EDITORIAL

Backup Your Data – The Need for a New Data Archival Format

***Randall J. Allemang**, Contributing Editor*

The admonition to backup your data is not a new concept, and whether we do it manually or automatically on our computers, the problem is two-fold. First, we must choose what to backup, and second, we must choose an appropriate media for the backup. In the computer (desktop/laptop) situation, the information that is selected for backup is normally contained in various folders and files incorporating different formats, some proprietary and some not. The media are generally another disc or a more permanent media like a CD or DVD or possibly a removable USB drive.

For short-term situations, the media are not generally a concern. Since there is a need to backup and retrieve information over a longer period of time, there is no solution for the media limitations except to reread the data from one media and to rewrite the data to a newer media, noting the demise over the years of paper cards and tape, magnetic 800 and 1600 bpi tape, floppy discs, optical discs, digital audio tape (DAT), etc.

The focus of my comments this month concern the backup of measured data and information derived from measured data that originates from experimental dynamic testing. These data are frequently unique and may have an extremely large financial value in terms of what it would cost to replicate the data by performing another test. We have the same issues for these data as we have with the information in our computers with respect to both what to backup and the appropriate media. Information today is generally in some sort of file or folder structure for this situation and is often in a proprietary format that requires the appropriate vendor library to decode.

As a user, I want the data that is backed up for long-term archiving to be in a nonproprietary format so that concerns about software compatibility are minimized. This also minimizes the dependence on a single vendor, a lesson learned the hard way during the days of proprietary minicomputer hardware. The time-sensitive nature of the media is unavoidable, and issues regarding rereading and rewriting the data to different media over time are the same for measurement data as for general computer data.

There are some distinct differences, though, when archiving measurement data. With respect to information in the form of measured data, beyond the financial value associated with replicating the test, the data are often required to be accessible over a lifetime of several years, sometimes stretching for 40 years or more (particularly in the aerospace and defense areas). We have to also realize that some tests cannot be replicated, which means that the existing data are essentially priceless. This means that periodically, the data will have to be reread and rewritten to be maintained in a stable media. This requires that the data endure beyond computer hardware, software and media in some sort of format that is independent of operating system, hardware and vendor.

Much of the data acquired in the digital age but prior to modern computer file structures (around 1980) is probably no longer available unless the specific computer, hardware and software is still operational or the users involved have moved from format to format in addition to moving from media to media. Even data taken up to the present time may have been archived in proprietary vendor formats that may not be accessible unless the associated computer hardware and software is still operational or backward compatible from current computer hardware and software.

In the late 1960s and early 1970s, this problem was largely recognized by the Structural Dynamics Research Corporation (SDRC) and essentially solved through developing a universal file data structure. Many other large companies developed their own internal data archival structure to handle the problem internally in their organizations. While the universal file structure developed by SDRC was not an official data archival standard accepted by any standards organization over the last 40 years, this data structure has become a *de facto* standard and is included as an export option in almost all vendor-based software in the dynamic testing area. Today, the universal file format (UFF) is the only widely accepted, nonproprietary format to store measured data and analysis results derived from measured data throughout the world. Certainly, the UFF structure can continue to be used for the foreseeable future.

However, there are problems with the UFF structure that indicate that this format needs to be augmented or updated to include information and archival methods that were not foreseen 40 years ago. At that time, some of the measurements and data processing methods we now use were not envisioned, and the possibility of storing gigabytes of contiguous time-domain data involving large numbers of channels (100-1000) could not have been anticipated when four-channel data acquisition systems cost $50,000 to $100,000 in 1970 dollars.

Vendors and users now want to store additional metadata associated with the measured data to more completely describe the test or data processing used to derive information from the measured data. While the UFF structure included a set of UFF functions that could be user defined, the addition of other information such as pictures and videos cannot be reasonably incorporated into the existing structure. For complete information concerning the many varied UFF structures, please visit http://www.sdrl.uc.edu/universal-file-formats-for-modal-analysis-testing-1.

If a new data archival structure is going to be developed, the existing UFF structure provides guidelines to what was done correctly and has endured. It also shows what was not done as well or what can be done better based on existing capabilities. For example, one of the most enduring aspects of the UFF structure is that it is ASCII based. While not the most efficient way of storing information from a file size consideration or a database retrieval viewpoint, textual information in the form of the written or printed word has endured for thousands of years. From a storage point of view, any inefficiency is offset by the reduced cost of storage media and methods over time.

A partial list of what needs to be corrected based on the current UFF structure includes: cross-power structures that clearly differentiate input-input, output-output and input-output data, clear procedures for handling read/write errors, inconsistent order-dependent unit issues, data structures based on a computer language (Fortran) which increasingly is unfamiliar to users/vendors, binary information stored in ASCII fields, white space errors (spaces versus tabs), limited ability to store long time record formats (serial versus interleaved), nongeneral, multidimensional matrix structures, addition of named value pairs for information rather than the current numerical codes, etc.

At this point, before we go to the time, trouble and expense of developing a new data archival format, the question of whether a newer format already exists needs to be addressed. Before that question can be answered, a few guiding principles probably need to be agreed upon to help define what any new data archival structure needs. A reasonable set of guiding principles include the following: any new data archival format should be an open format that is extensible and nonproprietary. The new structure needs to be as simple as possible and well documented so that vendors and users will adopt this new structure parallel to the existing UFF structure. The primary use of the data archival structure is archiving data or exchanging data from software package to software package and not as an internal database (so that speed of retrieval does not become a driving criteria). Recovering stored data is the primary concern, so a hierarchal rules structure for detecting and processing read errors must be developed. Finally, a structure that allows both vendors and users to add metadata to supplement the basic structures or to add additional, optional information is also desirable.

Existing data formats have been developed over the last 40 years, but when measured by the these guiding principles, the existing formats generally provide insight on what can be done but fall short in one or more ways. The existing data formats include: Neutral Files, ASAM/ODS, XML, Hierarchal Data Format (HDF) and the work of the IEST WG-DTE042. These data structures exist or are in development, and most were developed by a group of individuals representing vendors, corporate users or individual users. Unfortunately, unless some of the guiding principles can be eliminated, a new data archival format will need to be developed.

Fortunately, work has begun on a new data archive format that will satisfy the guiding principles identified above. Information has been solicited from users and vendors about their concerns and guiding principles. Most are very flexible and simply want the problems addressed. For those of you at IMAC this year, there will be a paper and presentation that summarizes these issues, the progress that has been made to date and the process that will be followed in trying to move to a solution over the next two to three years. I hope to include this paper as an article in an upcoming issue of *Sound & Vibration* magazine to provide more details concerning all aspects of this development. But at this point, there is still time to provide your concerns relative to existing or future data archival formats along with your thoughts on the stated guiding principles or some that have yet to be included. Please contact me if this subject interests you or if you have input concerning existing problems or abuses in current data formats as well as features you believe would be important to a new data archival format.

I hope this gives you something interesting to think about as we begin a new year. As always, I value your comments, please feel free to contact me at randall.allemang@uc.edu. Best wishes for the New Year.