Will the Web break?

Jonathan Zittrain

Harvard Law School, Harvard School of Engineering and Applied Sciences, and Harvard Kennedy School, Cambridge, MA, USA

What is the Web? What makes it work? And is it dying? This paper is drawn from a talk delivered by Prof. Zittrain to the Royal Society Discussion Meeting ‘Web science: a new frontier’ in September 2010. It covers key questions about the way the Web works, and how an understanding of its past can help those theorizing about the future. The original Web allowed users to display and send information from their individual computers, and organized the resources of the Internet with uniform resource locators. In the 20 years since then, the Web has evolved. These new challenges require a return to the spirit of the early Web, exploiting the power of the Web’s users and its distributed nature to overcome the commercial and geopolitical forces at play. The future of the Web rests in projects that preserve its spirit, and in the Web science that helps make them possible.

1. The past of the web

I have been asked to speak on the intriguing subject of: will the Web break? In order to approach that subject, we should consider an antecedent question: does the Web work? And a question even antecedent to that: what is the Web?

I look back to the ancient PC introduced in 1977. Data comes in and out through the various drives, on a CD-ROM or floppy disk, and you get output through a screen or speech synthesis or something. Then along comes the Internet, which has plenty of differences from the telephony system, but one basic similarity at first: you assume the stuff is already there, islands of information such as these PCs.

Thanks to the network, they are all hooked to one another synchronously, and you can trade data among them. The net is still basically invisible, you are just trading the data from one machine that has it to another that wants it. Now you have this overlay of the World Wide Web. What does the Web offer on top of this? One important thing it offers is the idea of not just having piles
of data on machines, but also a way of displaying it and having it organized so that even a small slice of that data can be uniquely identified. This is the basic concept of the universal resource identifier and under it the uniform resource locator (URL). Anything you want to say you have a link to it and if you say that link, what you are saying is what is on the other end of it.

That turns out to be an extremely powerful idea. It leads to the concept of simulating one seamless group of stuff that can link to each other, even though it is still underneath a bunch of independent machines all owned by independent parties. Of course, it is not ownership or restriction in the sense of, 'I would like to let this person see this but not that person see that'. It has rudimentary password architectures but, for the most part, it is expected to be a sort of collective, open system. If you have the link, you have the data. The closest we saw over the years of restriction as it came around is the robots.txt standard, which is not a standard at all and approved by no particular body.

robots.txt started out on a mailing list. In the crudest of ways, it can say: if you are a robot visiting this website, you should not be crawling the following directories. In other words, here is where the really good stuff is, do not look at it. The amazing thing is that there is no enforcement architecture. This just says ‘This is my preference. Please don’t look behind these curtains.’. Amazingly, as robots proliferated, chartered by Google or Microsoft or Yahoo or whoever, they respected robots.txt.

This is an extremely powerful idea because it is a companion to the sense of the Web as interconnected public stuff. ‘Public’ can have gradations. If there is stuff you want to put off limits in some way or make semi-private, then you can make it so a link will lead to it, but search engines would not index it, all made possible by an architecture for a naive request that actually works. Think about something like the founding of Creative Commons. Creative Commons said, great, we now have an architecture where you can share stuff. How do we actually make the legal side of this work? After all, under the Berne Convention, the act of putting something in a tangible form, including writing it in bits to somewhere, actually accrues copyright to it, even if, as the author of these bits, you did not intend to copyright it. When Creative Commons was started, the concept was to make a digital Library of Alexandria where people could deposit bits they really intended to share. The Web would be out there, and there would be the Creative Commons Library with the open bits. Luckily, people realized that the Web was that library, and the search engines were the indices to that library. Really, all Creative Commons needed to be was a very simple tagging architecture that would let people say ‘by the way, here is my feeling about how I would like these bits to be treated’.

That feeling is backed up nominally by law. Famously, Creative Commons licences are said to have three layers: the human-readable layer, Creative Commons Non-Commercial, etc., and you kind of know what that means; the lawyer-readable layer, which no one understands but it is supposed to be the legal instantiation of that request; then the key, machine-readable layer where you actually see tags that are consistent enough to instruct Google without Google realizing it. For instance, I could request all the images that match ‘Westminister Abbey’, and that are also tagged as Creative Commons licensed.

The amazing insight then was that the library was already there and waiting. It was just a question of a slight information architecture shift, just the way that robots.txt came along to indicate preferences about how data should be treated. If that is my sense of some of the most important definitional aspects of the Web, we then ask the question, has the Web worked? And, of course, to ask it is to answer it: Yes.

I look back to what you might call yesterday’s network of tomorrow: the CompuServe main menu, circa 1990. In a way, it has some elements that you can think of as hyper-texty. When you ask for news, you go to a news area, and so on. But of course all of it is curated, under one roof, by CompuServe. There was an amazing period of time when we thought the future of networks was going to be determined by the outcome of a big battle among CompuServe, AOL and Prodigy, whichever one was left standing would be the main network that we would all use. To our fortune, a much less curated platform, one that anybody could
The Web is dead. Long live the Internet
by Chris Anderson and Michael Wolff  August 17, 2010 | 9.00 am | Wired September 2010

Figure 1. Wired's 'The Web is dead, long live the Internet' graphic illustrates various technologies as a percentage of traffic volume, including the shrinkage percentage that is the Web. The Web is at the bottom. (Online version in colour.)

add to in the shape of the Web running on top of the Internet, put the CompuServe model to shame.

2. The Web of today

So then why do we see today things like Wired and its amazing cover story ‘The Web is dead’, with this graph (figure 1) showing that the Web is decreasing in size since the year 2000 as a proportion of total US Internet traffic. [1] Luckily, some others have come up with perhaps more helpful versions of this chart (figure 2; [2]). This is not just showing everything under a divisor of 100, but letting all traffic grow. You can see at the bottom, the Web, as a simple set of Web pages, is actually growing.

I actually find this depiction particularly helpful. What this is showing is the total proportion from 1990 to 2000 and then 2010. You can see that the amount of material going up on the Web has just been extraordinary. The fact that people are trading a lot of bandwidth intensive music that they should not be is, essentially, an irrelevancy next to the growth of the Web. It is also amazing thinking about the earliest beginnings of what we might now today call ‘Web science’. I think it was Robert Wilensky who said that it has been said that a million monkeys and a million keyboards could eventually produce the complete works of Shakespeare, now thanks to the Web, we know this is not true. It really just shows how many people you have got together producing all sorts of things and enabled to just put them up.

3. Will the Web break?

If I am offering, then, a so far optimistic view as opposed to quasi-superficial declarations of ‘the Web is dead’, let me now take up seriously the question of will the Web break? Here are a couple of causes for concern if we think about the Web in its core form and what it means to us. These
Figure 2. Boing Boing’s graphic used the same data, but showed overall traffic growth, not percentage by technology. The Web is at the bottom. (Online version in colour.)

are not just whether people will be using hyper-linked protocols, or if HTTP will be what people type or mean when they ask for a site, but simply that we are seeing the beginning of the end, or maybe even the middle of the end of the architecture and spirit of the PC and Web that were obtained from 1977 and 1992, respectively.

More and more of what is on our local machine no longer matters and for good reason. It is a real feature to be able to get to all your mail via Gmail, out in the cloud. If you lose your machine, you lose the value of the hardware and maybe some data or passwords cached on it, but you are no longer carrying around your trove with you. It echoes the transition when we went from the answering machine to voice mail, which happened without great tumult. It would be rightly unpersuasive of me to say that we really need to bring back the answering machine.

But there are some real issues that come with the transformation of the PC away from acting as a repository, and not just of data, but of code. The one thing about the PC was you could run any code you wanted. Looking at the platforms of today, I see the main menu of something like the iPhone and it seems to me an echo of main menus I have seen before. In fact, the first version of the iPhone did not allow any third-party applications—it really was just the CompuServe main menu. The second version, of course, has this intriguing, bizarre notion of a curated platform where third parties give you code to run, but Apple gets to be the funnel.

I do not just mean to pick on Apple here. I believe this model is a very powerful, tantalizing model. I think it could spread, not just throughout the Apple product line, but we now hear word, for instance, that Intel is putting out a new chip whose purpose, at the hardware layer, is to allow for code signing, so that the next PC you get will be one for which the manufacturer or the operating system maker can have a curated platform of just the sort that we are starting to see in the mobile zone. Microsoft has recently announced a similar system for curating applications for WINDOWS 8, and MACOS now boasts a curated App Store. This is one of the reasons I am interested in the battle between Android and iPhone, and others to see just how curated we want things to be.

To the extent that we have a more curated zone, it means that our experience is no longer as free to run the code we want or even to go to the destinations we want. What we see, at least on the iPhone and other platforms like it, is that the browser is secondary. It is nice to have a browser but, really, if you want to get somewhere, there is an app for that—even if the app is just a wrapper for a website. More and more content is bundled into the app and then Apple is not just curating code, but content as well, asking for you ‘is this offensive; is this worth your time?’. This rise of paternalistic, de facto, corporate content curation is one way in which it is worth asking if the Web is breaking.
Another way is to look at the Web’s persistence, link by link and destination by destination. Bit.ly and other URL shorteners have introduced a level of indirection that can be very convenient if, say, an arbitrary party has told you that whatever you have to say has to be 140 characters or less, but now everything is getting piped through Bit.ly. Should Bit.ly go belly up, all of those links break. We thought the original problem was a link points to somewhere, that somewhere goes offline, we do not have it anymore. But now, even if the somewhere is still there, we have introduced a new point of failure into the system. That is a very direct and technical way we know the Web might break.

Another example. I am in the UK attempting to watch MTV. And I am told in a wonderful use of the passive voice, copyrights restrict us from playing this video outside the US, though we really wish we could. We are showing this video to an audience of 350 million people, but you cannot see it because we have determined you are in the wrong location. I do not know if this counts as breaking because it is all using Web protocols, this is what it was designed to do, but it goes against what you might say is some of the spirit of the Web, which is increasingly geographically bound and uses other forms of content zoning.

Of course, it is not just the USA doing this. In the USA, my attempt to watch the BBC iPlayer results in a message that says ‘Currently, iPlayer programs are available to play in the UK only’. Why is this? Because the UK taxpayers paid for it and unless you are prepared to pay UK tax, do not put your hands on the BBC. Again, this has a sort of odd feeling of ‘things going away’. Here is another example. The New York Times published an article in which they talk about a criminal trial about to commence in the UK. And they got nervous that they could get in trouble in the UK for sharing information prejudicial to the defence of the defendant. It is not that they do not want to publish it, they would be happy to prejudice his defence, but they do not want to get in trouble in the UK. So we got this wonderful message ‘On advice of legal counsel, this article is unavailable to readers of nytimes.com in Britain’. This article is unavailable to readers of nytimes.com in the UK; this arises from the requirement in UK law that prohibits publication of prejudicial information. So okay, no prejudicial information to anyone in the UK, but everybody else can read it and on their honour should not email it to their friends who are potential jurors in that territory [3].

It is not just a question of can you get there or not. In the models I have shown so far, you at least know what you do not know. But here is an example of a different kind, comparing google.com and google.de on a search on the term ‘storm front’. If you search for ‘storm front’ on google.com, the first hit is for a neo-Nazi organization. If you look for it on google.de, that organization’s listing appears nowhere because Google’s interpretation of German law is that that site is simply not allowed to be displayed in Germany. You could even enter what amounts to the same URL, and, depending on where you are, that URL will point to different information.

No longer, then, can you say that your URL is ‘U’: it is no longer uniform. This is a phenomenon we are going to have to contend with. Now, if I were thinking about the Wired chart of the Web I would, in wondering about the future of the Web and its trends, think about that concept of uniformity and ask, how many URLs are there going to be? How many repositories of information are there that are only available to me or to my friends? Looking at something like Facebook, you realize there are more and more data organized for the sake of privacy that are no longer publicly accessible. Who can get them is now defined not crudely by geography but by very finely honed tools dictated by the user.

At that point I do not know if I should say that this is the Web. Is it a Web page? Of course. Is it using Web protocols? Yes. Is it on the World Wide Web? I guess so, it is on the World Wide Web for me. But as soon as I say something is on the World Wide Web ‘for me’, that is already starting to limit what we mean when we talk about a World Wide Web. There is also a sense in which, if you have every machine hooked up to the Web and each machine could potentially become a server, you see the Web is very spread out. Anybody can just put up their own server and away they go. Now of course Zipf’s law, the power law, has shown us otherwise.
Of the top 1000 sites as ranked by Google, three per cent of the top 1000 account for 80 per cent of the page views. At that point, you can say, wow, it is nice to have a safety valve. There could be a website in a forest that no one visits and I guess it is still serving. But that is sort of a philosophical question, rather than a reality that says more and more of the places we are going for data on the Web may look like those top three per cent. There are many different flavours of Google. But it is still sort of Henry Ford’s abjuration that your car can be whatever colour you want so long as it is black. We are reintroducing curators of content, and that is something I think we really need to keep an eye on.

I look at something like the Google Books project, which I think is a terrific project. I am very excited about the prospect for scanning all of humanity’s knowledge, as represented in books, to a commonly accessible zone. But when you go through to some of these scans you notice, say, this note in 1984 by George Orwell ‘Pages 20 to 21 are not shown in this preview, learn more.’ And when you ‘learn more’, you are not learning more about pages 20 to 21. It is just holes in the Swiss cheese because there is a settlement currently pending in the US courts by which maybe people in the USA will be able to pay money to see pages 20 and 21.

That starts to raise questions. To the extent that these scans become available widely enough that it is no longer worth it for libraries to stock the physical counterpart, we will start to rely more and more on Google Books. But it is not exactly the universal Web again because, depending on where you are, you will have different rights with respect to the content. Depending on how much money you have, you may or may not be able to see it. Most important, it has what I call the Fort Knox problem, which is that we then start to see data gravitationally clustering into single repositories.

It is funny. When we think of entropic systems, systems that tend towards entropy, the classical example is you cannot unstir things. You would think data is supposed to spread out, but somehow we see these clusterings going on in the Web where all of our goods end up in one place. When that happens, it makes it easy for anybody who objects to the existence of those goods to do something about it.

Under US law, there is a very little-known and little-used provision by which information that has been deemed copyright infringing can itself be impounded and disposed of. It is like a team of marshals raiding a warehouse before the infringing books can leave for the bookstores and the officers seize them all. That is now possible in a way that can become routine, and under the law, would be trivial. One example of this, that was just effected through private parties getting scared, with no case actually filed, was on the Kindle. 1984 was made available through Amazon by a third party, you could pay 99 cents and get Orwell’s 1984. It turns out that 1984 is in the public domain in Canada, where the seller was, but still under copyright in the US. Oops. Amazon panics and reaches in to every single Kindle that has downloaded 1984 and deletes it.

It was like ‘You don’t have 1984. You never had 1984. There is no such book as 1984.’. This is a very high-profile example. You are not going to be able to delete 1984 without catching flack, and Amazon apologized up and down for it. But that does not matter. Even if Amazon regrets it, it shows the power of an entity to go to a single set of collective curators, Amazon, Google, maybe two others and give an order to change or delete, maybe not an entire work, but just a piece of a work that is said to be defamatory or infringing. Suddenly, our entire historical record starts to become altered.

That is a very, very different configuration than the ones we have now. In most contraband material cases, someone gets penalized. You might pay damages, but only in the rarest of instances can the material itself be deleted or changed. So then what are some of the approaches we should be taking to preserve the essence of the Web? For that I start to think, sort of as a hat tip to Tim Berners-Lee’s former and current profession as a physicist, what some of the ideal physics of the Web are, in my view?

We say that the universe appears to be isotropic, which is to say wherever you are, you see the same stuff. There is not one place in the firmament that is special. It is not like the laws of physics
are different here to there. That should be an idealization for the Web, a goal to aim for. Wherever you are, that same link takes you to the same place.

I am not insisting that one size fits all; there should be different links for different flavours of things. But I ought to be able to send the URL that I see to my friend and say, look at this, see what I just saw and they will see the world through my point of view. The universe is also said to be time symmetric and persistent, which is how inference and induction work. What we did yesterday might be suggestive of what happens today and the experiment we run, under the same conditions, will always have the same results. I think that means it would be great if we had a Web for which when there is a link to a place, it stays there.

Generative is my own word to describe an environment in which unaffiliated third parties can produce stuff, and interesting and surprising things happen. As I understand quantum foam, that sometimes happens in the universe, too. If you start with a completely empty void, if there is such a thing, eventually something just appears. There is an amazing way then in which the tabula rasa can reset itself. I think that is a phenomenon we have seen on the Web precisely because the Web is so open to contribution from nearly anywhere. Going back to the observation that a million monkeys on a million typewriters could produce the complete works of Shakespeare, you could now say they are not producing Shakespeare because they are too busy producing all sorts of interesting things that do not even exist yet.

To push the physics comparison further, if there are many universes, it would be nice to have wormholes to travel among them rather than to have them be out there and inaccessible. Think of the many Wikipedias, one for each major language. They exist in their own bubbles. As automatic translation tools come online, it will be fascinating to see what the German language, the English language and the Chinese language versions of Wikipedia all have to say about the same topic. We will be able to set up different zones so we will not all share the same thing, but we will have bridges between them. This seems to be a really great value to aim for as we move forward with the Web.

Finally, there is this idea of the entropic, in the sense of, should the Web be spreading and moving outward and becoming distributed rather than agglomerating into just a few nodes? I think that entropic tendency might be a way to save the Web. So how do we do it? Too often, I think, we view the Web as an individual enterprise. We back-up our own data, we deal with our own merchants and other interactions and that is it. We sit with our shotgun trying to defend our own stuff. I would love to see a broader view by which, when we see the Fort Knox problem arriving and stuff coming together, we say no, we do not want that. There are a few examples of that happening on the Web itself.

In 1996, a guy named Brewster Kahle decided to start the Internet Archive. He crawls the entire Web and keeps a copy. He does it every so often, like painting the Golden Gate Bridge. As soon as he is done, he starts over at the other end. This is probably the most massive single act of copyright infringement ever. Still, Brewster is a nice and sensible guy. If you ask him to take something down, he does, in the sense of he restricts public access to it. He even allows you to do this automatically and retroactively. You can change your robots.txt file to say ‘I don’t want my Web site to be archived’. And it is so profoundly useful that it is hard to imagine the judge who would dare to get rid of archive.org despite the fact that it is brazenly illegal.

That is an amazing triumph of standards, practices and protocols over the literal doctrine of the law. It is also an excellent back-up so that if the original source goes down, I have got at least one other place to go to look for my content. I put the Google cache and some of the other search engine caches into this category. More formally, in academia we have seen projects such as ‘Lots of Copies Keeps Stuff Safe’ (LOCKSS) [4] and Slyck’s Database of file sharing programs [5], by which libraries with digital documents they want to preserve for the ages share them with other libraries and compare among themselves every so often to make sure that the integrity of the documents remains.

The biggest flaw in the Google Books project could be greatly addressed by having participating libraries acting not just as windows to Google’s goods, or warehouses for terminal
make your stuff available on the web (whatever format)
make it available as structured data (e.g. excel instead of image scan of a table)
non-proprietary format (e.g. csv instead of excel)
use URLs to identify things, so that people can point at your stuff
link your data to other people’s data to provide context

Figure 3. Tim Berners-Lee’s five stars of open linked data, courtesy of Ed Summers. See also Summers [6]. (Online version in colour.)

screens where you can look at Google’s books. Instead, these libraries should have actually have copies of the scans and compare their corpus with other participating libraries, and with Google, and the minute there is some change or reduction or alteration, it gets detected by a distributed group. To me that is the way of the Web and why this project is so valuable.

Now I think of some of Tim Berners-Lee’s current work, in which he offers a five-star Michelin-style system to incentivize posting data online (figure 3). If you just dump it on the Web, you get one star because all you have done is just put the data out there. If anything, the one-star approach actually tends to encourage the likes of Google because they have the PhDs and the computing power to crawl all of this junk and make sense of it.

What I see in Tim’s graduated star system is a way of saying that you should just put a little more effort into structuring your data. This will make it so the work that goes into tagging and structuring is done by the person most able to do it, and then it is much easier to write a search engine without the PhDs. You will notice here the fifth star is ‘link your data to other people’s data’. That is a wonderful way of trying to compromise on the Fort Knox problem. It is not saying link your data to just one canonical source that everyone will link to. Then you get the Bit.ly problem. Instead, we will let different people link to different notions of the data, and maybe there will be three or four idealizations of it, one of them perhaps in Wikipedia. If it gets cited to enough, we will have Brewster come along or ask Wikipedia to escrow that stuff so that those common links do not, a la Bit.ly, end up going dead should one of these emerging points of failure end up failing.

On the persistence front, I have been toying around with the idea of mutual aid as a way of trying to solve some of these problems, whether it is a denial of service attack or just stuff becoming unavailable. My thought was we could have a North Atlantic Treaty Organization (NATO) for cyberspace, a kind of a mutual aid treaty. If you are in trouble, I will send troops and if I am in trouble, you will send troops. Of course, the time I am in trouble is when you are least interested in the treaty and there is not necessarily a way to get you hold up your end.

But when I think of mutual aid as ‘clustering the fishing boats together to avoid pirates’, you can actually see ways of devising something that would help. So here is one instantiation of mutual aid that Tim has talked about before, and that I think has not received as much attention. There you are, clacking on your keyboard. You go to a server, it renders a Web page for you. That Web page has links to external sites. You go to visit that site and the site is not available. Either it is down or it has been blocked or there is a denial of service attack, who knows what?

Currently, there is no alternative. You are done. But, what if we had a new system implemented through the handful of Web servers that there are, Apache and Microsoft, by which Web masters could opt into a system whereby, when they render a page for somebody, they cache what is at the link. They grab a copy of whatever they are linking to. Then, when the person tries to get to the site and cannot, they can go back to the original site and say ‘Hey, I can’t get that link you just directed me to, would you mind telling me what was there?’.

All of a sudden, we have a scheme of mutual aid that naturally comes about, for which if I participate as a website I will know that others linking to me will also mirror my stuff. So there is an incentive for entities to participate in this. For the likes of Google, it can work out well because they are already indexing everything they see.
It also works if there is filtering going on. If it turns out the block is in between the person and the site, then you still end up getting to what you are missing. It is a subversive way of circumventing government filtering without having to advertise as some special human rights tool that will scare everybody.

When I think of an agenda for Web science, I think of the fact that once you get to a certain amount of data and a certain zoom in on your electron microscope, it no longer even makes sense to talk about what you see. There is too much, there is nothing that the naked eye could actually perceive. Everything is going to be interpretation of data. What Web science offers is the promise of actually being able to visualize data of the Web itself, perhaps even provided by the entities on the Web through mutual aid modalities, so that we can understand where the Web is going.

Then, we can see exactly how entropic it is, exactly how uniform it is, and apply on top of that the interdisciplinary zones of psychology, of business, of law, of policy, to realistically ask ourselves do we like where it’s going? If we do not, then what can we do to put it into a better place without having to become regulators?

My sense of the answer to that question is that the more the individual actually wants to help out, the better. That has been the key to making the Web succeed from its earliest days, somebody wanting to help. So I think of this xkcd cartoon (figure 4).

You actually have people eager to get out there and make things better. I hope that is the charge we take upon ourselves as we look to guide the World Wide Web into what will be its greatest days yet.

References