IMPACTS OF DATA FORMAT VARIABILITY ON ENVIRONMENTAL VISUAL ANALYSIS SYSTEMS

Richard Domikis*, James Douglas, Leonard Bisson
The Boeing Company, Springfield VA

ABSTRACT

This paper discusses the benefits and drawbacks inherent in data format variability in the context of environmental visual analysis systems. The current state of system performance can be improved through judicious selection and adoption of data formats. An end-user's ability to choose from a wide range of data formats to use based on his own needs is contrasted with the challenge created by diverse user format needs and the resulting solutions of multi-format devices or data conversion tools. A parallel between digital music format evolution and environmental data formats is drawn, and the lessons learned from the commercial digital music industry are shown to be applicable to environmental community.

1. EMERGENCE OF AUDIO FILE FORMATS

One of the principle benefits of the internet is the ability to find and share information between entities around the world. A natural and necessary precursor to this was the establishment of protocol standards assuring individual entities could understand the information they were sharing. This standards-based infrastructure set the stage for a common means to distribute primarily textual data over the internet. Offline sharing of files requires some level of file standardization. The need for accepted standards is dramatically compounded by the use by and size of the online community. A significant portion of the files and internet traffic is related to multimedia sharing. Until the recent popularity of streaming video, audio files dominated internet file sharing. MP3 remains the de facto standard for audio sharing.

The first half of the 1990's saw unprecedented innovations in personal computing. Processing power was skyrocketing, as were user expectations of software capabilities. Computers were becoming multimedia powerhouses capable of far more than simple calculations; they were digital encyclopedias, compact disc players, gaming stations, and a gateway to worldwide communication through the internet.

Media evolved from analog to digital and from physical to virtual, especially in the area of music.

Corresponding Author Address: Richard Domikis, The Boeing Company Mission Systems; email: richard.domikis@boeing.com

Also at this time, Compact Discs (CDs) had been around for over a decade, with users compiling extensive libraries of CDs. With more and more users relying on computers for everyday use it was only a matter of time before they wanted their music media in a digital format on their computers. Many methods for capturing this media in a digital format were developed. With the advent of the internet, digital music has taken to the web and is now shared seamlessly between users throughout the world on a variety of devices.

Music files require sufficient compression to accommodate the limited data throughput of the early internet and limited storage capacity of early computers. For example, a 700MB CD can store 80 minutes of audio. At this rate, a typical three minute song requires over 26MB of storage space. When files of this magnitude are moved across the relatively slow connections of the early internet, or stored on expensive portable flash memory, compression quickly becomes an important issue. Even in today's world of low storage cost and increased broadband internet access, compression is still a driving design factor. The average user who participates in music sharing typically has thousands of songs they wish to use and share. This enormous amount of data potentially being shared across the internet can quickly exceed the limits of the technology being employed by the users.

While many digital audio formats exist, MPEG-1 Layer 3 (MP3) is arguably the most successful in the digital music industry. Unlike most internet standards, widespread community adoption of the MP3 format was influenced more by the user community than a standards organization or corporate entity; its popularity was driven from the bottom-up. An audio-specific compression algorithm, MP3 capitalizes on psychoacoustic models to discard components of music less audible and less relevant to the listener. Efficient compression is a key factor in MP3's popularity.

In 1988, the MPEG audio standard had fourteen proposed coding schemes, two of which were Adaptive Spectral Perceptual Entropy Coding (ASPEC) and Masking-pattern Adapted Universal Subband Integrated Coding and Multiplexing (MUSICAM). Formal tests led these two schemes to merge, yielding a family of three coding schemes. In 1992 the Moving Picture Experts Group (MPEG) together with the International Standards Organization (ISO) specified MPEG-1, which included three codec formats known as Layers 1, 2 and 3, based on a simple variant of MUSICAM, an optimized version of MUSICAM, and ASPEC, respectively. The comparatively small hard disc storage of the time and the popularity of 28.8kbps modems led the relatively efficient Layer 3 to quickly gather momentum as a music storage format as a result of user demand and product compliance.[1]

By 1995 MPEG-1 Layer 3 donned it's now well-known moniker MP3. At this time MP3 also began to see significant adoption by the commercial sector, as it was the selected audio format for the WorldSpace satellite digital audio broadcasting system. Diamond Multimedia's introduction of the Rio portable music player in 1998 opened the eyes of American consumers to the idea of portable, solid-state music players. The first headphone stereo to play MP3 encoded files stored on memory internal to the music players, the Rio catalyzed a rush of portable compressed-music players into the market, as well as the development of new audio compression schemes. Hard drive-based portable music players began to show up on the market in 2000, as did MP3 compatible portable compact disc players.[2]

The pre-history of the MP3 format is much like the current state of the weather community. There exists a wealth of file formats for storing and transferring weather data. These formats range from open to proprietary, well-known to obscure, and can vary depending on who most recently possessed the data, what type of sensor produced the data, processing routines that may have been performed on the data, and countless other variables. There is no format currently accepted as the overall dominating standard for weather data. To utilize weather data, a user is burdened with converting or preprocessing data, slowing and challenging success. MP3 emerged as a result of the combination of competing standards and solutions coupled with a strong user base.

Unlike the MP3 user community, weather users don't have as strong an influence over the adoption of a community-wide standard. This may be a result of diverse user needs and new system formats. User needs may not be as diverse as they appear. In order to facilitate sharing of weather data across the community, a standard must be adopted. One way to do this is to through a "weather users group". Such an open group would be able to quickly find common ground to begin the specification of a standard. It seems that much of the metadata associated with any given weather product would be common amongst even disparate users. Once a common format is realized, awareness and sharing dramatically improve. A user tends to avoid the import and use of external weather data because of file format complications. While the impact of this varies on a case by case basis, there is no doubt that solutions are hampered by file format compatibility issues. In addition to the technical challenges, cost of building, maintaining and accommodating different formats is significant. A standard must be widely adopted by the weather community to facilitate any kind of mass data distribution.

Hierarchical Data Format 5 (HDF5), a relatively new data format, is designed to solve many of the problems faced by the scientific and engineering communities. It aims to store data in a hierarchical format, support arbitrary data set sizes, allow for smooth integration into a variety of development environments, and recognize certain complex data types. In short, HDF5 was intended to be "a completely new format and software library for data storage, management, exchange, and archiving of large and complex scientific, engineering, and other data.[3]"

This has proven its value to the science and engineering community, being adopted by entities including the Department of Energy, Lawrence Livermore National Laboratory, the Swedish Meteorological and Hydrological Institute, Research Systems, Inc., and Photon research Associates, Inc.[4]

While at first flexible and adaptable formats seem ideal for use in applications such as weather data transfer and processing they are not. It is certainly true that flexibility ensures that future needs can be accommodated; this comes with a price that many choose to ignore. A self defining standard such as HDF5 isn't really a standard but a framework that a standard could be defined within. The goal must be to define not only a framework but an actual standard that is complied with and extended only when needs and acceptance are confirmed. Some have criticized other standards as slow to adopt change. JPEG2000 3D for example has been evolving for some time. This criticism is poorly placed and does not appreciate the true challenge of standard definition and evolution, especially with a large base of users and goals of backwards compatibility.

While its value to these users and their communities is well-supported, HDF5 may not be the ideal data format for the weather community. Most importantly, HDF5 is designed to specify a standard framework for storing data in a hierarchical format. It is not designed to specify a standard for the particular organization of, type of, or metadata about stored data. Consequentially, while a user will be able to read the data in a vendor's HDF5 product, the user may not understand how to interpret the data. Without a guarantee of the organization of the data within the HDF5 file, the user has no predetermined way to ingest and analyze the data. Another issue for the weather community is compression. While HDF5 supports compression of data, it is inherent in the standard. This adds another level of complication to the user's interpretation of data, which may or may not need to be decompressed before it is analyzed.

"Open" has different meanings to different people. For most it implies non-proprietary with some degree of source access. The latter aspect challenges any format since if one can alter and augment the format source do you really have a standard format? Many would say no. There have been a number of excursions into self identifying and conforming standards, Common Object Request Broker Architecture (CORBA) for example. While it is certainly possible to achieve a "meet and greet" interface the common issue seems to be performance overhead. CORBA requires a CORBA engine and so while the interface is self defining the compatibility issue has not gone away but simply moved from the interface to the interface engine. This interface flexibility comes with a price of engine processing overhead and the resulting compatibility issues. It is easy to select a highly flexible "format" and then continue to define proprietary and closed implementations within that format. This is not a path to success for the weather community or any other data intensive area. What is required is a real multi-program weather data format standard that users help define and accept.

2. MP3 VERSUS OTHER STANDARDS

MP3 came into a world of existing audio storage and compression standards. Most of these standards were proprietary, and only a few, such as the Windows Waveform, released in 1991, were commonly seen on the internet. Waveform is not a compressed audio format. It produces relatively large files, challenging the typical internet connection speed of the early 1990s. After the introduction of the MP3, Waveforms declined in use, and are scarce today. The success of MP3 came from the culmination of industry standardization and popularity among users. Without the standard, users would have little interoperability. Without the popularity among users, MP3 may never have become the de facto standard in digital music.

A company that developed a portable music player that didn't support MP3 would experience a business disaster, as the libraries of music amassed by potential customers are the driving force for demand in support of audio formats. Many digital audio file formats were born after MP3, including Real Audio Media in 1995, QuickTime Audio in 1997, Windows Media Audio in 1997, Advanced Audio Coding in 1997, Audio Interchange File Format in 1999, OGG Vorbis in 2000 and mp3PRO in 2001, and most can be found to varying degrees on the internet, but none enjoy the current popularity and success of MP3. [5]

3. SUCCESSFUL DATA SHARING

Today, music file sharing on the internet centers around the concept of Peer-to-Peer (P2P) file sharing networks. The P2P paradigm is generally considered to have evolved through three generations. The first generation was based on a centralized list of files, which end-users would search for target files. Search results would link to target files hosted from other end-users' computers. This was the original basis of the design of the well-known Napster. This ultimately failed as United States courts ruled that the responsibility of any copyright infringement was placed on the entity that controlled the file list. P2P has become so popular that commercial vendors have included support for these concepts in their products.

The second generation saw the growth of networks without a centralized index server, and out of this grew Gnutella. All end-users, called nodes, were considered equal players in the network, which created bottlenecks from the immense user load. This problem was tackled by weighting nodes by their capacity, and allowing lower capacity nodes to branch off of the higher capacity nodes, which served as indexing nodes. This model is still followed by P2P networks today. Distributed Hash Tables (DHTs) were also introduced in the second generation. Various elected nodes would serve to index certain hashes, allowing for faster searching of files. A major drawback of DHTs is the inability to perform keyword searching, rather than simply exact-match searching.

The third and emerging generation offers what many view as a major advancement the introduction of anonymity. These networks are not yet as widespread as those of the second generation, but serve to protect the end-user from external snooping. By routing traffic through other clients, analysis of traffic patterns from origin to destination is resisted. Encryption is also commonly used. All of these added features introduce a new degree of processing overhead, further contributing to the slow adoption of this generation.

Marked increases in the volume of shared data led data brokers and libraries to give way to P2P sharing. This new idea of data swapping fostered concerns over the ethical issues of "sharing" copyrighted data. While certainly beyond the scope of this paper it must be recognized that even from unethical activities valuable concepts are created. Music was in a digital data format long before being shared over the internet, sold and distributed physically in the form of audio compact discs. It was natural that it would make the move from physical media to data files stored on a computer hard disc. It was also natural that vendors would seek to differentiate themselves to gain market share and a loyal customer base. This is not unusual, especially in the music industry, where numerous improvements in audio reproduction competed. Users were presented with products vying for acceptance. At first the individual investment in buying or converting music was accepted as necessary. This quickly gave way to a phenomenon similar to the "Hayes modem compatible" phenomenon where vendors recognized if a user's library was not playable as is they would not sell a new device. All of this was occurring while the quality of audio improved and file size decreased.

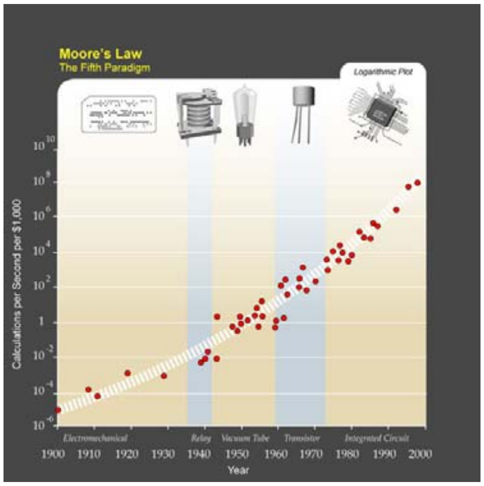
The current state of the weather community is much like that of the music community before the adoption of MP3. Numerous formats challenge users and systems. Some of these formats are better known than others, some are open, some are proprietary, and there is no clear standard. User needs, in combination with system acquisition authorities must define the path forward. To be successful, this path must include cross-system and cross-industry compatibility.

4. PERFORMANCE AND GROWTH

A major feature that MP3 lacks is the ability to store more than two channels of audio. MP3 was designed to support monaural and stereo audio data. It wasn't until later that the increasing popularity of "home theater" systems, often with five or six audio tracks, created a demand for an audio compression scheme which supported more than two tracks. This demand resulted in the creation of MP3 Surround, which maintains backwards compatibility with MP3. MP3 Surround is built around the Binaural Cue Coding (BCC) concept of parametric representation of spatial audio. BCC allows for audio composed of an arbitrary number of channels to be represented as a single track of data, combined with some side information. MP3 Surround maintains compatibly with MP3 by behaving as stereo data on a standard MP3 decoder. This required a deviation from BCC to represent audio data as two tracks instead of one. This stereo data can be scaled up using MP3 Surround ancillary data to support an arbitrary number of channels.

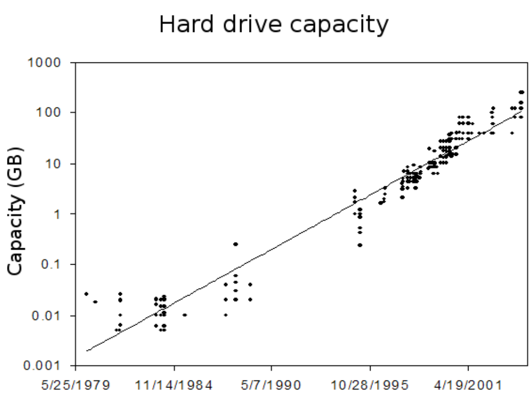
The lesson to be taken from MP3 is that a standard must be able to react to the needs of the users while concurrently maintaining compatibility. An imagery standard data format that parallels MP3 is JPEG2000. As the population of users expands so does the required needs of JPEG2000. Examples of responsiveness to user needs are the inclusion of 3D imagery, increased support of metadata, and an increase in the number of formats supported within the standard such as floating point numbers. The improvements to JPEG2000 may make this a more attractive format to standardize around. This is an area for further discussion and could be a starting point for a weather data users group.

Along with these improvements in the MP3 standard there are improvements being made in processing and storage. Moore's Law shows that processing power is doubling roughly every 18 months [6].



Moore's Law - 18 month processor doubling [7]

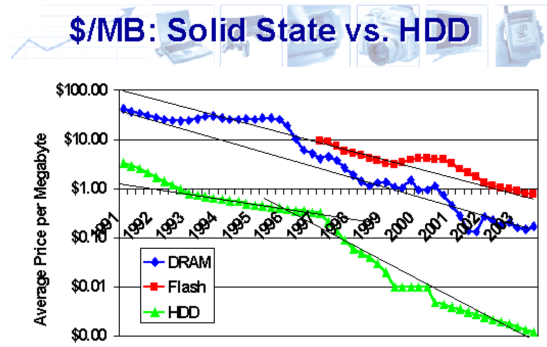
Similarly, Kryder's Law shows that storage capacity is doubling roughly every 23 months.



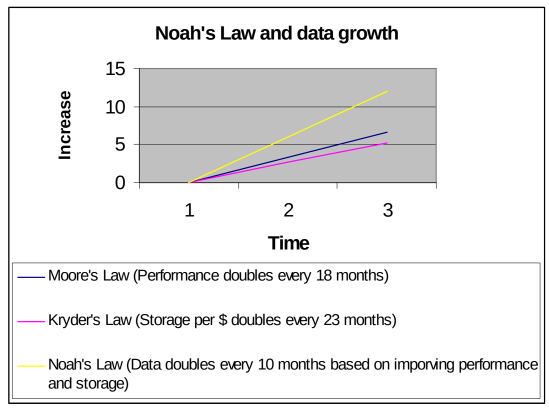
Kryder's law - 23 month storage density doubling [8]

While the amount of time it takes to double performance and storage may be argued it is still evident that both processing and storage are still increasing rapidly. The combination of the increasing processing power and decreasing cost of storage is enabling another phenomenon to occur. The result is an increase in the amount of data being created; we refer to this as Noah's law. Noah's law is supported by increasing internet traffic patterns, which according to the London Internet Exchange has increased by over three orders of magnitude over the last decade, and increasing storage capacities.

The figure below depicts the clear trend in decreasing cost of data storage over time.



Cost of storage over time [9]



Noah's law - 10 month data doubling

This increase in data described by Noah's law infers that existing networks and systems may become overwhelmed by the amount of data being shared. The historical solution of simply increasing the number of processors and storage is no longer a viable option. The antiquated solution of "throwing more iron at the problem" doesn't work. Users will need to employ more effective tools and methods to fully take advantage of the increased amount of data available.

5. CONCLUSION

The history of MP3 lays the foundation for a user-based emergent standard. The parallels between the music and weather industries make MP3 an interesting case study. The weather community is in a position to take advantage of a huge amount of data being generated. The adoption of a user-based standard is a key step in the facilitation of data sharing and the enabling of technologies associated with the weather community. Open data that is easily shared among users will be a catalyst for users on the fringe of the mainstream weather community. The enormous amounts of data that would be available would enhance efforts in many areas of research, including the use of data mining to improve weather forecasts as discussed in “Application of Decision Support Methods to Weather Sensitive Operations”, presented 2007

The weather community should appreciate that there is a larger set of potential users with equally challenging interface issues. These users are represented by the Open Geospatial Consortium’s (OGC’s) Geographic Information Systems (GIS) group [10]. Since weather is often visualized and Earth referenced the challenges of formats, processing and display are similar to current challenges in the GIS community. This community has provided a definition of “open” that the weather community should take note of: “Open GIS is the full integration of geospatial data into mainstream information technology. What this means is that GIS users would be able to freely exchange data over a range of GIS software systems and networks without having to worry about format conversion or proprietary data types[11].

Embracing this definition, “open” for the weather community would be: “Ability to use and freely exchange data over a range of weather systems and networks without having to worry about format conversion or proprietary data types.” While it is possible to create a weather user group, joining an existing group such as OGC offers immediate benefits and experience.

Open implies flexibility and support; it does not necessarily result in broken, user by user or system by system solutions. The community must seek unifying formats for cost, performance and interoperability.

The largest lesson to be taken from the evolution of MP3 is that proprietary and industry-mandated formats deter use and fracture the user community. The end users must adopt a common standard to support their diverse needs.

References

1 The History of MP3. MP3Licensing.com. 2 Nov. 2006 <http://www.mp3licensing.com/mp3/history.html>.

2 ibid

3 HDF5 Wins 2002 R&D 100 Award. National Center for Supercomputin Applications (NCSA). 2 Nov. 2006 <http://hdf.ncsa.uiuc.edu/HDF5/RD100-2002/All_About_HDF5.pdf>.

4 ibid

5 A Quick "Rip" Through Digital Audio File Formats. Ed. Ed Tittel. 30 July 2004. Informit.com. 2 Nov. 2006 <http://www.informit.com/articles/article.asp?p=212411&rl=1>.

6 Moore's Law. Ed. Ed Tittel. Wikipedia The Free Encyclopedia. 2 Nov. 2006 <http://en.wikipedia.org/wiki/Moore%27s_law>.

7 ibid

8 ibid

9 Flash Memory vs. Hard Disk Drives - Which Will Win? Ed. Jim Handy. 6 June 2005. SEMICO Research Corporation. 2 Nov. 2006 <http://www.storagesearch.com/semico-art1.html>.

10 Welcome to the OGC. Open Geospatial Consortium Inc. 2 Nov. 2006 <http://www.opengeospatial.org/>.

11 What is Open GIS? Open Geospatial Consortium Inc. 2 Nov. 2006 <http://gislounge.com/features/aa030600.shtml>.