



## Chaos of the Digital Universe

### Coming Soon to a Hairball Near You

*By Roger L. Kay*

*President, Endpoint Technologies Associates*

A couple of weeks ago, IDC put out a study (paid for by EMC) that estimated the amount of data produced in 2006 and projected for 2010 ([http://www.emc.com/about/destination/digital\\_universe/index.jsp?hpId=1](http://www.emc.com/about/destination/digital_universe/index.jsp?hpId=1)). The figures are astounding. All of us together, mostly individual users, created 161 exabytes of data in 2006. That's 161 billion gigabytes. In 2010, that figure will rise to almost a zettabyte. That's roughly 1,000,000,000,000,000,000 bytes. My eyes go cross just trying to count the number of commas.

Now, there are many scary things about all this data. For example, if you're looking for a needle in it, how will you find it? Will today's search techniques be up to the job? Unlikely. Particularly, if it is stored here, there, and everywhere.

Also, the quality of all this data will be — how to put this politely — uneven. Back in the 18<sup>th</sup> century, an educated person may well have been able to read everything that had ever been published because the archive was not all that large. And to get published, the thing had to pass some minimal standards at a publisher. Today, any fool can publish and does. Profusely. So who figures out what the good stuff is? Where is the trusted arbiter of quality? Brand names will figure prominently here. If you believe the Wall Street Journal, then you'll buy its version of things, and if you believe Jerry Falwell's version of things, then you'll buy into whatever the National Liberty Journal says (assuming either one will be around in any recognizable form).

Meanwhile, much of this data is transient; that is, generated once and thrown away. Streaming video, for example, which accounts for a huge proportion of the data explosion, is rendered, painted, and tossed. Once your eyes have seen it, the data's work is done. Meanwhile, redundant copies of the same video may be wandering around YouTube, Digg, and VideoSift at the same time. Are they the same? Slightly different? Oh, and who owns them? What about copyright, fair use, and royalties? Traditional publishers are hard pressed to deal with these phenomena.

Then, we get into areas like security, compliance, and availability, all falling more or less under the rubric of information management. How do we decide what information needs to be kept near at hand so that anybody can get it quickly? A digital copy of the Declaration of Independence is nice to have on Wikipedia for ready download, but how about all those digital vacation photos you took in Costa Rica? And what about after you die? Do we still need to keep them?

And we haven't even gotten into the legal liability issues facing corporations that store all this data on behalf of individuals who generate or lay claim to it. Say your whole medical record ends up online. There are lots of high resolution x-rays of your innards. They take up lots of space. Did some unauthorized person from the insurance company get in and backdate a scan? How do you know whether those records have been touched and fiddled with by only the right sort of folks? What happens to the company that was supposed to keep all this stuff if it loses some or lets it get tampered with? And what if that company goes

out of business? How can we be sure that people with the proper credentials can actually access these records?

And what about aging and retention obligations? Even extremely sensitive information can become worthless at some point. For example, the memo on the proposed merger is pretty ho-hum after the announcement. On the other hand, sometimes it's the other way around, and a file that was once available to everyone becomes restricted. For example, you can read the New York Times business section online today, but if you want to see the archives, you have to pay. Who is to go over this vast realm of stored data and figure out whether things classified should become open and vice versa? And when do you throw it away? There is no grand adjudicator for all this data, but from time to time, legal authorities may become very focused on certain interesting elements. How do companies keep things straight?

According to IDC, individuals will generate 70% of this vast quantity of digital information, but 85% of it will be stored in organizations, which will have some obligation to manage it with the right balance of security, compliance, and availability. Datacenter managers will find these issues becoming more and more urgent over time. And so we arrive at an explanation of why EMC commissioned the study in the first place. EMC has grown from its roots in enterprise hardware into the information management business through a strategy of organic growth, transformation, and acquisition. The company is positioning itself to address these problems on behalf of its core enterprise customer base.

What does the infrastructure that manages all this data really look like? For the most part, this note raises questions, but you can be sure that over time, as these issues loom larger and larger, EMC and other enterprise suppliers will be generating lots of answers.

*Roger L. Kay is the founder and president of Endpoint Technologies Associates ([www.ndpta.com](http://www.ndpta.com)).*

© 2007 Endpoint Technologies Associates, Inc. All rights reserved.