0.2 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.

0.2.1 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 journalism 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.

0.2.2 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.

0.2.3 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 the web itself.

0.3 The Fields