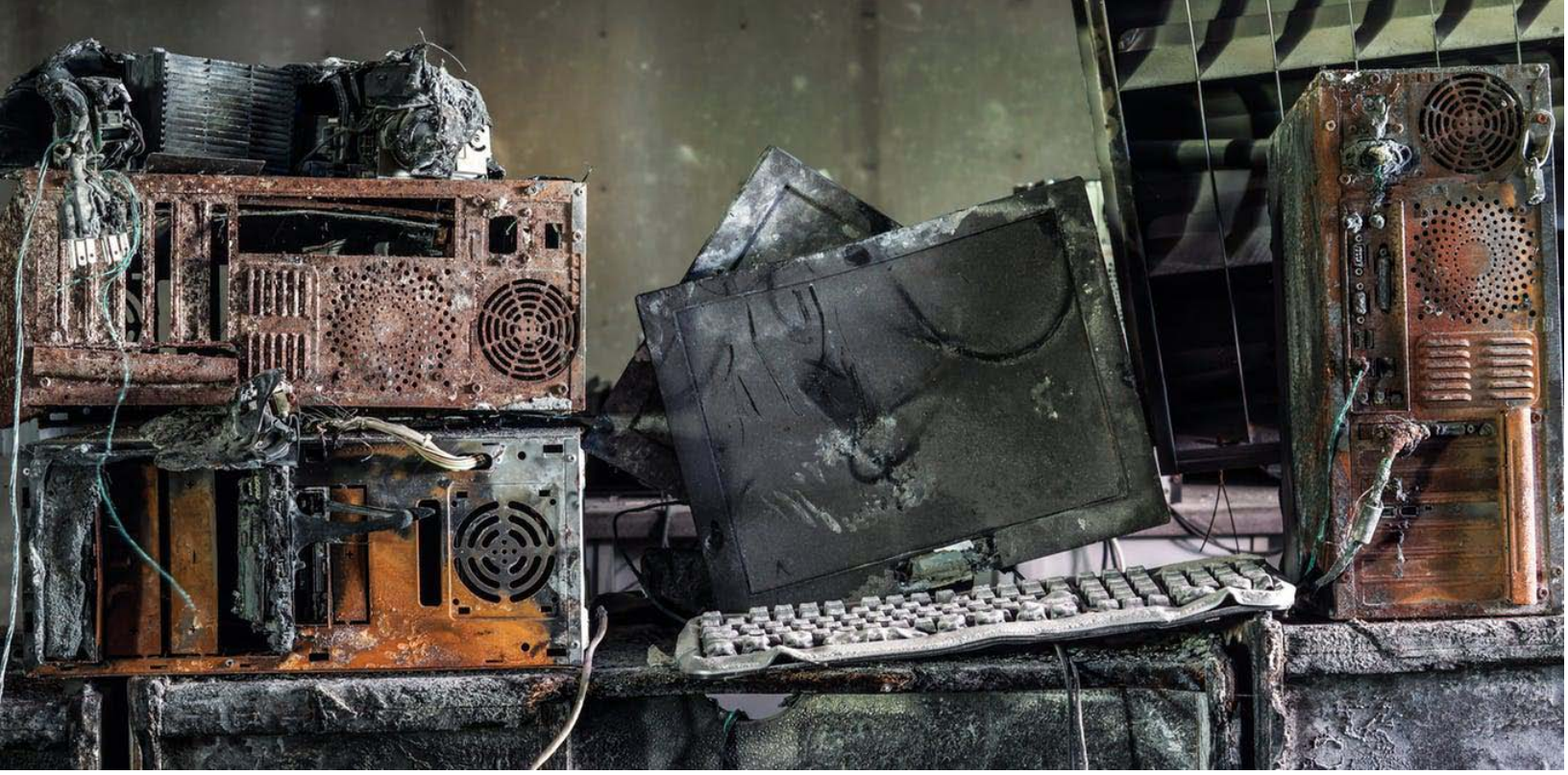
5-15-2019

Your internet data is rotting

Paul Royster
University of Nebraska-Lincoln, proyster@unl.edu



Your internet data is rotting

The Conversation
May 15, 2019, 6.47am EDT

The internet is growing, but old information continues to disappear daily.

Paul Royster, Coordinator of Scholarly Communications, University of Nebraska-Lincoln

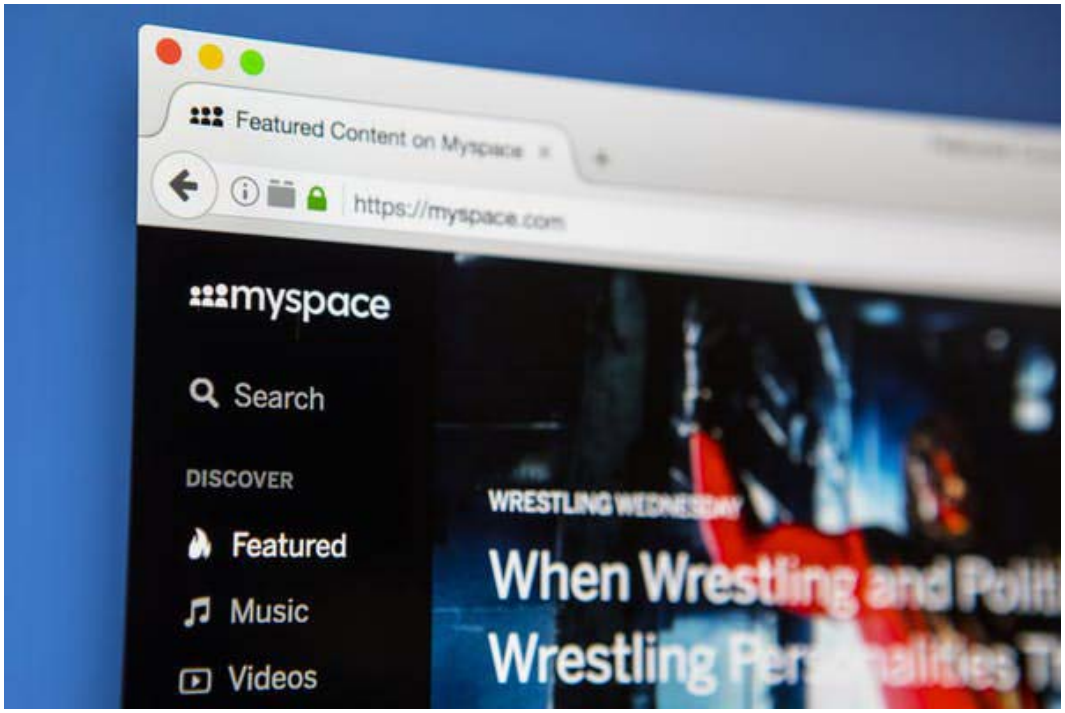
Disclosure statement: Paul Royster does not work for, consult, own shares in or receive funding from any company or organization that would benefit from this article, and has disclosed no relevant affiliations beyond their academic appointment.

Partners: University of Nebraska-Lincoln provides funding as a member of The Conversation US.

Many MySpace users were dismayed to discover earlier this year that the social media platform lost 50 million files uploaded between 2003 and 2015.

The failure of MySpace to care for and preserve its users' content should serve as a reminder that relying on free third-party services can be risky.

MySpace has probably preserved the users' data; it just lost their content. The data was valuable to MySpace; the users' content less so.



What happened to MySpace

MySpace is a social networking media site where performers could upload music or other content for access and distribution to its user community. It has always been a free site, with revenues coming from ads and programming that targets users for specific products.

Formed in 2003 in imitation of the social gaming site Friendster, MySpace grew rapidly and was purchased by Rupert Murdoch's News Corporation in 2005. By 2008, MySpace was the leading social networking site, valued at one time at US$12 billion But it declined in popularity – thanks to an overprevalence of ads, concerns about exposure of minors to sexual content and other issues. In 2011, News Corporation sold MySpace to Specific Media, who sold it again in 2016 to Time Inc., which was in turn bought by the Meredith Corporation in 2018.

So the company went through three changes in ownership over a 12-year period, and saw revenues and membership drop precipitously over that time. One sale might be fine, but three sales over short term suggests to me a troubled business that was not in a good position to watch over others' intellectual property.

Anyone using MySpace as a storage service who did not have alternate backup is simply out of luck. You left your intellectual property sitting beside the information superhighway, and when you came back 10 years later it was gone.

MySpace is not alone in encountering problems. Amazon cloud services, for example, also experienced a a substantial outage in 2011 and another in 2017.

Though temporary, and without actual loss of data, these outages left users without access to precious and important files for some time.

In a statement, Myspace said, 'We apologize for the inconvenience.'

A much bigger problem

Preserving content or intellectual property on the internet presents a conundrum. If it's accessible, then it isn't safe; if it's safe, then it isn't accessible.

Accessible content is subject to tampering, theft or other sorts of bad actions. Only content that is inaccessible can be locked and protected from hacking.

The internet currently accesses about 15 zettabytes of data, and is growing at a rate of 70 terabytes per second. It is an admittedly leaky vessel, and content is constantly going offline to wind up lost forever.

Byte sizes

The internet is approximately 15 zettabytes. By way of comparison, all the world's beaches combined have been estimated to hold roughly 10^{21} grains of sand.

zettabyte	10^{21}	1,000,000,000,000,000,000,000	sextillion
exabyte	10^{18}	1,000,000,000,000,000,000	quintillion
petabyte	10^{15}	1,000,000,000,000,000	quadrillion
terabyte	10^{12}	1,000,000,000,000	trillion
gigabyte	10^{9}	1,000,000,000	billion
megabyte	10^{6}	1,000,000	million
kilobyte	10^{3}	1,000	thousand
byte	10^{0}	1	one

Chart: The Conversation, CC-BY-ND Source: Paul Royster, University of Nebraska-Lincoln Get the data

Massive and desperate efforts are underway to preserve whatever is worth preserving, but even sorting out what is and what is not is itself a formidable undertaking. What will be of value in 10 years – or 50 years? And how to preserve it?

Acid-free paper can last 500 years; stone inscriptions even longer. But magnetic media like hard drives have a much shorter life, lasting only three to five years. They also need to be copied and verified on a very short life cycle to avoid data degradation at observed failure rates between 3% and 8% annually.

Then there is also a problem of software preservation: How can people today or in the future interpret those WordPerfect or WordStar files from the 1980s,

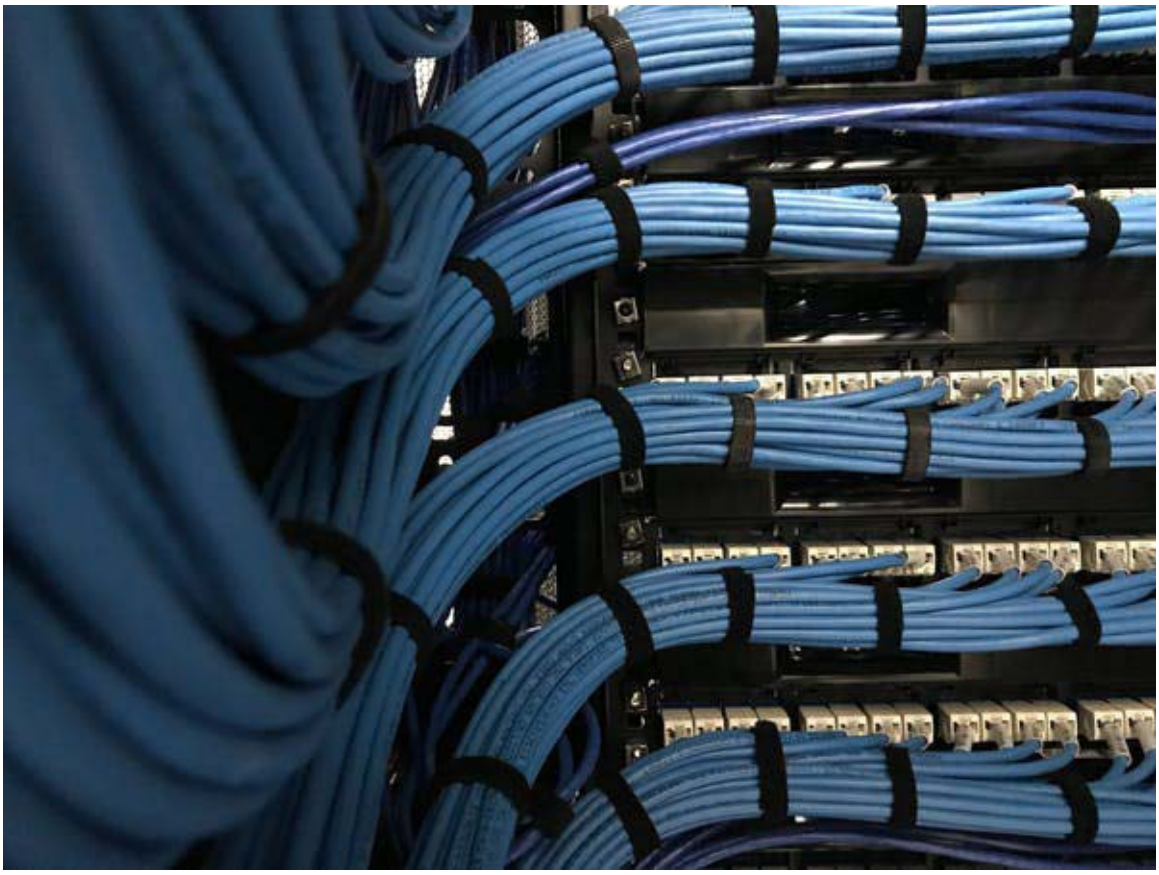
when the original software companies have stopped supporting them or gone out of business?

A nonprofit startup called The Internet Archive is preserving snapshots of the web on an ongoing basis, but mostly this is for top-level public HTML webpages such as The New York Times website and Facebook, not for underlying content files. As of last fall, its Wayback Machine held over 450 billion pages in 25 petabytes of data. This would represent .0003% of the total internet.

Universities, governments and scientific societies are struggling to preserve scientific data in a hodgepodge of archives, such as the U.K.'s Digital Preservation Coalition, MetaArchive, or the now-disbanded collaborative Digital Preservation Network. Preservation is hard and expensive in time, money and equipment. To be most useful, it not only has to be stored, but hosted in a form that is accessible and available for future reuse.

Actual storage costs less than $0.05 per gigabyte, but storage is only a small percentage of the costs of preservation. Acquisition, networking, maintenance and administration all require substantial and costly human labor.

Budgeting models suggest a 10-year preservation expense of around $2.50 per gigabyte, or $2,500 per terabyte, or $625,000 for the files MySpace failed to preserve.



Huge amounts of new content are uploaded to the internet every day.

A minute on the internet

Software company Domo calculated how much happens in a typical minute on the internet.

3,877,140	Google searches
2,083,333	Snapchats shared
473,400	Tweets sent
49,380	photos Instagrammed
79,740	Tumblr posts
1,944	Reddit comments

Chart: The Conversation, CC-BY-ND Source: Domo Get the data

Considering your own data

So yes, the internet is rotting, but archivists and digital librarians like myself knew it was rotten already, as did anyone who ever got a "404 File Not Found" error. Where there is economic incentive to keep and use data – such as user information, profiles or browsing history – it may exist for quite a long time. It has been said by many that data is the new oil, and corporations are anxious to drill and exploit this resource.

However, where content is less valuable to whomever owns the servers, there is less incentive to invest in preserving it. A survey of 10 million hits from 25 random sites in 2004 suggests that 404 errors occur at close to 3% of targeted URLs. The internet is growing much faster than it is rotting, but both things are happening at once. No giant internet company has your interests closer to its heart than its own.

One preservation network is known under the acronym LOCKSS – Lots of Copies Keeps Stuff Safe – and that's a good rule of thumb. Always have a backup, and always have multiple backups. Guard your privacy and guard your content, at least that content you may wish to have preserved, like photos, email, that screenplay or novel, or video and music files. Copyright rules do not prohibit storing content you may have purchased, as long as you don't put it out for public sharing.

Free storage is a great offer, but sometimes you only get what you pay for. The internet is neither secure nor permanent. It never promised to be, and users

should not assume that it will become so. Parts are rotting and corroding and collapsing as I type this. Just hope and plan to not be resting on that platform when it falls.

https://theconversation.com/your-internet-data-is-rotting-115891