DSHR's Blog

I'm David Rosenthal, and this is a place to discuss the work I'm doing in Digital Preservation.

Tuesday, March 13, 2018

The "Grand Challenges" of Curation and Preservation

I'm preparing for a meeting next week at the MIT Library on the "Grand Challenges" of digital curation and preservation. MIT, and in particular their library and press, have a commendable tradition of openness, so I've decided to post my input rather than submit it privately. My version of the challenges is below the fold.

The challenges of curating and preserving digital information are fundamentally economic. These are things we know how to do, but we don't know how to do at a scale commensurate with the problem we face. Even very optimistic estimates are that less than half of the academic papers or web pages that should be are preserved. The selection of, and metadata for, those that are is the subject of incisive criticism.

Assuming that the goal is for the digital information that is curated and preserved to be open access, two economic analyses are relevant:

- Cameron Neylon's *Sustaining Scholarly Infrastructures through Collective Action: The Lessons that Olson can Teach us* applies:

  > First, it illustrates that the problems of sustainability are not merely ones of finance but of political economy, which means that focusing purely on financial sustainability in the absence of considering governance principles and community is the wrong approach. The second key insight this approach yields is that the size of the community supported by an infrastructure is a critical parameter. ... Olson describes three different size ranges of groups: those that are small enough to reach an agreement to provide the collective benefit; those that are too large to do so; and those that lie in the middle, where the collective good may be provided but at a level which is below optimal.

  Only in small groups, whether of individuals, institutions, or governments are social pressures strong enough to prevent free-riding.
- Brian Arthur's *Increasing Returns and Path Dependence in the Economy* from 1994, which explains how the increasing returns to scale and network effects endemic to digital networks ensure that there will be one, or at most two, winners in each niche in the ecosystem.

But at a more granular level the economic challenges of curation and preservation are quite different.

Curation

The two main aspects of curation in this space are *selection*, and adding value by *enhancing metadata*, both human activities that don't scale. The Internet Archive's non-selective approach to collection is the foundation of its success. As I wrote back in 2013:

> What matters isn't the perfection of a collection, but the usefulness of a collection. Digital preservation purists may scorn the Internet Archive, but as I write this post Alexa ranks archive.org the 167th most used site on the Internet. For comparison, the Library of Congress is currently the 4,212st ranked site (and is up despite the shutdown), the Bibliothèque Nationale

Blog Rules

DSHR

DSHR in ANWR

Recent Comments

David.
AMD allegedly has its own Spectre-like security flaws by Alfred Ng at C|Net reports that:"CTS-Labs, a security comp…More

David.
Tom Simonite at Wired writes The Decentralized Internet Is Here, With Some Glitches:"David Pakman, a partner with v…More

David.
"Last Wednesday, the risks posed by Internet-facing memcached processes took on a new colour, when security vendor …More

David.
In YouTube, the Great Radicalizer Zeynep Tufekci writes:"It seems as if you are never "hard core" enough for YouTub…More

David.
"I believe more than $3 billion of all cryptoassets' volume is fabricated, and ... OKex, #1 exchange rated by volum…More

> de France is ranked 16,274 and the British Library is ranked 29.498. Little-used collections, such as dark archives, post-cancellation only archives, and access-restricted copyright deposit collections are all at much greater economic risk in the long term than widely used sites such as the Internet Archive.

Although the process of adding *technical* metadata has been automated, the value of doing so is negligible and even potentially negative. The value of other types of metadata, such as provenance and bibliography, can be significant, but it is doubtful that the cost of staff time to generate them is justified. It can be argued that machine learning could automate these processes. Human processes are biased, but the humans involved are far more diverse than the small set of geeks whose biases would be built in to the machine learning algorithms.

Via its "Save Page Now" feature, the Internet Archive crowd-sources part of its selection policy. Crowd-sourcing selection would be a good way to scale the selection part of curation. Alas, there are risks here too, as the bots and troll farms rampant in social media show. Two years ago Kalev Leetaru wrote:

> **of the top 15 websites with the most snapshots taken by the Archive thus far this year, one is an alleged former movie pirating site, one is a Hawaiian hotel, two are pornography sites and five are online shopping sites.** The second-most snapshotted homepage is of a Russian autoparts website and the eighth-most-snapshotted site is a parts supplier for trampolines.

Just as the cat videos don't impair YouTube's pedagogy, these facts don't impair the Internet Archive's usefulness. Its highly automated collection process may collect a lot of unimportant stuff, but it is the best we have at collecting the "Web at large". But for crowd-sourcing to be effective, it has to happen on a single platform.

Preservation

Preservation happens in three phases; ingest, preservation and dissemination:

- Ingest is the most expensive phase, and it is subject to strong economies of scale, both in terms of infrastructure (the Internet Archive sustains about 20Gb/s inbound), and in terms of staffing. Ingest at this scale would be technically challenging even if Web technology stood still. The rapid evolution of the Web from quasi-static hyperlinked pages to a JavaScript programming environment requires a highly skilled, fast-moving team to keep up.
- Preservation is less expensive, primarily because it is less staff-intensive. But it is subject to even stronger economies of scale, which have powered "the cloud" to displace on-premise storage at scales below about 10PB. The economic and business risks of cloud storage make it an inappropriate way to preserve our cultural and scientific heritage.
- Dissemination is typically the cheapest phase, at least at scales below the Internet Archive's 40Gb/s outbound. But that is because preserved collections other than the Internet Archive's get relatively little use, because network effects drive traffic to the dominant player in the niche (see Google, Facebook, Netflix, etc.).

Conclusion

Displacing the Internet Archive as the go-to resource for preserved digital content would require not merely building a bigger, better curated collection, but also capturing the mind-share that has kept it in the top 300 sites on the Web over decades. It isn't going to happen at a single institution, since the dollars that it would take would be far more effective in increasing access to preserved data at the Internet Archive.

Could a collaboration among many institutions using decentralized Web technology displace the Internet Archive? Chelsea Barabas, Neha Narula and Ethan Zuckerman's *Defending Internet Freedom through Decentralization* from last August surveyed the various efforts to decentralize the Web and, despite their efforts at optimism, showed that antitrust is the only feasible way to displace the FAANGs (Facebook, Amazon, Apple, Netflix, Google). Herbert Van de Sompel's Paul Evan Peters award lecture entitled *Scholarly Communication: Deconstruct and Decentralize?* describes a potential future decentralized system of scholarly communication built on existing Web protocols. But even he prefaces the dream with a caveat that the future he describes "will most

likely never exist", and I wrote a long exploration of the reasons for such skepticism.

Even if decentralized Web technology were ready for prime time (it isn't), it is inherently less cost-effective at delivering capacity and performance than centralized systems. Since the fundamental problem we're facing is lack of funds, this isn't a good direction to take.

Instead we should build upon the centralized system we have. The Grand Challenge of digital curation and preservation is to find ways to sustain and enhance the Internet Archive's capacities for crowd-sourced curation, at-scale preservation, and openly-accessible dissemination. An additional $10M/yr would be a big step in the right direction, but running ads against the preserved content, or mining cryptocurrencies in visitors' browsers aren't the ways to get it.

Of course, this all assumes that the Grand Challenge isn't to find ways to curate and preserve a Web locked up by Digital Rights Management, as specified by Tim Berners-Lee and the W3C.

Posted by David. at 8:00 AM
Labels: crowdfunding, digital preservation, distributed web, DRM, scholarly communication, web archiving

No comments:

Post a Comment

Links to this post

Create a Link

Home Older Post

Subscribe to: Post Comments (Atom)