Avoiding "Bit Rot": Long-Term Preservation of Digital Information

By VINTON G. CERF, *Fellow IEEE*

There is something ultimately satisfying about keeping information in digital form. It does not take up much space. It can be replicated for resilient preservation. It can be searched mechanically. It can be used to combine with other material using digital power tools. But this blissful outlook may not comport with the reality of digital information preservation and interpretation.

When we use sophisticated tools to create complex digital objects, we pay a price: our dependence on the software to be available and operational whenever we need access to the information. Of course, we also rely on the availability of the stored information as well. There has been a great deal of research and development devoted to high-density and reasonably sturdy digital storage technology, but we have already seen how quickly these media become superannuated. Who uses 8″ floppy disks (such as the ones used with the Wang word processor)? What about the 5.25″ floppy, or the more rigid 3.5″ disk? What about those old 14″ (or larger) optical disks? VHS has largely disappeared along with the 8-track audio cassette. Even DVD is being superseded by Blu-Ray. It is not merely the medium that becomes unusable, it is also the reading device.

To make matters more complex, the formats of the digital objects of our affection are very much software dependent. If we are lucky, the application software uses broadly available, license-free standards that promote information exchange and application interoperability. Even where this is the case, the applications themselves may not survive the evolution of operating systems or their demise. Where is CP/M today? DOS? TENEX? The list is long.

I think it is arguable that media may be invented that have intrinsic durability over the course of decades if not centuries. We have examples in the form of vellum that is readable today (if you speak and read the language) though written two or three thousand years ago. The ability to process such documents optically presents some potential for longevity. One could even imagine using a kind of artificial vellum to render material in digital form that could be scanned in the future by machines yet to be invented. Two-dimensional barcodes or other high-density digital representations might survive for a long time. Proper interpretation of such formatted information will depend first on being able to properly demodulate the optical patterns and, second, on the ability to correctly interpret the digital content.

Digital Object Identifier: 10.1109/JPROC.2011.2124190

One has to keep in mind that the digital objects we create may not even be presentable in purely optical form. They may include sound, imagery, software, various databases, and other digital information forms, all of which must be properly interpreted (by software) to correctly render the digital content in the form intended. If we cannot maintain the operability of the software and its ability to ingest the digital forms preserved, we may find the information impossible to render or at least difficult or impossible to interpret.

One may pose a similar question about the rapid movement towards "cloud computing," in which information is stored "in the cloud"—often in replicated form for reliability. What is important is not only that the information be accessible within the cloud but that it can also be extracted from the cloud and moved to another cloud or another resting place of the owner's choosing. One is immediately struck by the need for intercloud conventions to allow high-speed transfer of data from one cloud to another. Downloading from cloud to desktop or laptop ultimately may be fruitless if the data of any significant size. In the cloud world, an Exabyte is likely to be a fairly common unit of cumulative data gathering.

Absent an ability to correctly interpret digital information, we are left with files full of "rotting bits" that are of no value. Hence the term "bit rot" for this situation.

The question remains: What's to be done?

To begin with, standards in this area are vital. Without standards, every format may be application-unique and potentially protected by various intellectual property fences that prohibit reverse engineering to achieve compatibility. Open standards have the beneficial property that they can be made widely available and multiple implementations may be available, giving choice to users. Development and adoption of standards for information representation will help but cannot completely eliminate the problem, however.

New applications are developed constantly and, because they are new and because their developers may have had to develop nonstandard approaches to achieve new effects, there will always be applications of interest whose data is not in a standard form. Widely popular applications may use formats that eventually become *de facto* or even de jure standards, but they may not achieve that status or, if they do, there may still be barriers to unlicensed use of these formats. What happens if the owners of the rights to the software or the formats go out of business? What if they abandon the applications and the associated formats? What if users have widely invested in producing content using these formats? What if they have purchased content represented in proprietary formats but the software to interpret the objects is no longer available or cannot operate on newer (versions of) operating systems? No amount of standardization will necessarily assure continued access to superannuated formats, especially those that are treated as proprietary.

It seems to me that serious consideration should be given to legal frameworks that would allow proprietary formats to be made broadly accessible if the software associated with them has been abandoned by its rightsholders. "We will no longer support format X" is a conclusion that can reasonably be reached for sound business reasons by corporations and other entities responsible for the creation of software using format X (what ever that may be). Leaving users stranded with investments in these formats seems counterproductive. One would hope for voluntary but might contemplate compulsory situations in which abandoned formats are made accessible in some fashion. Reverse engineering, emulators of obsolescent hardware or software and other mechanisms might be used to extend the utility of "retired" data whose format isn't any longer supported. Escrow of source code is one means by which companies have protected themselves from failures of companies upon whose software products they may depend. The escrow closes (i.e., the software becomes accessible) if the party that "owns" it fails. One could imagine situations in which software becomes accessible through cloud implementations if it has been declared obsolete or the entity with rights has abandoned support or gone out of business (see Fig. 1).

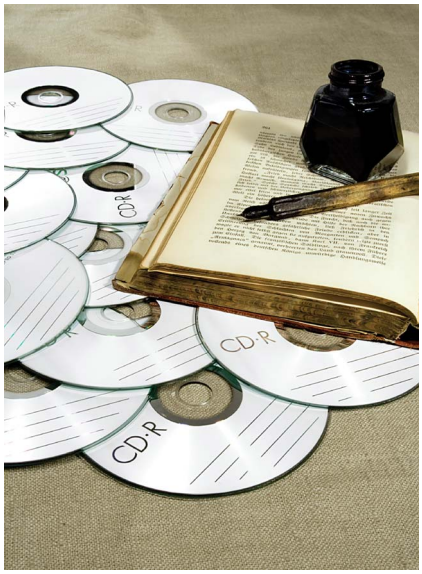

Fig. 1. ***It is not just the storage medium that data is saved on that will be important, but also the software that will help to determine availability and usability of stored information in the future.***

While contingency plans might be struck between parties on a contractual basis (such as with software escrow arrangements), it might be fruitful to explore standards terms and conditions under which proprietary formats and the software needed to interpret them might become broadly accessible. At the least, it seems prudent to encourage research in this area, so as to increase the likelihood that information produced today will be accessible and interpretable in the future. ■