



[<< Return to article](#)

The Myth of Doomed Data

The handwringing about obsolete formats is misguided. The digital files we create today will be around for a very, very long time.

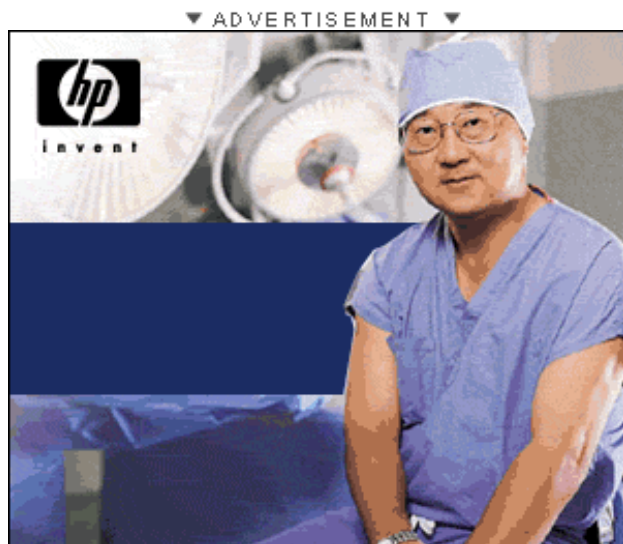


Domesday Book commissioned by William the Conqueror.

By Simson Garfinkel
[The Net Effect](#)
 December 3, 2003

Back in 1986, the British Broadcasting Company embarked on one of the most ambitious digital publishing efforts of all time. Its goal was to build a modern equivalent of the Domesday Book, that massive volume commissioned in 1086 by William the Conqueror to document all of the King's property and subjects. The BBC assembled a massive archive—two videodiscs crammed with photographs, video,

and more than 300 megabytes of text. In total, more than a million people contributed to the so-called Domesday Project.



This past March, a British Newspaper called *The Observer* made a disturbing observation:

16 years after it was created, the £2.5 million [\$4.3 million] BBC Domesday Project has achieved an unexpected and unwelcome status: it is now unreadable.... The special computers developed to play the 12-inch video discs of text, photographs, maps and archive footage of British life are—quite simply—obsolete.

As a result, no one can access the reams of project information—equivalent to several sets of encyclopaedias—that were assembled about the state of the nation in 1986. By contrast, the original

TRY DIGITAL

Get the same great magazine delivered to you without delay.



FEATURES

- Immediate access to current and back issues
- Latest issue delivered one week before print subscribers
- Keyword searches and ability to jump to articles and table of contents
- Ability to pass along issues for FREE

MIT'S MAGAZINE OF INNOVATION
TECHNOLOGY

SPONSORED LINKS

[HP notebooks and desktops.](#)
[Doctor-patient security.](#)

[RHT 2004 Salary Guide – The latest in salary trends!](#)

[Learn about the Qualcomm Launchpad™ Suite of application Technologies.](#)

[Is your salary competitive? RHT 2004 Salary Guide](#)

Domesday Book—an inventory of eleventh-century England compiled in 1086 by Norman monks—is in fine condition ... and can be accessed by anyone who can read and has the right credentials.

This ironic death of Domesday has been taken as a rallying cry for an increasingly vocal group of computer scientists and archivists who argue that we are in danger of losing our cultural heritage—or at least that part of our cultural heritage that we have been foolish enough to commit to electronic storage devices.

There's just one problem with this reasoning: it's wrong.

Recently I was at a conference with David Stork, chief scientist of Ricoh Innovations, the Silicon Valley research center of that giant Japanese office products company. We were there to talk about computer security, but all Stork wanted to discuss was his idea for the “Digital Lock Box”—an as-yet nonexistent service that would allow people to digitally archive information in such a way that they would “be guaranteed, with 99.9999 percent confidence, of being able to retrieve it at least 15 years later.”

As Stork put it to me at the conference, “a Word Pro 2.5 document on a [Macintosh low-density] 3.5-inch floppy, with an Illustrator 2.0 image with an out-of-date compression scheme, cannot be easily retrieved and viewed” even a few years after it is created. Building a system that can store and retrieve digital information with high fidelity is an engineering project that is “worthy of major government, corporate, and academic support.” Stork even has a slogan at his fingertips worthy of a bumper sticker: “Just save it!”

But consider the Domesday project. It's true that the original discs can be played only on a Philips VP 415 Videodisc Player—a system, designed by Philips specifically for the project, that could overlay every frame of extraordinarily sharp analog photo with 6 kilobytes of digital data. But advances in digital image compression technology made the VP 415 obsolete. Domesday was its first and last significant application.

That doesn't mean, however, that the data on the Domesday discs are gone forever. A group of dedicated engineers and electronic preservationists have painstakingly copied the information off the original discs and onto more modern systems. They have also created a computer program that emulates the BBC Micro, the special-purpose computer on which the Domesday system ran. This emulation allows today's standard PCs to play back the original Domesday videodiscs.

To be sure, this has all been an expensive and time-consuming process. But it has been done, proving that the process is possible. Not all digital material is worth preserving—most, in fact, is not. But Domesday was worth preserving and, as a result, it has been.

The real lesson of the Domesday Project is that nonstandard file formats carry a huge hidden cost. Because high-quality image and video compression hadn't been invented yet in 1986, the BBC saved a tremendous amount of money by putting the Domesday Project on a pair of videodiscs rather than stamping the data onto

perhaps a hundred CD-ROMs. But those savings must now be cast against the real cost borne by those who must migrate the data into a modern format.

Indeed, for every Domesday Project that has lost its data to proprietary equipment and file formats, it is easy to point to another project for which information created decades ago is still available. The Internet "Request For Comment" (RFC) series, started back in the 1970s, is readable on practically every computer on the planet today because the RFCs were stored in plain ASCII text. Similarly, you can download images sent back from the Voyager space probes 30 years ago and view them on your PC because NASA stored those pictures as bitmaps—pixel-by-pixel copies of the images without any compression whatsoever.

Some argue that it's impossible to look into the future and determine which of today's formats will survive and which will go the way of the VP 415. Poppycock! As a society we have a very good understanding of what will make one file format endure while another one is likely to perish. The key to survival is openness and documentation.

It is simply inconceivable that documents created today in Adobe's Portable Document Format (PDF), or images stored in the Joint Photographic Expert Group (JPEG) format, won't be decipherable on computers in the year 2030. That's because both the PDF and the JPEG formats are well-defined and widely understood. Adobe has lost control of PDF: there are more than a dozen programs that can create PDFs and display them on a wide range of computers. In other words, PDF is no longer a proprietary format. The same goes for JPEG. Yes, Adobe may fail and new 3D cameras may make two-dimensional photography obsolete. But we will always be able to read files in these formats, because the detailed technical knowledge of how to do so is widely distributed throughout society.

What about the physical media itself? Although there are many examples of tapes and floppy disks being unreadable five or 10 years after they are created, there are many counterexamples as well. Generally speaking, people who make an effort to preserve digital documents have no problem doing so.

Take, for example, the electrical standard (sometimes called IDE, now called ATA) that's used by the disk drives in most PCs. Developed in the 1980s, the ATA interface has been significantly enhanced over the past 20 years. Yet with rare exceptions, you can take a hard disk drive from the late 1980s or early 1990s, plug it into a modern desktop computer, and read the files that the disk contains. That's because the power cables, physical mounting brackets, data connectors, and even the electrical signals used by today's computers are compatible with the old drives. What's more, today's PCs, Macs, and Linux boxes all can read DOS file systems created in the 1980s. If the disk spins, you can frequently get back the data.

Consumer optical storage media has evolved into an even more stable standard. Music CDs and CD-ROMs created in the 1980s are still readable on today's DVD drives. When the next generation of optical storage comes out, it's likely to be backwards compatible as well. A disk drive unable to read old CDs would not be commercially viable.

Electronic archivists do have a significant challenge facing them: computer systems make it easy to put a tremendous amount of information in a single place. If you aren't careful, it's easy to lose all of this information at once. And today's

computer systems are so tremendously reliable that fewer and fewer users are properly backing up their data; people just don't remember the bad old days when a computer might fail at a moment's notice.

But on the whole, I think that electronic records are far more stable, more durable, and more likely to last than their paper equivalents. The technical problems are largely solved. We know how to create David Stork's Digital Lock Box. What's needed now is a plan to make long-term electronic archival services available to the masses.

Simson Garfinkel is an incurable gadgeteer, an entrepreneur, and the author of 12 books on information technology and its impact.

Copyright 2004 Technology Review, Inc. All rights reserved