377.

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

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

Trevino, James, and Doerte Doemeland. *Which World Bank reports are widely read?* WPS6851. The World Bank, May 1, 2014. Accessed May 10, 2014. http://documents.worldbank.org/curated/en/2014/05/19456376/world-bank-reports-widely-read-world-bank-reports-widely-read.

Walker, Jill. "Links and Power: The Political Economy of Linking on the Web." Baltimore, MD, June 2002.

Waterton, Claire. "Experimenting with the Archive: STS-ers As Analysts and Coconstructors of Databases and Other Archival Forms." *Science, Technology & Human Values* 35, no. 5 (September 1, 2010): 645–676. Accessed May 3, 2014. http://sth.sagepub.com.libproxy.mit.edu/content/35/5/645.

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*. ıPerfectly placed to tell us what′s really new about [the] second-generation Web.ȷ¯*Los Angeles Times* Business visionary and bestselling author David Weinberger charts how as business, politics, science, and media move online, the rules of the physical world¯in which everything has a place¯are upended. In the digital world, everything has its places, with transformative effects: ffl Information is now a social asset and should be made public, for anyone to link, organize, and make more valuable. ffl There′s no such thing as ıtoo muchȷ information. More information gives people the hooks to find what they need. ffl Messiness is a digital virtue, leading to new ideas, efficiency, and social knowledge. ffl Authorities are less important than buddies. Rather than relying on businesses or reviews for product information, customers trust people like themselves. With the shift to digital music standing as the model for the future in virtually every industry, *Everything Is Miscellaneous* shows how anyone can reap rewards from the rise of digital knowledge. **. New York, NY: Macmillan, April 2008.

Whitacre, Andrew. *Henry Jenkins Returns*. Accessed March 10, 2014. http://cmsw.mit.edu/henry-jenkins-returns/.

Willinsky, John. *Technologies of Knowing: A Proposal for the Human Sciences*. Boston: Beacon Press, January 1, 1999.

Wolf, Gary. "The Curse of Xanadu." *Wired* 3, no. 6 (June 1995). Accessed December 15, 2013. http://www.wired.com/wired/archive/3.06/xanadu_pr.html.

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"." ARTICLETYPE: research-article / Full publication date: Jan., 2003 / Copyright © 2003 University of Pennsylvania Press, *Journal of the History of Ideas* 64, no. 1 (January 1, 2003): 61–72. Accessed December 10, 2013. http://www.jstor.org/stable/3654296.

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. Accessed December 12, 2013. http://nms.sagepub.com/content/11/1-2/95.

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. Accessed December 8, 2013. http://papers.ssrn.com/abstract=2329161.