The Missing Links: Aggregating History and the Order of Data

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: Aggregating History and the Order of Data

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

My abstract will go here.

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 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.

Chapter 2

Layers of containment

We need 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's 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's taxonomy and Dewey's Decimal System.

Today's methods of classifying—and possibly, organizing knowledge in general—have radically changed. However, we increasingly need such generic words to describe the increasingly digital, ephemeral world around us. The software world 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 terms. For instance, several scholars have historicized and questioned the use of the word "data." Daniel Rosenberg charted the term's use through shifting contexts since the 18th century, noting that it was initially used to describe an indisputable fact or "given" in an argument (from Latin *dare*).[1] Annette Markham likewise questions the use of the word "data" in its modern context, suggesting that, "through its ambiguity, the term can foster a self-perpetuating sensibility that 'data' is incontrovertible, something to question the meaning or veracity of, but not the existence of."[2] Johanna Drucker suggests implementing its counterpart "capta," which highlights the inherently plucked and pre-envisioned nature of all information.[3]

Other contemporary words have been similarly historicized and questioned. John Seely Brown and Paul Duguid trace the history of the word "information" in *The Social Life of Information* and forthcoming research, highlighting its long history as an "unanalyzed term."[4] Likewise, Tarleton Gillespie draws attention to the word "platform" in the context of the software industry, focusing on the implications of the term's historical meanings.[5] 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

1. **rosenberg**.
2. **markham**.
3. **drucker**.
4. **duguid**.
5. **gillespie_politics**.

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]

Along with 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."[8] 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.[9]

Another, perhaps friendlier way to describe content is as a "link." Where content implies a container (containment), a link implies a connection (expansion), promising to break free from the contained information. Looking at the link favorably, if a publisher adds a hyperlink to an article, it purports to show not only erudition (the publisher has read and vetted the content within), but also altruism (the publisher is helping the content creator and the user reach one another). But here, the link surrounds 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 simulta-

6. **berners-lee**.
7. **nelson_one_liners**.
8. **jenkins**.
9. **searls**.

neous 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. In Franco Moretti's terms, the web encourages "distant reading" while treating its constituent parts broadly and generally as "content."

In this chapter I trace the lifecycle of "content" on the web, from its inception to its eventual storage in archives and databases. Treating it as a sort of biography, I show how the web exerts varying layers of simultaneous *containment* and *connection* on its intrinsic data. One could view the result as if in concentric circles, as a series of *wrappers* or *levels of abstraction* around the original source.[10] 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, the original converter into content, 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 layer of containment, the content acquires new metadata (or context), created by 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.

2.1 Theorizing containment

In crafting a biography of content, I am nodding towards the "biography of things" introduced by Igor Kopytoff. Like Kopytoff, I am interested in the passage of objects from one state to another, and the transitional moments that mark events in

10. **add figures here; one concentric-circle view**; **one more concrete view showing a specific sample webp**

a thing's history. This framework complicates the idea of any single, indelible act of categorization on an object—instead, an object is "classified and reclassified into culturally constituted categories."[11] But I also look to Kopytoff and the "social life of things" as they relate to commodities and exchange value. For Kopytoff, states of transition are equivalent to acts of exchange—in other words, the transitional is also the transactional. But if an object's "saleability" indicates its commodity status, what is the saleability of a digital object? Is content a commodity? The word, used so often in marketing contexts, implies a transaction of sorts—whereas the "link" implies exchange. How might the ways in which a piece of content is transferred, linked, and shared online reveal something about the web's culture and economy, and its notions of value? How does the language of free information and open source likewise affect or hide the transactions at play? Does the lack of a privileged "original" copy explode traditional notions of value and exchange? I am especially interested here in the act of *replication*, and the ease with which digital objects are copied. Any time you move a file, your computer actually copying it (and deleting the original); any time you watch a video online, your computer is actually reading a local copy. It is no accident that so much of modern computer architecture was developed at Xerox PARC, a research lab sponsored by the world's foremost copying company; Xerox was openly nervous about any file systems that did *not* employ copying as a central act.[12]

In following a piece of content, I am looking to analyze the whole; a single online photo might turn up in far-reaching corners of the web, and imply many acts of exchange and use around it. Specific acts of classification play into a greater whole that is interlinked by societal understanding of what constitutes a category, and how an object should be categorized. Kopytoff recognizes the need for healthy and cohesive classification as well: "Both individuals and cultural collectivities must navigate somewhere between the polar extremes by classifying things into categories that are simultaneously neither too many nor too embracing. In brief, what we usually refer

11. **kopytoff**.
12. **lanier**.

to as 'structure' lies between the heterogeneity of too much splitting and the homogeneity of too much lumping."[13]

Here I am informed by the study of infrastructure, and especially Geoffrey Bowker and Susan Leigh Star 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 affect the entire structure of technology and culture. 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—standards that overlap and depend on one another in important ways.[14] *Sorting Things Out*, along with Star's companion article "The Ethnography of Infrastructure," point to the large-scale effects of small-scale sorting. They highlight the problems and limits with traditional classification, and suggest ways to render it more dynamic and responsive. But Star also points out, quoting Gregory Bateson, "What can be studied is always a relationship or an infinite regress of relationships. Never a 'thing.'"[15] In other words, in studying a single thing, it is important to recognize its embeddedness; content is never just content, and to describe it is to also describe its containers.

The classification standards in place, as suggested by Bowker and Star, give varying levels of agency to the systems and the humans working within them, which suggests an approach informed by actor-network theory (ANT) and its treatment of ontology and agency.[16] While the web itself does not impose any singular ontological

13. **kopytoff**.
14. **bowker_star**.
15. **star_ethnography_of_infrastructure**.
16. **latour**.

framework (except, to an extent, in the case of the Semantic Web), the databases and platforms that draw from it use its structure to organize their own knowledge. Moreover, many of these databases—such as the search indexes run by Google—apply advanced algorithms that filter and divide content at massive scales, untouched by human hands. In other words, the interplay between nonhumans and humans is crucial in the production and distribution of content on the web (in fact, many of the web's end users are also machines, understanding and accessing data through HTML markup or APIs rather than words and images). My framing of content/container may seem to imply a hierarchical or one-way relationship between the whole (the web or the archive) and the part (the piece of content), forgoing the "flat" network proposed by ANT. However, as Bowker and Star suggest, the part always influences the evolution of the whole, and the layering of the web turns the container back into content. Debates about the border between content and context (or data, or metadata) fall apart, as they often collapse into one another.

2.2 Stage One: The URL

Content has a rich backstory before it arrives on the web, but I am treating its birth âĂŞ the moment when an object or artifact becomes a piece of content âĂŞ as the moment in which it is uploaded to the web. It now has its first container, and its first piece of web-native metadata: the URL. Even before it is connected to other URLs (at which point it becomes a "link") an end user can access it online via a string of text. As Tim Berners-Lee tells it, the Uniform Resource Locator was one of the most difficult concepts to develop and understand as he began to weave the web. 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. The virtual space of the web here seems to mimic physical space in the world, suggesting that one can find a document at 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 text 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" domain, 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 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 that vague, homogenizing definition does not hold up.

Eszter Hargittai points to the need for technical expertise in order to properly understand a URL and the implications behind it. 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.

URL "shorteners" such as those employed by the firm bit.ly 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.

Perhaps the explosion of the URL was an inevitable byproduct of the web's very structure. It is infinitely distributed and highly networked; it shuns hierarchical organization schemes, which seems to go against 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. Where knowledge was once shaped like a tree, on the web it looks more like Deleuze and Guattari's rhizome: an infinitely "intertwingled" mass. One cannot make sense of it using URLs alone, but links offer a start.

2.3 Stage Two: The Link

The birth of the "link" occurs at a second level of containment, after an object becomes "content" with a URL. The link wraps the URL in an HTML element that allows it to be quickly accessed from another page. Without links, the web would just be a series of disconnected nodes; with links, the web becomes a network. Bowker and Star suggest that links have the power to classify without any human agency or intervention, which 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." 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." 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 tagâĂŤthe string "<a href>âĂę" and everything in betweenâĂŤin an HTML document). 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 destinationâĂŤback 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 windowâĂŤso in a literal sense, the link contains the URL.

The link is forever associated with (and perhaps plagued by) the footnote. Nelson's hypertext manifesto Computer Lib/Dream Machines praises the screen for permitting "footnotes on footnotes on footnotes," 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 âĂŞ its manifestation in a browser, known as the anchor text âĂŞ 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.

The anchor text has no innate relationship to its target, and it is only pointing to the target's address. As a result, the link can be seen as a sign. Analyzing the link's anchor text through a semiotic frame reveals a number of interesting conventions and uses, each of which bears underlying motives. The many flexible uses of the link 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. When a link's text is identical to its address, like "http://www.google.com," it seems to be removing this layer entirely. However, there is nothing stopping someone from putting a completely different address into the anchor text, further emphasizing the lack of relation between anchor and target, or signifier and signified. This distance is what allows a scam artist to direct an unknowing user to a phony bank website, even if the stated URL is for their real

bank. It is also used for more playful and innocuous ends, such as with "rickrolling," a meme where someone provides a purportedly useful link, but it actually leads to a video of Rick Astley's 1987 hit "Never Gonna Give You Up." Whether playful or nefarious, both of these uses are enabled by the structure of the link, and the lack of relationship between the text and the target.

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. 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. Studies and practices that analyze and aggregate links would do well to closely analyze the text of the link. There have been very few large-scale studies of the semiotics of linking, or the way in which the anchor text helps to gain insight into the target resource or the connection being made. One exception comes from a small 2006 study of automated blog classification, where the researchers determined that the anchor text was in fact the best signal for improving classification. One of the researchers now studies the text of tweets to gain insight into the links they embed, once again treating the users' descriptions of links as more important than what networks are sharing it.

But for now, most studies simply take an aggregate view of link sharing, treating each connection as equal regardless of context. This has vast implications for the news media and has undoubtedly affected content creation and discourse. Anyone who shares an article inevitably, and perhaps inadvertently, raises the article's profile and algorithmic rank whether they liked it or not. 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. This is best exemplified by Rusty Foster's "Today in Tabs" newsletter, which popularizes the idea of "hate-

reading" and links to some of the most infuriating articles in the news. It is not clear to algorithms whether or not someone liked an article (let alone why they liked it)âĂŤit is only clear that they are talking about it. This may be because there is no straightforward way for an automated system to understand the many cultural nuances behind a link.

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

Another alternative to the web's form of linkage comes from Ted Nelson, a long-time critic of the web's architecture. As the original hypertext visionary, his scheme, called Project Xanadu, floundered for decades, and was never truly 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 PREVENTâĂŤever-breaking links, links going outward only, quotes you can't follow to their origins, no version management, no rights management." 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. 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. Jaron Lanier believes this simple omission has profoundly affected culture and economics, which forms a cornerstone of his recent book Who Owns the Future?

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.

There has been much recent controversy surrounding the restructuring of the news headline in the hyperlinked age. 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." The headline is "click bait," playing to a user's innate curiosity (Atlantic writer Derek Thompson calls it the "curiosity gap") 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 just as 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. The website Upworthy is a canonical example of click-baiting headlines, and even its organizational structure is revealing; the company creates no original content, but instead employs people to trawl the web, find content, and put a new headline on it. The team is not creating new content, but new containersâĂŤand

it is one of the most popular and successful media business efforts of recent years. Despite this, Upworthy has been mocked frequently, such as via the joke news site "Clickstrbait," which leads users down a rabbit hole of curiosity-inducing headlines without guiding them to actual content.

Interestingly, Upworthy is one of the first websites to attempt to move beyond the simple "pageview" metric, heralding a new measure of success called "attention minutes." These metrics will make privacy advocates cringe; by monitoring which browser tab is open, where the mouse is pointing, or how much of a video the user has watched, Upworthy hopes to understand user behavior more deeply. Upworthy's blog claims that "this is a metric focused on real user satisfaction," but it is still a measure of behavior as a proxy for emotion, and the end goal (a like? a share? a donation to a worthy cause?) remains unclear.

In all of these cases, the layers of containment could be seen as layers of signification. I stated earlier that the birth of content (the transformation of an object into content) occurs at the moment it is uploaded to the web, and accessible via a URL. Here is where it moves from essential object to sign and message, in Jean Baudrillard's terms. Looking at the web as a designed artifact with a specific, graspable structure, Baudrillard proves fruitful in emphasizing the web's political and philosophical origins and ramifications:

The semiotic revolutionâĂęconcerns virtually all possible practices. Arts and crafts, forms and techniques both plastic and graphicâĂęwhich until then were singular and distinct, are synchronized, and homogenized according to the same model. Objects, forms, and materials that until then spoke their own group dialect, which only emerged from a dialectical practice or an original style, now begin to be thought of and written out in the same tongue, the rational esperanto of design. Once functionally liberated, they begin to make signs, in both sense of the phrase (and without a pun): that is, they simultaneously become signs and communicate among themselves. Their unity is no longer that of a style or practice, it is that of a system.

Replacing "design" with a word like "information" or "data" reveals the homogenizing force of the web and its ability to squash varieties of creative works (photos,

videos, text, music) into data, which allows for easy exchange, commodification and reuse.

2.4 Stage Three: The Feed, the Index

Links rarely exist in isolation. For one, links contain links themselves, as I touched on in the last section. But another form that the link 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, allowing for another layer of containment beyond the link. 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. The table usually truncates the content into a headline, and perhaps an image or opening paragraph. 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 feedsâĂŤin other words, 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, completely powerless without it. This is evidenced by Upworthy itself, who recently found its content potentially demoted in Facebook's algorithm with no meaningful explanation, shrinking its immense traffic to half of its previous size. Another major content distributor, the lyrics annotation website Rap Genius, recently found its pages move from the top hit on Google to its seventh page, due to changes in Google's algorithm. These content aggregators can move around large swaths of content (millions upon millions of interlinked pieces) via slight changes in their codebases, with no obligation to inform anyone of the reasons or even that it is occurring. This is perhaps the highest level of containment, and few (if any) actors can claim to contain these sites in turn.

To be fair, 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. In effect, 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."

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. A large portion of website hacks and attacks are merely to insert hidden links on the targeted sites. In the process, Google has had to remain one step ahead of the advertisers, with the link as the battlefield, influencing the web and changing its structure in turn. But this battle has mostly been played out by machines, which are responsible for a substantial amount of the links created âĂŞ as well as the links browsed and followed âĂŞ on the web. Besides a generic, easily replaceable piece of metadata in a web request, it is in fact impossible 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.

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 1999, and the "model of user behavior" must undoubtedly be changing as the web's capabilities and browsing habits change. This bizarre mixture of human and nonhuman, as well as the substantial influence that links have on the information we encounter as everyday users, speaks to the usefulness of actor-network theory in framing the political economy of links and linking.

While links are shared for a variety of reasons âĂŞ some of them more nefarious than others âĂŞ 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 free information sharing. Looking at it through the lens of Marcel Mauss's writings on gift exchange, however, it seems to carry more weight than this: "Exchanges and contracts take place in the form of presents; in theory these are voluntary, in reality they are given and reciprocated obligatorily." This can be seen beyond the blogs of Web 2.0; users exchange links on Twitter and retweet, favorite, or like posts on various social media platforms. In each case, a link or like on a social media post is performative and transactional, with the implicit expectation of a future like in return.

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

bubbles." Many studies of links have traced political homophily, public debate, blogs and global flows of information; if we take these at face value and treat hyperlink usage as a proxy for importance, impact, and communication, then link-sharing can turn conversation inward, allowing searchers to see only blogs that have overtly linked to one another (blogs which, presumably, have similar views and opinions). While the Internet may allow for a more heterogeneous group of voices to surface than in traditional media (and indeed, this is one of the ways in which the medium is widely celebrated), one must still take part in link sharing with a particular group in order to be found, leading bloggers into already-established and tightly wound networks. This phenomenon is most expertly outlined by Philip Napoli, who calls it "massification": in the editorial and algorithmic decisions that determine where links are placed and directed, there is a distinctive replication of old "mass" media patterns.

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 a series of layered containers. Moreover, the information, story, or "piece of content" that they may be after is 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.

Repetition plays a strong role in Freud's definition of the uncanny. I wouldn't suggest that a user is frightened by search results, but there is a sense of unease or anxiety in finding the same content repeated ad infinitum. Google's "search by image" feature provides a list of "visually similar images" that reveal hundreds of nearly identical photos. For example, an image search for "office meeting" turns up the same stereotypical figures; businessmen and businesswomen in suits, seated around a table, poring over seemingly identical documents in a typical conference room. The photos are mere signifiersâĂŤit seems clear that the subjects are actors, and no business is actually being done. The emptiness of the content itself, and its

endless repetition, is highly unsettling. Freud's notion of the uncanny has also been applied to the context of online advertising. Often when a user visits a product page, the same product is then re-presented to them in the sidebar of a completely different siteâĂŤoften mere minutes later, but other times it takes days or months. This is a more direct application of the uncanny, as it does make the user feel as if they are being watched.

2.5 Stage Four: The Archive

Content's final resting place is in the database, or archive. But all the same, it is not fair to call it "final," since the context and metadata surrounding it is always subject to change. Moreover, this lifecycle is vastly oversimplified; often the content reaches a database as soon as it is accessible via a URL (for instance, with a photo that is uploaded to Flickr or Instagram). Then as the content moves around the web, affected by other creators, distributors, aggregators and indexers, it is placed in an untold number of databases, with varying structures and associated metadata. So in a sense, the four stages that I have outlined here collapse on one another, and the framing that I have offered is far too neat, simple, and narrative-driven for the distributed, infinitely networked, rhizomic web.

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, works with it. While users increasingly treat the web as a database âĂŞ what we know familiarly as the "cloud" âĂŞ it is less distributed and hypertextual than that metaphor seems. There is no single database, but rather very many, housed on servers around the world. Each of them faces the same 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 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. Jacques Derrida traces this phenomenon in Archive Fever, where he calls the archive "a pledge, and like every pledge, a token of the future."

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. Derrida highlights this challenge as well: "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. Order is no longer assured." Derrida thus relates the archive to a prosthesis, a built and artificial entity that mimics but does not replicate the infinitely rich sensoria of reality. Claire Waterton, citing Michael Taussig, also uses the border metaphor to describe the increasing diffusion of information: "the border zone of representation is currently expanding, proliferating, and blurring, becoming permeated by itself."

Seen in this way, the database is perhaps the only truly containing force; the previous 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. But all the same, 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. Character strings and encoded bits are devoid of context and semantic meaning. They certainly do no justice to the richness of the original content, which points to a proliferation of associated narratives (for instance, each photograph has a photographer, a subject, a setting, a camera, and all of the processes that formed the "becoming" of these entities, and which we implicitly consume and consider as we look at that photograph).

My aim is not to suggest any overarching solution to the limitations of the archive; it is, in fact, 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. Classification is an inherently epistemological, performative act that is always embedded in a certain community and always subject to change. In short, it is socially constructed. In order to fully recognize and remember this, Bowker and Star 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." 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 this 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.

Likewise, I intend to propose that 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 taxonomy and contemporary chaos through its use of user-generated ontologies. In order for machines to understand a network, everything must be definitively categorized, but the categorization scheme itself is subject to change. Certain standards arise, but each individual or community is free to create its own form of linked data. All the same, the slow adoption of the Semantic Web may have to do with its reliance on these "ontologies"; even if multiple ontologies can coexist, they are still trying to compromise the web's disorder.

Derrida's "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 showsâĂŤthey 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, each system is limited by the borders of the database, and aside from folders and tags, none provide a useful way to store meaningful associations between these documents.

In all cases, the problem of "information overload" is paramount, and the virtual piles of documents and content get increasingly difficult to wade through and make meaning of (unless one takes the high level, "big data" viewâĂŤa perspective that has its own pitfalls, as danah boyd and Kate Crawford show). But information overload is nothing fundamentally new. Ann Blair finds a complaint about a "confusing and harmful abundance of books" as early as 1545 (in Conrad Gesner's attempt to catalog all known books), and many other scholars have historicized information overload and management strategies (such as commonplace books, scrapbooking, "stringing," and the encyclopedia). One of the most canonical methods of organizing too much information is the card catalog, in use by libraries for more than a century; early hypertext systems, such as Xerox PARC's NoteCards and Berners-Lee's ENQUIRE, are noteworthy in their remediation of the affordances of this old tool.

However, 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 response to this, users store everything, in hopes of using the archive's power of containment to understand it. But containing is not understanding, and by turning rich multimedia into bits of text, containment in fact furthers the distance between the user and the real, lived experience that the content aims to capture and describe.

2.6 Postscript: Erasure and Afterlife

Content has an afterlife when it goes through a sort of reversal of the stages outlined above; it must be plucked from an archive by a search algorithm, which is in turn responding to a request by a user. Some content never does live again; for instance, nearly one-third of all reports on the World Bank's website have never once been downloaded. This does seem to run counter to the "pack-rat" mentality of users and institutions proposed earlier, but it also points to the vast amounts of knowledge we are creating that may require a new format to be rendered useful. This is not to say that the knowledge contained in the World Bank's documents has been utterly forgotten (the document could be distributed by email, or presented at a conferenceâĂŤthe inability to track it is the crux of the problem); only that it is (literally) not helping anyone in its current structure.

Other content may in fact be useless, or worse, detrimental. Knowledge, and even facts themselves, have been rephrased, rewritten, and reversed for as long as facts have held public influence. Some content is outright false, misleading or slanderous, and other content is simply embarrassing. Certain regrettable pieces ought to be remembered (such as a racist comment made by a powerful public figure); others are more likely best forgotten (such as a teen's suggestive selfie that gets distributed around the internet). But the question is not who determines what deserves a place in history and what should be erased; even if one deletes the content, it is not likely to disappear entirely.

The user's experience of deleted content is the broken or dead link, the ubiquitous 404 "Not Found" error page. While in some cases a dead link signifies a necessary removal (such as the teen's photo), in others it is a stand-in for lost knowledge. There's no doubt that content does disappear; studies have found that 30-50 percent of citations 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," which the Hiberlink initiative is currently researching the extent of). There is, of course, a substantial chance that the content still lives

somewhere on the web, often in multiple places; but if it is no longer at the path specified by the link, it will be much more difficult to find. To combat this, archivists and cultural heritage institutions aim to preserve the web's history for later retrieval. The Internet Archive crawls and stores as many websites as possible, while the Archive Team aims primarily to preserve discussion forums and old blogs. Unlike other groups, the Library of Congress saves websites worth saving by manually choosing them, often taking an "aggregate the aggregators" approach and storing large text databases. In each of these cases, groups are establishing an archive that is perhaps less financially motivated than a company's database, aiming instead to preserve the knowledge and associations within for public benefit.

2.7 Conclusion

Hypertext is built on the premise of collapsing traditional, hierarchical categorization schemes, felling the tree and digging to find the rhizome. This information structure certainly has its historical precedents; a reference book, such a dictionary or encyclopedia, is a classic example. Organized alphabetically (which is to say arbitrarily), it is always referring to other words and terms, requiring the dedicated reader to jump from one page to another, following any thread at will. Michael Zimmer connects Diderot's EncyclopÃľdie and its use of renvois to the hyperlink, noting its ability to subvert hierarchical knowledge distribution and censorship in the process. Much of this language is echoed by Nelson, Berners-Lee, Paul Otlet and the many early champions of the web, who saw it as a democratizing force leading towards social good.

However, the web's highest-level platforms now encompass, embed, and contain all other media, a phenomenon that is difficult to see as we users browse one page at a time. While this structure affords certain advantages, there should not be a one-size-fits-all model for experiencing media, communication, and culture. The web provides no built-in way to "zoom out" and see overarching link structures; it does not allow curious users to trace content to its origins; and it is a disorganized mass

that various actors have spent a massive amount of time (and often earned a great deal of money) sorting out. As Bowker and Star remind us, each act of sorting has consequences, and that we rely on sites like Google to do it for us, with no obligation of transparency, is a dangerous reality to live in.

The web and the archive's acts of containment, on every level, likewise have real economic consequences. There has been much lamentation of the demise of the "creative class," reducing rich and multifarious works to the act of "content creation." Similarly, there is much trepidation about big data companies that ingest this content and our interactions with it, making billions of dollars in order to grant us access. I would suggest that these trends are (not necessarily caused, but) enabled by the structure of the web itself. Looking to new structures and forms of classification would do well to counteract the containing, homogenizing forces of computation, and the "big words" (data, information, document) that come with it.

Chapter 3

The Size and Shape of Archives

3.1 Introduction

3.1.1 The Networked Archive

The word âĂIJarchiveâĂİ brings to mind a stuffy room full of old books and manuscripts, closely guarded by librarians. In the traditional archive, 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 (or manuscript), even though books themselves contain multitudes of media (text, images, maps, diagrams) and the bibliographies and indexes that offer a window into a bookâĂŹs constituent parts remain limited by space and language. If your archive dive takes you beyond the books in the current room, youâĂŹll have to leave the room.

But archives come in many forms. More recently, an archive is likely to be digitized, stored on networked servers in databases. Here the archiveâĂŹs 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 archiveâĂŹs scope, it may even be able to reach for information outside of it. In short, the digital affords new abilities for linking or networking the archive, allowing it to dynamically expand, contract, and change shape. In the networked archive, we can forge new connections and create more nuanced context for the information stored inside.

Most of todayâĂŹs 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.

A networked 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-networked 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.

The act of networking 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. Moreover, the web is only one of many digital hypertext systems, and the hyperlink âĂŞ the primary object and mechanic for network formation on the web âĂŞ has its own limitations that early hypertextual systems bring into full relief, inviting close analysis of the webâĂŹs archival affordances.

3.1.2 Defining the Archive

The notion of the archive has exploded even beyond its new digital meaning. Foucault uses the term to refer to âĂIJsystems of statementsâĂİ that consist of the âĂIJhistory of ideas,âĂİ the entirety of sayable things and their referents. FoucaultâĂŹs epistemological archive subsumes both the stuffy room and the digital database into itself. So is the archive literal, digital, or figurative? What size and shape does it take? Does it represent an individualâĂŹs memory, or collective history?

This shifting notion of archive varies based its shape and its scope. An archive can

be personal, institutional/collective, or universal. Despite the vast difference between, say, a studentâĂŹs 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, like the referent, 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 indexesâĂŤa 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 BorgesâĂŹ exact one-to-one map of the world. 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 archiveâĂŹs 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. TodayâĂŹs digital archives are not so neatly or hierarchically drawn. Jacques Derrida uses the border metaphor to describe the recent diffusion of archives: âĂIJthe 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.âĂİ Claire Waterton likewise suggests that the border zone is âĂIJcurrently expanding, proliferating, becoming permeated by itself.âĂİ Reflecting the postmodern skepticism towards standard categories and hierarchies, the networked archive morphs and munges its contents 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 itemsâĂŹ organization and discovery, with the aim of helping people find information, structure ideas, and do research. This includes the systems surrounding collection itselfâĂŤorganizational, structural, and sociocultural. To put it in Lev ManovichâĂŹs terms, âĂIJdata structures and algorithms are two halves of the ontology of the world according to a computer.âĂİ I am interested in an archiveâĂŹs data structures (specifically with regard to its itemâĂŹs 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 âĂIJarchiveâĂİ is similar to the concept of the âĂIJdatabaseâĂİ 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. 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. As Marlene Manoff says, âĂIJThe 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.âĂİ Following Jerome McGannâĂŹs insights, I see the database as a technical instrument used for the structuring and enabling of archives; it is not the archive itself.

Like McGann and Manoff, I also use the word to emphasize a lineage. TodayâĂŹs information management tools continue to inherit many ideas and techniques from traditional archives and note-taking systemsâĂŤa fact that âĂIJdatabaseâĂİ doesnâĂŹt 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. These archives are also deeply embedded within and linked to one another through APIs, further complicating the picture.

The rise of the knowledge economy 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 âĂIJwe are all miniarchivists ourselves,âĂİ calling the information society an âĂIJinformation management society.âĂİ Belinda Barnet considers it a âĂIJpack-ratâĂİ mentality, while Derrida succinctly and famously titles the phenomenon âĂIJarchive fever.âĂİ My use of the term encompasses traditional archives, modern databases, and the algorithms and interfaces in betweenâĂŤthe indexes, note-taking systems, bibliographies and encyclopedias that first forayed into networked information.

3.1.3 Outline of Themes

Most histories of the proto-web 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-LeeâĂŹs 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 âĂŞ whether they existed long before computers or after the rise of the web âĂŞ that point towards three corresponding themes. These themes recurrently surface when dealing with digital archives and information management.

The first section addresses the spatialization of memory and knowledge. Here I consider the use of visual metaphors for information and the associations between memory and physical space. The spatial and dimensional nature of knowledge is at odds with the âĂIJflatteningâĂİ effect of indexes and the collapsing of dimensional space that non-hierarchical linking affords. Cycling through Ephraim Cham-

bersâĂŹ Cyclopaedia, I will then examine Paul OtletâĂŹs vision of the âĂIJradiated libraryâĂİ and his architectural inspirations.

The second section turns to the intersubjectivity of knowledge, or the relationship between personal memory and collective history. An individualâĂŹs personal archive has markedly different properties and requirements than a groupâĂŹs or institutionâĂŹs, 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 webâĂŹs algorithmic recommendation systems that conflate the individual and the collective as they traverse the links of the web. In this section I will examine Vannevar BushâĂŹs memex machine, which wavered between personal study aid and collective knowledge generator.

The third and last section analyzes the encyclopedism of knowledge systems: the constant striving to expand beyond the archiveâĂŹs horizon, and to achieve total comprehensiveness. Where intersubjectivity is concerned with the amount or makeup of a systemâĂŹs users, this section concerns the amount and structure of the information in the archive. Many efforts to document, index, or link the world have truly attempted to map the world âĂŞ every piece of information about everything âĂŞ or have at least appealed to an impulse to do so. In this section I endeavor to explain this encyclopedic impulse, and suggest some of its promises and pitfalls.

3.2 Spatialization

3.2.1 The Dimensions of Memory

Memory is inherently spatial. Even when we donâĂŹt remember something, we often know where to find it. A recent study asked participants to save statements into

various folders with generic names (such as FACTS, DATA, INFO, and POINTS). Despite the unmemorable folder names, âĂIJparticipants recalled the places where the statements were kept better than they recalled the statements themselves.âĂİ The researchers found that âĂIJâĂŸwhereâĂŹ was prioritized in memory,âĂİ providing preliminary evidence that people âĂIJare more likely to remember where to find it than to remember the details of the item.âĂİ They conclude by suggesting that we may be using Google and Wikipedia as memory extensions that then rewire our 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 âĂŞ Barnet might call them âĂIJmemory machines,âĂİ John Willinsky âĂIJtechnologies of knowingâĂİ âĂŞ to help summarize, filter, sort, and select. Computer systems are only one piece of this longer history of tools and practices. David WeinbergerâĂŹs âĂIJthree orders of orderâĂİ 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.

A theme across all of these orders of order is a reliance on spatial memory (the âĂIJwhere to find itâĂİ in the Columbia study). Archival and classification schemes use terms like âĂIJborder,âĂİ âĂIJdomain,âĂİ and âĂIJkingdomâĂİ (is it a coincidence that these terms all carry connotations of politics and power struggle?). We visualize network schemes as trees and as rhizomes, represented on maps, graphs, and diagrams. It seems that proper spatial visualization of an archive might not only help us remember where something is saved, but also give a high-level understanding of the archive itself, improving browsing and serendipitous search. The ancient practice of constructing âĂIJmemory palacesâĂİ (and Giulio CamilloâĂŹs memory theater of the Renaissance) âĂŞ outlined in Frances YatesâĂŹ The Art of Memory âĂŞ strongly emphasizes memoryâĂŹs reliance on spatial orientation and fixed dimension. 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 reorderedâĂŤonly added to. This completes a grander spatial metaphor that Peter Burke notices âĂŞ that of the course, which a student must run, envisioning and memorizing images in places along the route towards knowledge.

This reliance on spatial 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 libraryâĂŹs stacks and expect to find it again. A similar concern arises with online reading and writing. Ted Nelson calls hypertext âĂIJmulti-dimensional,âĂİ and Stuart Moulthrop says it aims to be âĂIJwriting in a higher-dimensional space,âĂİ 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, âĂIJI remember reading that sentence near the top of the page in the left columnâĂęâĂİ). 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 âĂIJnavigateâĂİ down âĂIJpathsâĂİ in given âĂIJdomains.âĂİ

This last fact is surprising given that a common theme among hypertextâĂŹs pioneers, including Berners-Lee, is a desire to break down traditional linear and hierarchical classification schemes. A hierarchical scheme âĂŞ like LinnaeusâĂŹs biological taxonomy or DeweyâĂŹs decimal classification âĂŞ immediately suggests a tree view, and we can find many old examples of tree graphs in the Renaissance and Enlightenment. On the other hand, an alphabetical scheme offers a linear view, one that âĂIJflattensâĂİ 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?

3.2.2 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 ChambersâĂŹ 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 âĂIJcapable of the advantages of a continued discourse.âĂİ The word âĂIJencyclopediaâĂİ literally means âĂIJcircle of learning,âĂİ 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 âĂIJintellectual entropyâĂİ and âĂIJepistemological urgencyâĂİ of the time. 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 âĂIJegalitarianâĂİ forms of knowledge organization. For instance, religious topics would be placed alongside secular ones. Alphabetical organizational also turned the system into more of a âĂIJquick referenceâĂİ 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 âĂIJa confusing and harmful abundance of booksâĂİ as early as 1545 (Chambers even called this complaint âĂIJas old as SolomonâĂİ).

All the same, Chambers felt he needed an overarching scheme. In the encyclopediaâĂŹs preface, he included a diagram and listing of forty-seven categories (called Heads), complete with cross-references to the entries. In ChambersâĂŹ words, âĂIJthe 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.âĂİ In order to truly demonstrate a âĂIJcontinued discourse,âĂİ Chambers needed a graph, a map. Each of the Heads in the diagram contains a footnote that lists that headsâĂŹ terms (known as Common Places).

ChambersâĂŹ use of Heads and Common Places followed Phillipp MelanchthonâĂŹs 1521 subject division into loci and capita (Peter Burke suggests that these would now be called âĂIJtopicsâĂİ and âĂIJheadings,âĂİ less strong and physical metaphors). Loci (âĂIJplacesâĂİ) bring to mind memory palaces, but also the âĂIJcommonplace bookâĂİâĂŤto which Chambers was knowingly attaching himself. Many scholars used commonplace books as information management devices to store quotes, summaries, aphorisms, and so on, and these often had specialized systems for retrieval. Richard Yeo sees ChambersâĂŹ use of the term as directly appealing to the popularity of commonplace books at the time. 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 âĂIJno explicit reference to one another.âĂİ 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 DiderotâĂŹs celebrated EncyclopÃľdie (Diderot started by translating Chambers). DiderotâĂŹs work made further use of renvois (references) to question and subvert traditional knowledge structures and authoritiesâĂŤincluding the bookâĂŹs 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, âĂIJguiding the reader to radical or subversive knowledgeâĂİ while evading the eyes of the censors. Zimmer directly ties the renvois to the hypertext link, suggesting that Bush, Nelson, and Berners-Lee all âĂIJintended to free users from the hegemony of fixed information organization in much the same way that renvois did for the readers of the EncyclopÃľdie.âĂİ

It is clear that Diderot fully recognized and built upon ChambersâĂŹ developments in linking references, but I call into question the notion that the prior âĂIJfixedâĂİ organization systems had no detractors or provisional solutions (moreover,

the renvois are âĂIJfixedâĂİ themselves). Carolus Linnaeus, the author of perhaps the prototypical taxonomy, knew well that classifications are âĂIJcultural constructs reflecting human ignorance.âĂİ Leibniz also understood its limitations; his Plan for Arranging a Library included a âĂIJmiscellaneousâĂİ section, a tacit acknowledgement that the system is in some way imperfect or incomplete. Leibniz also praised his famous Note Closet, developed by Thomas Harrison, for this same ability: âĂIJA single truth can usually be put in different places, according to the various terms it contains âĂę and different matters to which it is relevant.âĂİ

Moreover, multiple hierarchies can coexist and offer competing schemes. Some of these schemes were already organized not as much around content as context. 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. This is, crucially, an early citation-based network.

3.2.3 Paul Otlet and the Radiated Library

Along with Vannevar Bush, Paul Otlet bridges the second and third orders of order. Born in Belgium in 1868, Otlet predated Ted NelsonâĂŹs 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 âĂIJthe city where the world comes to take notesâĂİ 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 âĂŞ particularly his biographer Boyd Rayward âĂŞ 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 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 worldâĂŹs information. Second, following his belief that books were redundant and arbitrary agglomerations that obscure the data held within (which is the object of a researcherâĂŹs true inquiry), he suggested a universal decimal classification system that built on DeweyâĂŹs system to incorporate an itemâĂŹs metadata, its references and constituent parts. Its entries read less like library call numbers and more like modern databasesâĂŹ structured queries. Finally, in his most striking prediction, he proposed a âĂIJradiated libraryâĂİ 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: âĂIJCinema, phonographs, radio, television, these instruments taken as substitutes for the book, will in fact become the new book.âĂİ OtletâĂŹs concept of a âĂIJradiated libraryâĂİ and a âĂIJtelevised bookâĂİ 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 TraitÃĩ de documentation a study of the âĂIJarchitecture of ideas.âĂİ The first names for the Mundaneum âĂŞ the universal repository Otlet and La Fontaine set out to build âĂŞ were alternately âĂIJcity of knowledgeâĂİ and âĂIJWorld Palace.âĂİ In the end, the Mundaneum âĂŞ like the archive itself âĂŞ bridged the physical and the digital, as Otlet called it at once âĂIJan idea, an institution, a method, a material body of work, a building and a network.âĂİ 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, âĂIJFor Otlet it was important to leave space for transformation and modification in response to the unforeseen and unpredictable.âĂİ Leibniz had conceived of the âĂIJlibrary without wallsâĂİ long before, but OtletâĂŹs radiated library went many steps further.

His 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 OtletâĂŹs 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, OtletâĂŹs 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 OtletâĂŹs Universal Decimal Classification. The universal repository was, of course, never completed, but it came closer to fruition than the memex or Xanadu.

3.3 Intersubjectivity

3.3.1 From personal memory to collective history

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 scholarâĂŹs 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 OtletâĂŹs Mundaneum) had the world-encompassing scope of the web.

Jeremias Drexel stated in 1638 that there is no substitute for personal note-taking: âĂIJOneâĂŹs 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.âĂİ 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 individualâĂŹs memory make it difficult to generalize categories. An individualâĂŹs thought process might be reminiscent of BorgesâĂŹ Chinese encyclopedia, which offers a taxonomy of animals divided by absurd traits, such as âĂIJThose that belong to the emperor, embalmed ones, those that are trained, suckling pigs, mermaids, fabulous ones, stray dogs,âĂİ and âĂIJthose that are included in this classification.âĂİ 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. BorgesâĂŹ categories remind us that classifications will always be compromises, between individuals and groups, or between groups and a collective whole.

Markus KrajewskiâĂŹs 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 donâĂŹt turn out to be for the same uses. Krajewski says âĂIJThe difference between the collective search engine and the learned box of paper slips lies in its contingency.âĂİ 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 âĂIJa pledge, a token of the future.âĂİ 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. 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 BenjaminâĂŹs 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 âĂIJalgorithmically rearticulate the relationship between individual and aggregate traits.âĂİ The communities and demographics that form around individuals can in turn be aggregated and intersected into a single, massive whole. At each stage, memory is abstracted further and further from us.

TodayâĂŹs 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. 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.

3.3.2 Vannevar Bush's Memex

Unlike OtletâĂŹs radiated library, or NelsonâĂŹs Xanadu, Vannevar BushâĂŹs 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 in the Atlantic called âĂIJAs We May ThinkâĂİ conceived of the memex

machine, an automated microfilm device that could store an entire library in one drawer and retrieve any item within seconds. 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 BarnetâĂŹs words, that âĂIJthere is a conviction, without dissent, that modern hypertext is traceable to this article.âĂİ

Bush begins by arguing that, âĂIJThe summation of human experience is being expanded at a prodigious rate,âĂİ but suggests that our methods for retrieving such experience are hindered by âĂIJthe artificiality of systems of indexing.âĂİ He points out the limitations of keeping data only in one place, and of using strict formal rules to access it: âĂIJthe 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.âĂİ His proposed solution, the memex, aims to mechanize âĂIJselection by association, rather than by indexing.âĂİ

The memex is built for personal use; BushâĂŹs model is âĂIJthe human mind,âĂİ after all, and not âĂIJhuman mindsâĂİ (as Barnet notes, he follows the cybernetic tradition of the time in modeling computation on human thought, along with Wiener, Shannon, Licklider, and others). 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 âĂIJfor insertion in his own memex, there to be linked into the more general trail.âĂİ

Bush goes on to suggest that the memex will hold new forms of encyclopedias and ready-made trails, along with âĂIJa new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record.âĂİ But he does not dwell on this to consider where this âĂIJcommon recordâĂİ 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 âĂŞ or in BushâĂŹs case, each act of association âĂŞ increasingly fraught and violent.

BushâĂŹs work relies on the trail, a closely curated path where one document directly associates with another. Ted Nelson instead suggested âĂIJzippered lists,âĂİ which would operate like trails but without BushâĂŹs emphasis on sequence. 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 BushâĂŹs âĂIJMemex Revisited,âĂİ he begins to tease out the possibility of the memex forming trails for a scholar, suggesting that it could âĂIJlearn from its own experience and to refine its own trails.âĂİ Here the influence of WienerâĂŹs cybernetics and feedback theory are clear, and it begins to point to the machine learning and automated classification that occurs today. Most intriguing is BushâĂŹs 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 BushâĂŹs memex is not without its limitations. John H. Weakland offered two criticisms of the memex in response to the Atlantic article. He asks âĂIJhow personal associations of the general record could be generally useful,âĂİ as well as how a researcher can find things they donâĂŹt know about already. It appears to me that the second challenge is an extension of the first: associative indexing may be more inherently fuzzy and idiosyncratic than content-based indexing systems like text search and tagging. It sacrifices fixity and consistency at the expense of individuality and nuance.

Another limitation of the memex, offered by Belinda Barnet, is that âĂIJBushâĂŹs model of mental association was itself technological; the mind âĂŸsnappedâĂŹ between allied items, an unconscious movement directed by the trails themselves.âĂİ

Bush himself recognized this, pointing out that the human memory system is a âĂIJthree-dimensional array of cellsâĂİ that can gather, re-form, and select relationships as a whole or a subset of a whole. While later hypertext systems and the Semantic Web come closer to such a three-dimensional structure, like the memex they are often constrained to âĂŸsnappingâĂŹ 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 BushâĂŹs proposed adaptive memex would be modeled against the individual researcher; in machine learning terms, its âĂIJtraining setâĂİ would not be formed in the aggregate like modern-day recommendation systems, but rather from the unique trails formed by an individual.

3.4 Encyclopedism

3.4.1 The Endless Archive

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 âĂŞ an archive built for everyone is more likely to collect everything âĂŞ but I divide them here to highlight the tendency for content to stretch towards complete and total comprehensiveness, or what I am calling encyclopedism.

When building an archive, where do you stop? Paul Otlet wanted to index all of every book. In his notes, he insists, âĂIJI 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.âĂİ This persistent, obsessive quest for comprehensiveness is part and parcel of the archive âĂŞ you either want to collect and connect everything, or everything worthwhile, within a given scope.

Once again this conjures up a Borges story: the 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.

Encyclopedism has its roots at least in the Renaissance, as Ann Blair notes: âĂIJit is reasonable to speak of encyclopedic ambition as a central ingredient of the Renaissance obsession with accumulating information.âĂİ Even in 1548, Conrad Gesner began compiling a âĂIJgeneral bibliographyâĂİ 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. Some critics, like Jesuit scholars Francesco Sacchini and Antonio Possevino, recommended an âĂIJaggressively purgedâĂİ rather than universal library, throwing out any redundant or misleading texts. Gesner disagreed, but his reasoning was telling: âĂIJNo 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.âĂİ 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 todayâĂŹs technological language reflects this drive. WikipediaâĂŹs mission is âĂIJto give freely the sum of the worldâĂŹs knowledge to every single person on the planet.âĂİ which is reminiscent of GoogleâĂŹs: âĂIJto organize the world's information and make it universally accessible and useful.âĂİ The worldâĂŹs knowledge, universally accessible, to every person. The goal is impossible; capturing âĂIJthe sum of the worldâĂŹs knowledgeâĂİ is akin to BorgesâĂŹs aleph âĂŞ a point that contains all points âĂŞ or to his one-to-one map of the world.

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 âĂŞ Wikipedia, Google Books, Project Gutenberg, Amazon âĂŞ may have come closer than any previous efforts to capturing the worldâĂŹs 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.

3.4.2 Ted Nelson and Xanadu

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

Nelson was deeply inspired by BushâĂŹs memex, referencing him frequently in presentations and even including the entirety of âĂIJAs We May ThinkâĂİ in his book Literary Machines. Building on BushâĂŹs ideas, Nelson suggested âĂIJzippered listsâĂİ instead of trails, which could be linked or unliked as its creator desired, advancing beyond BushâĂŹs âĂIJprearranged sequences.âĂİ But his biggest development was to reintroduce the global ambition of Otlet into BushâĂŹs associative vision: the idea of a universal, networked, collectively managed hypertext system.

The result would be, as Barnet says, âĂIJlike the web, but much better.âĂİ In NelsonâĂŹs 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 âĂŞ 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.

NelsonâĂŹs idea for Xanadu might resemble Wikipedia; one of WikipediaâĂŹs core tenets is âĂIJNo Original Research: donâĂŹt create anything from scratch, just compile,âĂİ reflecting the principle of NelsonâĂŹs transclusions. 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. NelsonâĂŹs

transclusions are more akin to the Windows shortcut, Mac alias, or Linux softlink. The WebâĂŹs default, on the other hand, is to copy rather than link. Jaron Lanier suggests that copying-not-linking is a vestige of the personal computerâĂŹs origins at Xerox PARC, whose employer was quite literally in the business of copying, and was inherently wary of ideas that bypassed it.

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 webâĂŹs 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 CenterâĂŹs NewsLynx project laments, âĂIJthe inefficiency of one-way links left a hole at the center of the web for a powerful player to step in and play librarian.âĂİ

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. There is a level of hubris in the encyclopedic impulse that Nelson doesnâĂŹt hide. His proposed system is top-down and brittle in certain ways, including rigid security and identification systems. And his proposal for online âĂIJmicropaymentsâĂİ 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.

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. The web is an ephemeral stream, and you wonâĂŹt step into the same one twice; Barnet equates web surfing

with âĂIJchannel surfing.âĂİ Xanadu promised to treat its contents like an archive rather than making us build archives around it. Comparing it to the web raises interesting questions: how much structure, organization, and control should we place on our networked information systems? How much is desirable, and how much is technically and economically feasible? And if we consider the archival capabilities of each, how are they building, sorting, and selecting our information?

A skeletal version of Xanadu (still without its two-way links) was finally released on the web, after more than 50 years of development, in summer 2014. It has joined the myriad archives and knowledge systems embedded inside the web. Many of the later, âĂIJsecond-generationâĂİ hypertext systems were geared towards personal and institutional uses (systems like NoteCards, Guide, WE, or AppleâĂŹs HyperCard). 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.

3.5 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 âĂIJinformation overloadâĂİ 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 networking 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 new intertwingled mass 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.

ArchivesâĂŹ 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 networking of the archive requires knowing your archive âĂŞ 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 webâĂŹs massive structure suggests total comprehensiveness âĂŞ a true universal library âĂŞ 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 networked archives begin to network 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, âĂIJBig 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.âĂİ A full understanding of the structures and challenges in network- and archive-building gives us one view into what boyd and Crawford call the âĂIJmodels of intelligibilityâĂİ and âĂIJinbuilt limitationsâĂİ 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 the archiveâĂŹs size, shape, and scope.

Chapter 4

Publishers and the Archive

In the previous chapters, I have outlined the ways in which the archive, and critical notions of it, has shifted from a fixed and graspable entity to a suite of interconnected parts, constantly shifting and morphing and adapting to new information. As Soumen Chakrabarti puts it, the web is "an active and evolving repository of knowledge," rather than a fixed, bordered entity or set of categories. This chapter hones in specifically on the structure of news archives, and the ways online publishers and legacy news outlets are treating their digital and digitized archives.

The archive has historically been known as "the morgue" to newsrooms, but new technologies and conditions have lead to many recent attempts to reanimate the news archive. Nicole Levy wondered if 2014 is "the year of the legacy media archive" in a *Capital New York* story about *Time* magazine's new "Vault".[1] 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.[2]

The Times' celebrated *Innovation Report*, an internal document leaked to the press in May 2014, emphasizes the archive's potential: "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

1. The Vault can be found at http://time.com/vault.
2. .

"we can be both a daily newsletter and a library." The report suggests that arts and culture content, more likely to be evergreen, could be organized "more by relevance than by publication date," and the topic homepages should be more like guides than wires. The report goes on to enumerate successful experiments with repackaging old content in collections, organized by categories and themes. They even suggest building a CMS widget to create collections– 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.

What makes this moment rich for focusing on digital archives, and their potential value to publishers both old and new? What are these outfits' plans for measuring success? How can legacy media best engage their old archives, and how can digital media prepare itself for the archives of the future?

The archival focus is one of several related movements in today's online news industry. Coupled with trends towards "explainer" and "data" journalism, we see a pattern among some news outlets attempting to evade and reconsider the news cycle's obsession with speed and feeds. News is typically delivered in a stream format, full of boilerplate text that is repeated across every story related to a given theme. By experimenting with new forms of what a news story can be – incorporating lists, photos, videos, graphs, quotes, interactives – explainers and data journalists aim to find a new digital

The resulting media is rich with the potential for re-use. Many of these pieces stay relatively "evergreen", the news world's term for stories that are less directly tied to the news cycle. They also incorporate collections of media that, taken together, form a network or ecosystem of resources that allow for a new view into digital archives, focused on what you're citing as much as what words you're using.

4.1 Introduction: Networking the news

Trends happen in cycles, and the news follows; the news is often repeating itself. Even something seemingly new like data journalism is a holdover of "precision journalism,"

something that few news stories explaining the trend will point a reader to. Sometimes old documents, whether leaked or declassified, can refuel an old story and paint it in a new light. Other times, new cultural events will conjure up the old; when *12 Years a Slave* was released, the New York Times unearthed a story about its namesake, which then went viral on Gawker.

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 digital content, and the ways that news publishers can continuously keep their archives relevant and context at hand.

4.1.1 Paywalls

While many traditional media publications are recognizing the potential role of archives in their shift to digital, they have tended to silo away their archival resources behind paywalls. This is the most literal interpretation of gaining value from one's archive, and it is an understandable choice given the few avenues for them to make money on the web. But paywalls are just one of many possible ways to extract value from archives, and there's more than one way to make a paywall.

One problem with the way that publishers deal with their paywall now is that they have few ways of measuring "successful" archive diving. Like many publishers, Time.com requires a paid subscription in order to access their online Vault. But say a registered user is clicking through the Time Vault. Did she sign up for the subscription just to see the old issues? Is she there on a dedicated research project, or is she just browsing?

Archive paywalls tend to be very limiting; while publishers will occasionally lift the paywall on relevant archival stories, it is usually impossible to even preview an archived story. This prevents interested users from even testing out the interface or getting curious in the archive in the first place. As online publishers experiment with paywall methods, they should keep in mind their older stories.

Some archives are more open than others. The New Yorker opened its archives

for the summer of 2014, free of charge, as they built the paywall system. The archive is now back behind walls, but the summer experiment seemed interesting. How much were users diving into the archives? How were they browsing the archives–through search or serendipity? Did opening the archives encourage people to think and write more about the New Yorker's rich history?

A publisher's archive will often turn up in specific articles geared towards history. At Time.com's Vault, editor/curator Lily Rothman digs out stories and quotes from the history of Time, ranging from historical interests ("Read TIME's Original Review of *The Catcher in the Rye*") to ephemeral oddities ("13 Weirdly Morbid Vintage News Stories"). Time assistant managing editor Samuel Jacobs likened Rothman to a radio D.J., highlighting singles from the archive to entice readers to pay for the whole collection.[3]

Other historic deep dives might occur in weekly columns or "long history" forays by individual journalists. A Sunday Times article, for instance, might take a historic look at a particular person, neighborhood, or community. These projects have a chance to draw attention to the past through curation as well, by drawing out and resurfacing old stories, photographs and statistics.

These projects are a promising start, but they tend to be isolated endeavors, relegated to a single story. Sometimes willfully nostalgic, they can carry an air of "eating your news vegetables." They do not bring the archive fully into dialogue with ongoing events, whether in the research or design process. Journalists don't have easy, seamless access to their publication's past knowledge and institutional memory. Topic pages suffer from lack of organization and explanation. Readers cannot easily dive into old content.

More generally, news' perpetual focus on the *new* and *now* keeps news publishers in the mindset of serving as information providers rather than knowledge repositories. This results in news delivered in story format, through wires, feeds, and streams. These are one-dimensional channels and metaphors, delivering information in a straight line and single direction. But information is more understandable, sus-

3. .

tainable and useful if it's in more than one dimension, structured as a tree, rhizome, or web. Internal news archives have the potential to provide richer, faster resources and connections than the web as a whole, or its indexers (like Google or LexisNexis). This can drive traffic back to the site in turn, and bolster the publisher's authority as an information source. Legacy news publishers are rightly proud of their long histories of stories, photos, and notes; they could do well to show it off rather than hide it away behind paywalls. For publishers, opening access to their past could contribute to solidifying their status in the future.

4.1.2 The stages of news archives

The development of a digital news archive goes through three stages. These mirror Pavlik's stages of development in online newspapers in general; in 1997, Pavlik observed that newspapers started by copying from the print edition, then supplemented the copy with interactive features, then in the third and last stage, started writing copy specifically for the online version of the story. The development of a digital news archive likewise runs through three stages, which I refer to as digitizing, atomizing, and networking.

The first stage for any legacy publisher – anyone who creates physical newspapers or magazines – 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.

Most publishers have reached this stage; 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 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." So it is crucial to remember at

this stage that the digital archive has different potential from its physical counterpart. The physical record should be the starting point of the digital archive, but as a source for linking, rethinking, and remixing, not as a stodgy artifact to be modeled after.

It is also telling that many of the digitization projects, begun decades ago, focused exclusively on salvaging the text. This ignores substantial information in the archive, of course, and speaks to the shortsightedness of many projects aimed at digitizing the past. Images, advertisements, maps and formatting were all lost, and many of them are being re-scanned by publishers, at great expense, in order to capture the details that they ignored years ago.

Historical images are one of the greatest potential sources of engagement and revenue for news archives. It could be trivially easy for some news archives to sell old photographs of historic local events. Some projects, like the New York Times' advertisement tagging project, are aiming specifically at images in attempts to gain more insight from them. It is therefore surprising that images struggle to be monetized.

Researcher Kalev Leetaru took this approach to the Internet Archive. The Internet Archive's OCR software threw out images, and Leetaru's would save whatever it threw out as an image file. He has since put 2.6 million of these Internet Archive images onto Flickr for open use. "They have been focusing on the books as a collection of words," he told the BBC. "This inverts that."[4] Newspaper and journal images provide a richer glimpse of history, and one that might prove more engaging to digital readers than dated text. You also get a sense of the visual language and associations of the time; as any cultural studies scholar can tell you, advertisements provide a revealing window into culture and history.

Old images are an untapped resource for new stories as well; these millions of images on Flickr or in a news archive can be used freely to enrich a story. Networking an archive also involves making its images structured and discoverable.

The digital publishing space is increasingly moving toward a "post-text" world, to use a term Felix Salmon used when he announced joining Fusion. Digitization projects focus on text, not least because it is the easiest to computationally glean

4. **bbc_leetaru**.

insight from, but future efforts to digitize and categorize need to take a holistic, post-text approach to their media assets and web resources as a whole. This will allow the signals and not the formats to determine the best way to tell stories and present information.

The second stage is to *atomize* the archive; to break these scanned pages into their consituent parts. Given the newspaper's inherently hypertextual nature (discussed later), this is a major challenge at any scale. What metadata is worth saving? The text, the subtext, the pictures? The photo or pullquote on the side? Is the image in the center of the page associated with the article on the left, the right, or both?

Newspapers are rich archival documents, because they store both ephemera and history. Journalists sometimes divide these types of news into "stock" and "flow"; the constant stream of information built for *right now*, versus the durable stuff, built to stand the test of time. Newspapers also have advertisements, classifieds, stock quotes, and weather diagrams. 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 articleâĂŹs 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 paperâĂŹs writing style or ad placement change over the course of a decade?) Still others may be hoping to deep-dive into a particular story across various journals. In each case, we can glean information from where and when it was published on the page.

The project of atomizing the archive should take advantage of the signals built into newspapers in the first place, accumulating metadata from its size, shape, and context. An atomized archive should also provide a solid interface for viewing the original in its context. When legacy news publishers refer to a "linked" record in a digital archive, they are refering to the ability to see the original source page; if a text-searchable index is connected to a scan of the original, it is considered "linked." Some publishers do not even have linked records for their entire archive. This makes

context difficult to grasp for interested researchers.

But legacy publishers with atomized archives also often have clunky ways to access them. The Boston Globe uses Methode, a content management system that does not allow mass printing of articles, does not carry the publication date with it as an article is printed. These small errors add up, and no one likes to use the archive as a result, instead relying on LexisNexis. Atomizing the archive is not helpful without a good interface.

The final stage is to *network* the archive, which few publishers have fully done; indeed, as the drive towards encyclopedism implies, it is likely that this project could never be *fully* done. A better question is, where is the best place to start in networking the archive? How can we balance value with manageability?

Networking the archive requires tagging, annotating, and connecting the items together. This includes both explicit and implicit references, and both manual and automatic means. Whether an editor makes a note on a specific article, or a bot tags an old image with "Poughkeepsie, N.Y.," each act can help in the networking of the archive.

The end goal is to bypass the need for dedicated search to surface archival content. We want to use a "push" rather than "pull" method of archival materials. This is as true for reporters and dedicated researchers as it is for casual browsers and fans. A user doesn't always know exactly what he or she wants, and a networked archive can work with a user to surface it.

The "networked archive" borrows from, but is distinct from the notion of "networked journalism." A term popularized by Jeff Jarvis to refer to the growing citizen journalism movement, networked journalism has also led to Jarvis's succinct motto of "do what you do best, link to the rest."[5]. Building on this notion, Charlie Beckett posits that linking between sources leads to editorial diversity, connectivity and interactivity, and relevance.[6]

A networked archive borrows from this notion but turns the conversation inward;

5. jarvis
6.

links that point inward are vastly different from those that point out. Few news organizations have truly embraced networked journalism, and even fewer (if any) have considered a networked archive.

However, as Juliette De Maeyer points out, there is increased interest among news organizations in the power of linking.

4.2 Context in context

4.2.1 The Scoop Effect

The time is ripe for news and history – content and context, feeds and archive – to collide. News outlets have long obsessed over the "scoop", being the first to break a story, and indeed these breaking stories still drive a great deal of traffic. But publishers are increasingly scooped in turn; stories break immediately on social media, rather than the next morning in the newspaper. Emily Bell argues that social media and "these super platforms ARE the free press, taking over many of the functions of the mainstream media. Social networks are now attracting the same pressures and challenges at a much larger scale that journalism and civic media has wrestled with for years." This is having a profound effect, of course, on publishers; where does the media fit in here? But it is especially affecting the research process, skills and news lifecycle for both journalists and editors/strategists.

For journalists, it has increasingly destroyed the stereotypical image of the reporter with a notepad in city hall. The increasingly real-time nature of scooping has led to reporters scouring Twitter as much as being in the field; even communication with sources increasingly occurs via email or tweet. The increasing presence of "explainer" and "data journalism" likewise speaks to this need; reporters must to wade through massive amounts of information in fast, efficient ways in order to uncover possible news stories, which requires very strong digital research skills. These are skills that librarians have been practicing for centuries, and a well-organized and linked archive can help reporters immensely with this research process. The data journalists thus

emerges as an amalgam of reporter and librarian.

For editors and newsroom strategists, it has shifted the role of the journalist and the news publisher to explainer, data-gatherer, and context-provider. Picture a newsworthy event occurring as the epicenter, and the reporting that occurs around it as a set of concentric circles around the event. Towards the center, one might find tweets, wire reports, and quick announcements. At the edge, there are longform pieces, explainers, multimedia work and data-oriented stories that help draw immediate events into larger phenomena. While the scoop remains crucial and breaking news draws traffic, news outlets can no longer serve as raw information providers, with no context. For a publisher to stand out, it is crucial to bring ongoing stories into a larger dialogue and conversation.

4.2.2 Explainers

This focus is not limited to legacy media, as the rise of "explainer journalism" and context-based reporting emerges as the other side of this coin. *The Nation*'s editor and publisher Katrina vanden Heuvel suggests that "a clever use of archives is kind of an explainer 2.0."[7] The goal is to provide knowledge, not news.

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

Rosen argues that while journalists are paid to report the news and not explain it, they should also be giving background and context to larger stories. Journalists "donâĂŹt do a very good job of talking about the beginning and what got us to this point where it became news," according to Alex Blumburg of *This American Life*.

7. .

Even the occasional explainer that gets it right ends up in the flow of the same old information; Rosen argues that explainers like David Leonhardt's credit crisis piece in the New York Times "should have been a tool in the sidebar of every news story the Times did about the mortgage mess." The little "what's this?" link is "not about web design. That's a whole new category in journalism that I fear we do not understand at all."

Rosen also points out that such explainers are helpful for other reporters as well as the public, influencing news and information flow across the pipeline. A Times explainer, for instance, can reach a reporter who is informed by it as he or she interviews local officials. Calling it a "scaffold of understanding," Rosen suggests that we "start with clueless journalists" in the path towards providing context, and went on to create explainthis.org, for people to admit what they don't know to journalists who are "standing by."[8]

Explainthis.org, now defunct, was like a library reference desk, staffed by the public and monitored by journalists. A peer of StackOverflow and ancestor to Quora, it is organized around questions rather than topics, discussed by the public and monitored by journalists. It requires someone to be curious enough to ask the question, however. Rosen touts the ability of explainers to generate interest in a topic, but here we're already expected to be interested.

At a South by Southwest panel in 2010 called "Future of Context," Rosen outlined the reasons explanation is needed and why it wasn't taking off. He cited both design and institutional problems; the prestige and real-time excitement of breaking (rather than explaining) news, as well as the explainer format getting lost in the shuffle of other news items.[9] Metrics like clicking, watching, and even spending time on a site are not measuring the level of understanding or knowledge gained.

The panel opened with NPR's Matt Thompson, owner of former contextual news blog newsless.org, arguing that we need more "systemic information, not episodic info." Systemic information could include lists, charts, and maps that stay valu-

8. **rosen_2008**.
9. **rosen_2010**.

able well after the episodic news is irrelevant. Tristan Harris of Apture says, "my background is computer science. You never do work that you can't re-use." He suggested an "object-oriented" approach to journalism with an eye towards sustainable, continuously updating tools and widgets that keep a reader informed. News is organized around stories rather than objects, resulting in streams rather than systems. A systemic, object-oriented approach to news places the context in the center.

The panel concluded with Harris suggesting a wiki-like approach to journalism, which a big news organization like the New York Times would have the power to sustain. When Kramer asked "how is this more than links?" Thompson replied "Links can be part of it." So can wikis, embeds, collections, and related articles.

"The context should be the foundation. The systemic stuff should be what you can access first. The episodic stuff is what should be the more info. We âĂIJghettoizeâĂİ topics pages on our sites, by creating a topics section. When the public just finds just a random collection of links on a so-called topics page, âĂIJthe quest for context everywhere is set back,âĂİ Thompson argues. What would a site look like if it were structured around systems instead of stories?

Journalists may think, weâĂŹre doing so much and now you want to provide context!? Think like an engineer. Make it an imperative to do work you can reuse to provide context. You can use that subduction plates info graphic again and again with every story you write about earthquakes. ItâĂŹs redefining the notion of âĂIJtodayâĂİ value. YouâĂŹre writing something TODAY thatâĂŹs only appending something thatâĂŹs already valuable. Engineers donâĂŹt do work they canâĂŹt reuse. Do work you can use next time."

Chuck Peters, CEO of the Cedar Rapids Gazette:

"I canâĂŹt see providing that context without changing how we create information in the first instance. Any factual element (photo, incident, quote, data, etc.) can be relevant to numerous contextual narratives. So each of those elements needs to both âĂIJstand on its ownâĂİ and be tagged with as many potential relationships as possible... We usually create information today in locked-down packaged articles, which block the easy flow of the elements between and among narratives."

Finally, in an article entitled "Swimming lessons for journalists," PBS's Amy Gahran asserts, "todayâĂŹs journalists can âĂŤ and probably should âĂŤ consciously shift away from jobs that revolve around content creation (producing packaged âĂIJstoriesâĂİ) and toward providing layers of journalistic insight and context on top of content created by others (including public information)."

4.2.3 Vox

At the start of 2014, Ezra Klein left his position at the head of Washington Post's Wonkblog to start Vox, a news site that aimed to make context a first-class citizen of web journalism. Vox's mission: "to create a site that's as good at explaining the world as it is at reporting on it."[10] Vox hopes to take a step back from the immediate news event and place it in a larger phenomenon. Taking the long view on stories also gives them an eye towards sustainability; Vox's topics are built around what they call "card stacks."[11] Cards have titles like "Everything you need to know about marijuana legalization," or "9 facts about the Eurozone crisis," and each card is divided into question-driven subsections like "What is marijuana decriminalization?" Readers can navigate sequentially, or dive from question to question, going through Vox's explanations and photos. The final option is always the same: the "Explore" button takes the reader back to the top of the stacks.

Vox's card stacks house a growing and morphing repository of knowledge. They are a public archive, like a Wiki but with more authorship intact. At the end of each card, Vox offers a link to email the author/curator of the card stack. For Vox reporters, starting a stack is also a pledge to maintain it. Vox also give a summary of changes made to the card (full versioning, they say, is coming soon).

The goal is not to replicate Wikipedia, but more like a wiki "written by one person with a little attitude," as Vox co-founder Melissa Bell put it. It's obeying the rules of journalism rather than "no original research." Klein has Wikipedia in his sights, suggesting in the New Yorker that "I think it's weird that the news cedes so much

10. .
11. See http://www.vox.com/cardstacks.

ground to Wikipedia. That isn't true in other informational sectors." By combining incremental news with an evolving repository, Klein hopes to gain the best of both worlds: "the card stacks add value to the news coverage. And the news coverage creates curiosity that leads people to the card stacks." This follows Rosen's idea that explaining the news can generate future interest in incremental updates. For Klein, "the biggest source of waste is everything the journalist has written before today."[12]

For Klein, there is a distinct need, like Rosen saw, for a website that takes a step back and explains the news. In his words, "The more folks in the media feel like it's beneath them to answer questions like "What is marijuana?" or "What is Ukraine?" the more we don't have to compete with them."

Vox has accompanied other "explainer" and data-focused websites, like Nate Silver's FiveThirtyEight, and The New York Times' The Upshot. Soon after Vox's launch, Craig Silverman wrote "Why Vox (and other news orgs) could use a librarian," suggesting that Vox had "a huge challenge, due to the rapid decay of facts.".[13] Some of these facts may not even be obviously newsworthy, such as if an academic research paper changes a fact in an explainer on Alzheimer's care. Who is going to keep everything up to date? "Someone at Vox is going to need to know which card stacks to update when," and how to keep the explainers updated with minimal maintenance.

4.2.4 News Libraries

While someone needs to maintain all of these card stacks, it may not be a librarian. There are fewer and fewer librarians in newsrooms, which places their responsibilities increasingly on the reporter instead. Amy Disch, chair of the Special Libraries Association News Division, speaks to the traditional division of skills between reporter and librarian in the newsroom: "We can find the information in a lot less time because we know how to drill down in a database. We know good sources to go to where you can quickly find information, so we can cut a lot of time for [reporters] and leave them

12. **nyt_vox_melding**.
13. .

to do what they do best, which is interviewing and writing. I have my specialty, and they have theirs."[14]

Most legacy newsrooms have a library, but their librarians are "a dying breed," with librarians getting laid off from a variety of institutions after the recession. Over 250 news librarians lost their jobs in the U.S. from 2007 to 2010, and membership in the Special Libraries Association News Division has steadily dwindled. Some news libraries and research centers have been completely shut down, outsourced to vendors like LexisNexis.[15] Not only do their reporters' research abilities suffer, they cease to be a steady provider of useful and updated information for readers.

At a 2001 summit on news libraries, futurist Arthur Harkins suggested that in order to stay relevant, news librarians should "leave the information management functions to automation" and instead focus on "the ability to put knowledge into context and to synthesize information." The librarians focused on solutions like structuring incoming stories, helping merge mixed media operations and create new revenue opportunities from older assets. teaching journalists the necessary research and technical skills. Finally, the librarian's task is largely to tag; to structure stories for future discoverability and reuse, by both journalists and the public. Some news libraries, like the Boston Globe, used to manually tag their stories, but no longer do.

Today, many of these skills are expected of new journalists at the outset. Such reporters arrive armed with years of internet research skills, though some of it with Google over specialized databases. Leslie Norman, former librarian at the Wall Street Journal, suggested, "I see the news library as it once existed as probably dying, but in many newspapers, it's evolved into something else."

Although news libraries are a dying breed, some libraries and cultural heritage organizations are making promising digital inroads into news. Old newspapers provide a rich archive of both historic resources and incidental ephemera like sports scores, weather reports, advertisements and small human interest stories. This gives historians a glimpse of a day, with the major phenomena of the day mixed in with everyday

14. .
15. .

events.

Most of these projects are aimed towards the serious researcher, but they also point towards ways to engage casual browsers and fans of history. The National Library of Australia's Trove collection features 370 million resources; primarily, Australian newspapers ranging from 1803 to 1954.[16] Their API allows programmatic access, which in turn leads to the TroveNewsBot, an irreverent Twitter bot that can search the collection and yield a personalized result. Similarly, the Digital Public Library of America's DPLA Bot and British Library's Mechanical Curator both post random resources from their collection, aiming to inject a serendipitous sense of the past into the present.

Newspapers would do well to merge increasingly with digital cultural heritage institutions and library APIs.

or they argue that "the software newsrooms have adopted in the digital age has too often reinforced a workflow built around the old medium."

4.3 The structure of stories

News 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 newspaper's front page reads like a website homepage, 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 po-

16. http://help.nla.gov.au/trove/building-with-trove/api

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

However, 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. Citation is "as old as written language itself,"[17] and it is *itself* a language, with its own idioms, syntaxes and exceptions. The footnote has its limitations (as discussed earlier somewhere), but newspapers don't even get footnotes. So while a newspaper's layout and seriality might afford a news story more insight than an academic article, its flatness and lack of citation conventions lead to limitations in computationally gathering insight from a newspaper archive.

Digital news publishing has the potential to change the conversation through the networks it creates; its hyperlinks, its embedded media, and the media that links *to* it. This can form "hyperlink-induced communities."[18]. Hyperlinks allow for a new standard of citation, reference, and context provision for news. At smaller scales, 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 also for creating a traceable dialogue that can improve a story's discoverability in the future.

Much of the work done on mining digital archives has focused on Natural Language Processing (NLP), or the science of converting human languages to machine language and back. This is a very fruitful avenue for research and archival enlivenment, but it is only one stone to unturn; here I might propose developing a Citation Language Processing (CLP) system.

The use of citation analysis to determine impact, weight, or ranking is an old

17. Chakrabarti
18. Chakrabarti

practice, especially for the sciences. Known as bibliometry, the practice has a long history with strong conventions, which I will dive into more closely in the following chapter. The online version is sometimes known as "webometrics."

But even a Citation Language Processor requires a standardized Citation Language. The closest that we have come to such standards involves the Semantic Web and, specifically, the rNews initiative. But our work here is easier, because it is presuming a single publisher within one domain; the ability to change its own language, if not the language of others, enables vastly improved inlinking and recommendation from one's own archive.

4.3.1 Ontologies and tags

In technical terms, stories are usually objects in a database that have associated text, images and tags. Stories contain multitudes; they are an agglomeration of multimedia objects. Any link from one story to another must then refer to the story as a whole, rather than a salient part of that story (whether it be a certain paragraph or an interactive chart within it).

What is the atomic unit of information for news? It has traditionally been the article in a feed or stream, but Apture's Tristan Harris suggests that "Because journalism is structured on the article, it doesn't accomodate the full extent of information we need." An article tends to pull paragraphs from one source, photos and charts from another. The news app Circa organizes its content around "atoms" of news: single facts, quotations, statistics, and images that can be reaggregated and remixed as needed. Systems like Vox and Circa aim to create a baseline repository to build upon rather than recreate from scratch every time.

This approach rethinks how we organize news items and structure stories. A "story" need not be a fresh new original piece of reporting every time; instead it can be a collection, ecosystem or dialogue of items. A journalist can still create, but also curate, collect, and contextualize. Thinking of a story as a collection or mash-up offers a new framework of a story as a highly linked entity, one that can start to organize itself.

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

Google's success rides on their reliance on *crawling* in the weeds rather than *categorizing* from on high.

Shirky shows that "you could have a lot of links. You don't have to have just a few links, you could have a whole lot of links," and if you have a lot of links, "you don't need the hierarchy anymore. There is no shelf. There is no file system. The links alone are enough." For Shirky, links can be formed by the tags that users create; tags are crucial in organizing the web. "Tags are important mainly for what they leave out. By forgoing formal classification, tags enable a huge amount of user-produced organizational value, at vanishingly small cost."

Some studies show that even at web scale, with users tagging items for personal and idiosyncratic reasons, distinct and simple patterns emerge that allow for collaborative classification..[19]

The New York Times also sees tagging as core to their business, and the main reason that they have remained the "paper of record" for decades. In short, well-structured stories and tags have helped the Times remain a library, information hub and general authority on contextual information. This can be true for all publishers, both legacy and digital, and it is surely at the heart of what Vox aims to build with its card stacks. But the Times' Innovation Report sees them falling behind, as they adhere to the needs of the legacy Times Index rather than the modern affordances of

19. **catutto_semiotic_dynamics**.

digital search. The Innovation Report notes that it took seven years for the Times to start tagging stories "September 11." Evan Sandhaus suggests that they are organizing their stories counter to the way people navigate news, and the Innovation Report proposes a new set of tags – terms like "Timeliness," "Story tone," and "Story threads" (such as "crisis in Ukraine," which they admit would require them to "Better organize our archives").

4.3.2 Linked tags

But as Stijn Debrouwere points out, even "tags don't cut it," as the title of his blog post says. "Each story could function as part of a web of knowledge around a certain topic, but it doesn't." Tags – which are often inconsistent, outdated, and stale – provide the only window into content at the metadata level. "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." The reason tags don't cut it relates to the level of *intersubjectivity* that newsrooms are dealing with.

Debrouwere suggests using vocabularies; set people, places, organizations, events and themes. He also advocates for relationships over tags, which borrow 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." Finally, Debrouwere advocates for "entities, not labels." This point is the most important; "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."

In short, as Debrouwere suggests, "we have the ability to transform those rudimentary link dumps into valuable landing pages and content hubs that collect all the content for a person, an organization or an event." Relationship cascades, synonyms

and homonyms. Debrouwere is suggesting returning a level of structure to the open, massively linked and networked web. Why would this be useful? Wasn't the web averse to strict taxonomies and structures?

The answer is a question of scope; when deciding what to do with archives, it is crucial to remember the archive's size, shape,and scope. At a massive scale like those of Twitter or Flickr, tags work well because there are still more than enough matches to go through. At an individual scale like someone's blog, tags can work because there are only a handful of tags and one person to tag them. But publishing archives are closer to the intersubjective scale, between the personal and universal. At this scope, tags can be a challenge.

At large scales, tagging resembles a form of *collaborative filtering*, which looks to the properties of users, their preferences and behaviors, in order to categorize and ultimately recommend items to others. Tagging as publishers perform it, however, is not collaborative in the same way. When a user bookmarks and tags an article or photo on the web, he or she is doing so for personal reference; when an editor tags a news story to prime it for publication, it is for the article's discoverability. In the former case, tags could be "Articles I read last night" rather than the standard categorical tags.

Shirky likewise discusses when ontological classification works well, suggesting that despite the success of tags on the web, a small corpus with expert catalogers should be organized around a taxonomy. But the web – with its large corpus, lack of categories, and amateur catalogers – is not a good fit for ontologies. News archives fall in between these extremes, but as they are increasingly digitized, atomized, and networked, they are increasingly moving towards a weblike structure.

In order to prove worthwhile, newsroom archives must afford greater value than a simple Google or LexisNexis search. While these services provide text search, a well-linked archive can include images, videos, charts, maps, statistics, quotations, comments or annotations. The stories that result can be a combination of these, referencing an ecosystem of media that already link, reference, and embed one another. Publishers can also leverage smart entity recognition and linked data tools to aid in

automatic and rich tagging. While newsroom librarians are increasingly disappearing, and publishers cannot upkeep ongoing stories in their archive, why is text search the only way to search for stories?

Debrouwere envisions a tagging system where a tag can double as a full card or widget, linked in turn to other cards and widgets in a network of knowledge. Coupled with more automated, dynamic, and context-aware tagging methods, well-structured news archives could become sustainable repositories of knowledge in their own right, turning publishers into information providers and authorities on the level of libraries and information technology companies. There is value in this well beyond charging a reader per-article for archive access.

4.3.3 Promising starts

Some initiatives and organizations are taking promising steps towards linking their archives. The New York Times RD lab recently released a tool called Madison that aims to crowdsource insight about the ads in old Times issues. Starting with the 1950s, the tool asks users questions about the ads that they see, with the aim of adding structured metadata to otherwise difficult-to-parse texts. The team even released the underlying crowdsource platform as an open-source project, allowing others to run their own crowdsourcing endeavors. This is a promising opportunity for publishers who have the scale and user base required for a crowdsourcing project.

Vox's Chorus platform is another promising endeavor in structuring archives. Beyond the sustainable mind to data that is Vox's card stacks, Chorus helps reporters better structure their stories, from adding smart tags to media and widgets.

Some promising endeavors are coming from small starts. Take a collection of already-related stories and build nuanced links between them. Or hone in on a single aspect of a linked collection – for instance, geographic data – and aim to structure that first. Projects like MapCake – which automatically creates maps out of structured location tags – show how effective it can be to focus efforts on structuring one media type. Sites like National Geographic are likewise well primed for beautiful photo archives, which can be watermarked and travel with articles and social media posts.

It makes sense for such sites to focus on structuring their photo archives first and foremost. Radio and podcast-oriented sites could do well to take advantage of services like PopUp Archive

Finally, there are many digital libraries beginning to offer resources and services; quotes, images, and videos. These have the potential to expand the borders of a news archive. The Digital Public Library of America, JSTOR, Flickr, PopUp Archive; all of these sites have useful resources for both reporters and readers. Many also have APIs that would allow easy integration with existing content discovery systems.

4.3.4 Next steps

Chapter 5

Linking the News

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.

5.0.5 From tags to links

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.

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

1. .

continuously animated and relevant.

5.1 Mapping Links

5.2 Legacy media and links

5.3 Digital media and links

5.4 Honing in a couple topics ("controversies")

5.5 Automated NER and linked data potential

Chapter 6

Technologies and Tools