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

© Massachusetts Institute of Technology 2015. All rights reserved.

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, all-encompassing subject; I keep running into the very problems I want to address. I've encountered potentially perfect 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 resources and ideas, which confirms that I'm far from alone. Finally, I've spent sleepless nights trying to organize my own 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. New material arrives faster than we can contain it. For centuries, librarians and archivists have collected and sorted out our history and access to the past. Now we've turned to newer and untested proxies and heuristics 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 only ones doing it.

Where did this new paradigm come from? There's no doubt it's partly technological; there are new challenges and affordances. The archive has exploded and networked to an unmanageable scale. Machines help to sort out the results, but this has pushed the power of archivist to the technically capable, who have in turn changed the way that archives work. Instead of speaking about archives, Google and Facebook talk of networks; instead of categories, they rely on links.

The link is an ideally situated object 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 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 do more than categorize. They measure importance and impact. *In fact, this might be the link's primary function. Only secondarily does it create a category.*

I first recognized the power of links as a software engineer and backend web developer. I've worked with a variety of organizations on news and event curation and monitoring applications, using many different programming languages and frameworks; my job, essentially, has been Link Wrangler. I corral news articles, tweets, and events to improve classification and recommendation systems. In the process, I've grown frustrated with the URL; it's treated as a unique identifier and as information's atomic unit, when it's truly neither.

The URL is not unique The same article or event listing reappears under dozens of URLs, and any attempts to find a "canonical URL" are expensive and inconsistent. Sometimes – like in the case of a wire service that gets aggregated by several publishers – there's no singular home for it.

The URL is not atomic URLs point to multitudes of resources, or none at all. Links have text, pictures, videos, audio, other links, and annotations on all of the above. They can change depending on who's asking for them and when they're asking. They can give your computer a virus. They can (and often do) cause money to change hands between unseen actors. They can open in a new tab or window, or open your email client. In other words, the *Uniform Resource Locator* is not uniform, nor is it necessarily a resource.

Most of all, information travels in both hard and soft ways; it gets copied, remixed, and recycled. Sometimes it's explicitly cited, and sometimes the connection is not

[1] There are centers in a network, though.

even obvious.

Beyond the URL, I measured the link itself. Links come in many forms and contexts. It is technically easiest to treat URLs as unique and links as equal, but the reality was more complicated.

In this thesis I aim to unpack what links do – and what they fail to do – for 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 folds in on itself. I'm speaking especially to the bridge between news and history. My goal is to speak to news and media organizations who hope to enact and enliven their own archives, and on the other hand, towards libraries and archives endeavoring to inject some of their historical resources and context to current events and social issues.

The hyperlink might seem like an innocuous and even trivial object to study. Conversely, it could feel overly academic and only tangential to the news/archive divide that I hope to spin together. But the link is exciting to me precisely because it, by definition, points in all other directions. In focusing on the link, I hope 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. One could say that I've taken the "comparative" part of Comparative Media Studies too seriously.

We all make links. Whenever we blog about, email, tweet, like, or search 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 all make links, and links make history. But we can't understand or determine the rules under which that link has influence. We can't opt out.

Chapter 1

Introduction

Bibliography