



# Future Imperfect



**Vinton G. Cerf**  
Google

**A**s the second decade of the 21st century dawns, predictions of global Internet digital transmissions reach as high as 667 exabytes (10<sup>18</sup> bytes; [http://en.wikipedia.org/wiki/SI\\_prefix#List\\_of\\_SI\\_prefixes](http://en.wikipedia.org/wiki/SI_prefix#List_of_SI_prefixes)) per year by 2013 (see <http://telephonyonline.com/global/news/cisco-ip-traffic-0609/>). Based on this prediction, traffic levels might easily exceed many zettabytes (10<sup>21</sup> bytes, or 1,000 exabytes) by the end of the decade. Setting aside the challenge of somehow transporting all that traffic and wondering about the sources and sinks of it all, we might also focus on the nature of the information being transferred, how it's encoded, whether it's stored for future use, and whether it will always be possible to interpret as intended.

## Storage Media

Without exaggerating, it seems fair to say that storage technology costs have dropped dramatically over time. A 10-Mbyte disk drive, the size of a shoe box, cost US\$1,000 in 1979. In 2010, a 1.5-Tbyte disk drive costs about \$120 retail. That translates into about 10<sup>4</sup> bytes/\$ in 1979 and more than 10<sup>10</sup> bytes/\$ in 2010. If storage technology continues to increase in density and decrease in cost per Mbyte, we might anticipate consumer storage costs drop-

ping by at least a factor of 100 in the next 10 years, suggesting petabyte (10<sup>15</sup> bytes) disk drives costing between \$100 and \$1,000. Of course, the rate at which data can be transferred to and from such drives will be a major factor in their utility. Solid-state storage is faster but also more expensive, at least at present. A 1-Gbyte solid-state drive was available for \$460 in late 2009. At that price point, a 1.5-Tbyte drive would cost about \$4,600. These prices are focused on low-end consumer products. Larger-scale systems holding petabyte- to exabyte-range content are commensurately more expensive in absolute terms but possibly cheaper per Mbyte. As larger-scale systems are contemplated, operational costs, including housing, electricity, operators, and the like, contribute increasing percentages to the annual cost of maintaining large-scale storage systems.

The point of these observations is simply that it will be both possible and likely that the amount of digital content stored by 2010 will be extremely large, integrating over government, enterprise, and consumer storage systems. The question this article addresses is whether we'll be able to persistently and reliably retrieve and interpret the vast quantities of digital material stored away in various places.

Storage media have finite lifetimes. How many 7-track tapes can still be read, even if you can find a 7-track tape drive to read them? What about punched paper tape? CD-ROM, DVD, and other polycarbonate media have uncertain lifetimes, and even when we can rely on them to be readable for many years, the equipment that can read these media might not have a comparable lifetime. Digital storage media such as thumb drives or memory sticks have migrated from Personal Computer Memory Card International Association (PCM-CIA) formats to USB and USB 2.0 connectors, and older devices might not interconnect to newer computers, desktops, and laptops. Where can you find a computer today that can read 8" Wang word processing disks, or 5 1/4" or 3 1/2" floppies? Most likely in a museum or perhaps in a specialty digital archive.

### Digital Formats

The digital objects we store are remarkably diverse and range from simple text to complex spreadsheets, encoded digital images and video, and a wide range of text formats suitable for editing, printing, or display among many other application-specific formats. Anyone who has used local or remote computing services, and who has stored information away for a period of years, has encountered problems with properly interpreting the stored information. Trivial examples are occurring as new formats of digital images are invented and older formats are abandoned. Unless you have access to comprehensive conversion tools or the applications you're using continue to be supported by new operating system versions, it's entirely possible to lose the ability to interpret older file formats. Not all applications maintain backward compatibility with their own versions, to say nothing of ability to convert into and from a wide range of formats other than their own. Conversion often isn't capable of 100 percent fidelity, as anyone who has moved from one email application to another has discovered, for example. The same can be said for various word processing formats, spreadsheets, and other common applications.

How can we increase the likelihood that data generated in 2010 or earlier will still be accessible in useful form in 2020 and later? To demonstrate that this isn't a trivial exercise, consider that the providers of applications (whether open source or proprietary) are free to

evolve, adapt, and abandon support for earlier versions. The same can be said for operating system providers. Applications are often bound to specific operating system versions and must be "upgraded" to deal with changes in the operating environment. In extreme cases, we might have to convert file formats as a consequence of application or operating system changes.

If we don't find suitable solutions to this problem, we face a future in which our digital information, even if preserved at the bit and byte level, will "rot" and become uninterpretable.

### Solution Spaces

Among the more vexing problems is the evolution of application and operating system software or migration from one operating system to another. In some cases, older versions of appli-

---

**If we don't find suitable solutions to this problem, we face a future in which our digital information will "rot" and become uninterpretable.**

---

cations don't work with new operating system releases or aren't available on the operating system platform of choice. Application providers might choose not to support further evolution of the software, including upgrades to operate on newer versions of the underlying operating system. Or, the application provider might choose to cease supporting certain application features and formats.

If users of digital objects can maintain the older applications or operating environments, they might be able to continue to use them, but sometimes this isn't a choice that a user can make. I maintained two operational Apple IIe systems with their 5 1/4" floppy drives for more than 10 years but ultimately acquired a Macintosh that had a special Apple IIe emulator and I/O systems that could support the older disk drives. Eventually, I copied everything onto newer disk drives and relied on conversion software to map the older file formats. This worked for some but not all of the digital objects I'd created in the preceding decade. Word processing documents were transfer-

able, but the formatting conventions weren't directly transformable between the older and newer word processing applications. Although special-purpose converters might have been available or could have been written – and in some cases were written – this isn't something we can always rely on.

If the rights holder to the application or operating system in question were to permit third parties to offer remote access in a cloud-based computing environment, it might be possible to run applications or operating systems that developers no longer supported. This kind of licensing would plainly require creative licensing and access controls, especially for proprietary software. If a software supplier goes out of business, we might wonder about provisions for access to source code to allow for support in the future, if anyone is willing to provide it, or acquisition by those depending on the software for interpretation of files of data created with it. Open source software might be somewhat easier to manage from the intellectual property perspective.

### Digital Vellum

Among the most reliable and survivable formats for text and imagery preservation is vellum (calf, goat, or sheep skin). Manuscripts prepared more than a thousand years ago on this writing material can be read today and are often as beautiful and colorful as they were when first written. We have only to look at some of the illuminated manuscripts or codices dating from the 10th century to appreciate this. What steps might we take to create a kind of digital vellum that could last as long as this or longer?

Adobe Systems has made one interesting attempt with its PDF archive format (PDF/A-1; [www.digitalpreservation.gov/formats/fdd/fdd000125.shtml](http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml)) that the ISO has standardized as ISO 19005-1. Widespread use of this format and continued support for it throughout Adobe's releases of new PDF versions have created at least one instance of an intended long-term digital archival format. In this case, a company has made a commitment to the notion of long-term archiving. It remains an open question, of course, as to the longevity of the company itself and access to its software. All the issues raised in the preceding section are relevant to this example.

Various other attempts at open document

formats exist, such as OpenDocument format 1.2 (and further versions) developed by OASIS (see [www.oasis-open.org](http://www.oasis-open.org)). The Joint Photographic Experts Group has developed standards for still imagery (JPEG; [www.jpeg.org](http://www.jpeg.org)), and the Motion Pictures Experts Group has developed them for motion pictures and video (MPEG; [www.mpeg.org](http://www.mpeg.org)). Indeed, standards in general play a major role in helping reduce the number of distinct formats that might require support, but even these standards evolve with time, and transformations from older to newer ones might not always be feasible or easily implemented. The World Wide Web application on the Internet uses HTML to describe Web page layouts. The W3C is just reaching closure on its HTML5 specification (<http://dev.w3.org/html5/spec/Overview.html>). Browsers have had to adapt to interpreting older and newer formats. XML ([www.w3.org/XML/](http://www.w3.org/XML/)) is a data description language. High-level language text (such as Java or JavaScript; see [www.java.com/en/](http://www.java.com/en/) and [www.javascript.com](http://www.javascript.com)) embedded in Web pages adds to the mix of conventions that need to be supported. Anyone exploring this space will find hundreds if not thousands of formats in use.

### Finding Objects on the Internet

Related to the format of digital objects is also the ability to identify and find them. It's common on the Internet today to reference Web pages using Uniform Resource Identifiers (URIs), which come in two flavors: Uniform Resource Locators (URLs) and Uniform Resource Names (URNs). The URL is the most common, and many examples of these appear in this article. Embedded in most URLs is a domain name (such as [www.google.com](http://www.google.com)). Domain names aren't necessarily stable because they exist only as long as the domain name holder (also called the *registrant*) continues to pay the annual fee to keep the name registered and resolvable (that is, translatable from the name to an Internet address). If the registrant loses the registration or the domain name registry fails, the associated URLs might no longer resolve, losing access to the associated Web page. URNs are generally not dependent on specific domain names but still need to be translated into Internet addresses before we can access the objects.

An interesting foray into this problem area is called the Digital Object Identifier (DOI; [www.doi.org](http://www.doi.org)), which is based on earlier work

at the Corporation for National Research Initiatives ([www.cnri.reston.va.us](http://www.cnri.reston.va.us)) on digital libraries and the Handle System ([www.cnri.reston.va.us/doa.html](http://www.cnri.reston.va.us/doa.html)) in particular. Objects are given unique digital identifiers that we can look up in a directory intended to be accessible far into the future. The directory entries point to object repositories where the digital objects are stored and can be retrieved via the Internet. The system can use but doesn't depend on the Internet's Domain Name System and includes metadata describing the object, its ownership, formats, access modes, and a wide range of other salient facts.

vital that we solve the problems of long-term storage, retrieval, and interpretation of our digital treasures. Absent such attention, we'll preside over an increasingly large store of rotting bits whose meaning has leached away with time. We can hope that the motivation to circumvent such a future will spur creative solutions and the means to implement them. □

**Vinton G. Cerf** is vice president and chief Internet evangelist at Google. His research interests include computer networking, space communications, inter-cloud communications, and security. Cerf has a PhD in computer science from the University of California, Los Angeles. Contact him at [vint@google.com](mailto:vint@google.com).

**A**s we look toward a future filled with an increasingly large store of digital objects, it's

**cn** Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

## How far have we come?

See *IC's* Millennium Predictions (Jan/Feb 2000 special issue)

- "Guest Editors' Introduction: An Internet Millennium Mosaic":  
<http://doi.ieeecomputersociety.org/10.1109/MIC.2000.815848>
- "Millennial Forecasts":  
<http://doi.ieeecomputersociety.org/10.1109/4236.815849>

## Where will we go?

See more from our *IC's* Internet Predictions issue (Jan/Feb 2010)

- "Guest Editors' Introduction: Internet Predictions":  
<http://doi.ieeecomputersociety.org/10.1109/MIC.2010.11>
- "Internet Predictions":  
<http://doi.ieeecomputersociety.org/10.1109/MIC.2010.12>

**IEEE**  
**Internet Computing**  
[www.computer.org/internet/](http://www.computer.org/internet/)

This article was featured in

# computing **now**

ACCESS | DISCOVER | ENGAGE

For access to more content from the IEEE Computer Society,  
see [computingnow.computer.org](http://computingnow.computer.org).



IEEE  computer society

Top articles, podcasts, and more.



[computingnow.computer.org](http://computingnow.computer.org)