



HHS Public Access

Author manuscript

J Clin Neurophysiol. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

J Clin Neurophysiol. 2016 October ; 33(5): 403–413. doi:10.1097/WNP.0000000000000257.

A Proposal for a Standard Format for Neurophysiology Data Recording and Exchange

Matt Stead, MD PhD and

Department of Neurology, Mayo Clinic, Rochester, MN, USA

Jonathan J. Halford, MD

Department of Neurology, Medical University of South Carolina, Charleston, SC, USA

Summary

The lack of interoperability between information networks is a significant source of cost in healthcare. Standardized data formats decrease healthcare cost, improve quality of care, and facilitate biomedical research. There is no common standard digital format for storing clinical neurophysiologic data. This review proposes a new standard file format for neurophysiology data (the bulk of which is video-electroencephalographic data), entitled the Multiscale Electrophysiology Format, version 3 (MEF3) which is designed to address many of the shortcomings of existing formats. MEF3 provides functionality which addresses many of the limitations of current formats. The proposed improvements include: (1) hierarchical file structure with improved organization; (2) greater extensibility for big data applications requiring a large number of channels, signal types, and parallel processing; (3) efficient and flexible lossy or lossless data compression; (4) industry standard multi-layered data encryption and time obfuscation which permits sharing of human data without the need for deidentification procedures, (5) resistance to file corruption; (6) facilitation of online and offline review and analysis; and (7) provision of full open source documentation. At this time, there is no other neurophysiology format which supports all of these features. MEF3 is currently gaining industry and academic community support. We propose the use of the MEF3 as a standard format for neurophysiology recording and data exchange. Collaboration between industry, professional organizations, research communities, and independent standards organizations are needed to move the project forward.

Keywords

Neurophysiology; EEG; VEEG; electroencephalography; video-electroencephalography; standard format; common data format; compression; encryption

In 1999, the Mars space orbiter exploded when incoming data were misinterpreted. The incoming data was assumed to be formatted in Standard International (SI) units when they were actually in U.S. customary units (Kush et al., 2014). Standard data formats are very important in medical science as well. Standard formats for storing data create interoperability between health information systems which promotes efficiency and

Conflicts of Interest: Neither author has conflicts of interest to disclose.

Conference Presentation: none

facilitates collaboration in scientific pursuits. The lack of interoperability between information networks is a significant source of cost in healthcare. In 2005, it was projected that full interoperability between healthcare providers could potentially save \$77.8 billion per year in the United States alone (Walker et al., 2005).

There is no common standard digital format for storing neurophysiological data, the bulk of which is video-EEG (VEEG) data. Many physicians who practice clinical neurophysiology are not aware of this or don't consider it a pressing issue. This is because the lack of a common format does not impede typical clinical neurophysiology workflow, which consists of recording, reviewing, and archiving VEEG data. But, like the Mars space orbiter in 1999, in certain situations, the lack of a common neurophysiology recording standard can cause productivity to "crash and burn" because of poor interoperability. Neurophysiology researchers are more aware of this problem because it causes them extra work and frustration due to poor interoperability between the equipment of different original equipment manufacturers (OEMs) and difficulties in exchanging data between institutions. But perhaps the voices of researchers are not heard as well by the OEMs since the market size for research applications is less than that for non-research medical applications. This article describes basic background information about the importance of medical data storage formats, a brief history of the main data formats in clinical neuroscience, and proposes the use of a new format, the Multiscale Electrophysiology Format version 3 (MEF3), as a new universal standard format for neurophysiological data.

FILE FORMATS: DEFINITIONS AND TECHNICAL ISSUES

A file format is a standard way that information is encoded for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats usually have a published specification describing the encoding method, and most EEG vendors will provide a format specification description upon request. There are multiple characteristic features of neurophysiology file formats:

1. Flat versus hierarchical file system

Most older EEG formats were "flat" formats in which all of the data was placed into one computer file. The advantage of this approach was simplicity. However, there are two major disadvantages to a flat file system approach. First, as the file size increases, a single data file can get so large as to be difficult to use. Second, if the single data file gets damaged, this can cause the loss of the entire VEEG recording instead of the loss of only one of many files. Most modern formats store data in a hierarchical system involving an organized structure of many files for each recording.

2. Binary versus eXtensible markup language (XML)

Some file formats encode data in binary form, meaning that data is encoded in a compressed form more akin to computer language that cannot be read by typical text editor software. Binary files have the advantage of being able to store data more compactly and can therefore be transmitted more quickly. On the other hand, XML formatted data is easier for humans to read and edit, theoretically decreasing the cost to develop software to read the format.

Although the use of XML has been popular in medical data storage formats for the last few decades, it has several significant disadvantages in comparison to binary formats. First, an XML format produces larger (less compressed) files. Second, human readability/editability may not be a significant advantage for XML because modern VEEG recordings are so complex that they are probably not accessible to anyone but a neurophysiology-savvy programmer anyway. Third, it is more difficult to code programs which can process the text found in an XML format as efficiently as binary data can be processed, leading to decreased system performance.

3. Compression

In information theory and computer science, data compression involves encoding information using fewer bits than the original representation, making it possible to store data using less storage space. Most conventional video and audio formats use efficient data compression but most of the current EEG formats from the OEMs do not. This is a problem considering the growth of continuous EEG monitoring, which is greatly increasing the amount of EEG data which must be stored.

4. De-identification

As with most medical recordings, EEG files have protected health information (PHI) embedded in them. This PHI usually consists of patient names, birth dates, and medical record numbers but it also includes date and time of the recording. Many common vendor formats have PHI placed in multiple locations within the data file structure of each individual recording, often redundantly. This PHI is often not handled separately from the rest of the EEG recording such as channel specifications, annotations and EEG signal data, making it difficult to remove during a de-identification process.

5. Encryption

This is the process of encoding messages or information in such a way that only authorized parties can read it. Encryption does not in itself prevent interception, but denies the message content to the interceptor. Although health data encryption has not been a requirement of the Health Information Portability and Accountability Act (HIPAA) in the past, it has been encouraged but not enforced. In 2013, the federal HITECH Act added additional penalties for accidental release of non-encrypted PHI (Rath, 2010). Most neurophysiology OEM formats do not currently provide for encryption of PHI. This may change in the future as companies seek to provide data formats which minimize the risk of HIPAA violations in the US.

WHY A STANDARD NEUROPHYSIOLOGY FORMAT IS NEEDED

Clinicians frequently see patients who have had an EEG or VEEG performed at another practice or institution. If the patient brings the previous study on portable media, the data often arrives without software which can allow it to be visualized. Many physicians don't have access to Persyst™ or other software which can allow them to view VEEG recorded in alternative formats. Even if software for visualizing the VEEG comes with the data, often this software is difficult to install due to operating system issues. If the software can be

successfully installed, the clinician cannot use his or her customary montages because they are usually not provided. Attempting to recreate these montages in this unfamiliar software environment is difficult and time consuming. The typical outcome of this scenario is that the physician does not get access to the past VEEG data or is forced to perform a suboptimal review the VEEG data in a software system that is unfamiliar. A common VEEG format would allow any clinician to call up any VEEG data in their own familiar EEG viewer and/or load it into their existing VEEG database.

Even within a single practice or institution, the lack of a common format can make it difficult to view EEG recordings. This usually occurs when new VEEG equipment is purchased from a different vendor. Within a few years, the older VEEG equipment is taken out of commission and it is difficult or impossible to view old recordings made with the previous vendor's equipment. A common recording standard would allow older data to be called up using the new system and loaded into the new VEEG database, even if it was from a different vendor. Physicians are sometimes hesitant to change from one VEEG vendor to another because of the problem of format interoperability. Consequently, the lack of a standard format is at times anti-competitive.

In the field of clinical neurophysiology research, the lack of a common recording standard causes multiple problems and inefficiencies. First, since most large EEG research projects involve sharing of data between institutions which often have EEG systems from different vendors, unless the research team includes computer programmers who have an understanding of EEG storage formats, all EEG data has to be translated into a common format in order to be shared or third-party software must be purchased which can read multiple formats. If all data is to be translated to a common format, the European Data Format (EDF, or EDF+) is often chosen. But translation of all data into this format takes considerable effort and the format has significant limitations including no support for video, no compression, limited dynamic range, and poor extensibility. If the project leaders cannot parse the VEEG formats themselves for input into data analysis platforms such as Matlab or Python, they must purchase expensive third-party software (from such vendors as Persyst Development Corporation or Compumedics) to view and analyze the VEEG recordings. Sometimes even VEEG recordings from a single OEM cannot be viewed by the current VEEG review software from that OEM, only using third party software, because the recordings were acquired using an earlier version of its equipment which used a slightly different VEEG recording format. Second, because methods for de-identification of recordings are not standardized, functions for VEEG data de-identification provided by the OEMs may not work as advertised. De-identification of VEEG data has to be verified carefully and this is made more difficult by existence of so many different neurophysiology formats. Recently, researchers performing a multi-site National Institutes of Health (NIH) study discovered that the de-identification function of one of the market-leading OEMs was removing PHI from only parts of their VEEG files but not others, leading to research protocol violations. Disclosures to institutional review boards (IRBs) had to be made and requests had to be submitted to the OEM to correct their de-identification function. (The changes to the software were eventually made but took years to take effect.) Third, if researchers attempt to write software which parses the proprietary formats of VEEG OEMs, they sometimes find that it is a harder task than they had anticipated. OEMs will usually

provide a specification file which describes how their VEEG data is encoded in their file format, but this file can be very incomplete and not describe how data is encoded in all situations. This requires that the researcher expend effort to read many VEEG files and discover the missing parts of the file format specification through a trial-and-error approach.

STANDARD FORMATS IN MEDICAL IMAGING AND ELECTROPHYSIOLOGY

There have been multiple projects to create standard data formats in other fields of medicine. The efforts in the fields of electrocardiography and imaging are particularly informative as these fields are so large. Previous efforts have also been made to create a standard neurophysiology format.

Electrocardiogram (ECG) Formats: Partial Success but No Universal Standard

Despite being the most common electrophysiologic medical recording, there is not a standard clinical ECG data format. There have been multiple attempts at creating a standard. The first attempt was in the 1980s and culminated in the creation of the Standard Communication Protocol for computer assisted ECG (SCP-ECG) format. In the 1980s and early 1990s, the European Union funded the creation of a standardized ECG database to train automated ECG analysis software development (Willems et al., 1989). Researchers in this project, led by Jos Willems MD, PhD, worked in a collaborative manor to achieve consensus on the design of a common ECG recording format among manufacturers, researchers, and physician end-users in the US, Japan, and Europe. In 1993, this format was named a standard by the Association for the Advancement of Medical Instrumentation (AAMI). A special consortium was formed, called OpenECG, to promote the consistent use of the SCP-ECG format by all manufacturers and researchers. This organization used a multifaceted approach to promoting the standard: (1) A web portal was created where news about the format could be disseminated and software tools could be downloaded, (2) format development and implementation guidelines were developed, and (3) programming contests were held with awards for the best format-compliant tools (Chronaki et al., 2002). But limitations to the format forced developers to revise it in 2005 and the format was not approved by the International Organization for Standardization (ISO) until 2009 (Trigo et al., 2012). The SCP-ECG format never took hold as a universal format because it was adopted primarily for short-duration ECG recordings and because most of its proponents were concentrated in Europe. Yet there was great interest and use of this standard format in the scientific community and it is one of the largest initiatives in the history of medical informatics standardization (Zywietz, 1998).

Another standard ECG format initiative was launched by the Food and Drug Administration (FDA) in 2001. Because ECGs were frequently used to help evaluate the safety and efficacy of candidate drugs, ECG recordings needed to be collected and inspected by the FDA. The variety of different ECG formats was a hindrance to this work, so the FDA partnered with the Health Level Seven (HL7) medical standards organization to create a new standard ECG format, called HL7 aECG. The FDA then mandated that researchers submit their ECG data to the FDA in this format. Why the FDA did not use the SCP-ECG format instead of creating their own is not clear, but it may be because the FDA had decided that they

preferred an XML format rather than a binary format (like SCP-ECG), due to greater ease-of-use. The developers of SCP-ECG made changes to their annotation nomenclature so that it better conformed to the HL7 aECG format and software conversion tools between the two formats were also created (Schloegl et al., 2007).

For a time, SCP-ECG and HL7 aECG were the only two standardized ECG formats, but two other formats were later adopted and used broadly as well. Although the Digital Imaging and Communications in Medicine (DICOM) format was primarily intended for medical images, new features were added to support other diagnostic modalities, including ECG. In 2000, a cardiovascular working group for DICOM released DICOM Supplement 30, which has been used by some ECG equipment vendors. Its usefulness as a standard format has been controversial since implementing a DICOM system takes considerable effort and may not be cost-effective (Trigo et al., 2012). Another standard format, called Medical Format Encoding Rules (MFER), was developed by a Japanese group (Medical waveform description Format Encoding Rules, January 2003) because they wanted a format which could encode all medical signal data including ECG, EEG, and respiratory data. The MFER format became an ISO standard in 2007. There are many other less-broadly adopted ECG formats, including those particular to storing long-term ECG recordings from Holter monitoring. Overall, despite considerable efforts in the ECG research community, one common standard recording format never took hold and there continues to be multiple formats in use worldwide.

The DICOM Imaging Format: One Format to Rule Them All

In contrast to the story in ECG, medical imaging has successfully developed a common standard format which has been broadly implemented. Plans for standardization began earlier than for ECG, in the early 1980s, which may partially explain the triumph of standardization in this field. In 1983, the American College of Radiology and the National Electrical Manufacturers Association (NEMA) started to collaborate with the goal of developing a standard format. The result of this joint effort was the ACR-NEMA standard, which was released in 1985. This standard was unsuccessful, mostly because it did not contain methods for network communication. A new standard format, called DICOM, was developed (using some lessons-learned from the previous format) and first released in 1993. This standard used an object-oriented design which grouped different types of imaging data into “information objects” which enabled the format to include a wide diversity of imaging data. The standard also included a network communication protocol, which specified exactly how data was to be transmitted between DICOM-compliant devices. DICOM was first released in 1993 and it became a European pre-standard in 1995.

DICOM is not just a standard format but also is an independent international standards development organization administered by NEMA’s Medical Imaging and Technology Alliance. All current digital image-acquisition devices produce DICOM images and communicate through DICOM networks. DICOM files and messages use more than 2000 standardized attributes (specified in the DICOM data dictionary) to convey various medical data such as patient name, image color depth, and current patient diagnosis. The number of vendors and devices which use the DICOM format is enormous. The DICOM organization

does not certify devices as DICOM-compliant; vendors self-certify that their devices are compliant. Perhaps the reason that radiologists and medical imaging device manufacturers adopted unified standard is because it is necessary considering the complexity of medical imaging. There are so many devices in medical imaging facilities that must communicate to acquire, process, store, and display data that without a standard, efficient clinical practice would not be possible. Even with a common standard, there are often interoperability problems between imaging devices because often device manufacturers do not implement all aspects of the complex DICOM standard (Pianykh, 2008).

Neurophysiology Formats: Previous Attempts but Continued Poor Interoperability

There have been three past efforts at EEG format standardization. None have produced a format which has been used for VEEG recording by equipment manufacturers. The formats have been used as a method of transferring data between researchers. The first standard EEG format, the American Society for Testing and Material (ASTM) E1467-92 standard, was developed by US researchers (Jacobs et al., 1993). This was a comprehensive standard developed to be HL7 compliant. Later, two additional formats were developed by European workgroups. The Extensible Biosignal (EBS) format was similar to the ASTM format except that it had the additional advance of using the international character set ISO 10646 (making it more appealing to international users) and supported a more efficient binary encoding scheme (Hellmann et al., 1996). The European Data Format (EDF) (Kemp et al., 2003), developed to store EEG and sleep medicine recordings, has achieved the broadest use. Unlike the ASTM and EBS formats, the EDF format was a simple binary format which was designed to be a primary recording format. A few vendors adopted it as their primary format, but most developed their own proprietary formats. Nevertheless, it is currently the most prevalent format for researchers to exchange EEG data. The EDF format has multiple short-comings which prevent it from being considered as a viable modern standard including lack of support for video, no data compression, restricted dynamic range, and lack of encryption for PHI.

NECESSARY INGREDIENTS FOR A STANDARD NEUROPHYSIOLOGY DATA FORMAT

Multiple features are needed for a standard format suitable for neurophysiology data storage in the 21st Century (see Table 1). Because the amount and types of data recorded are indubitably going to continue to grow, a format must be able to scale for current and future big data applications. This means that the format should allow the signal data to be stored in high resolution, allowing for future improvements in sampling rate, video resolution, channel types, and number of channels. Due to increasing storage requirements, efficient lossy and lossless data compression functions should be provided. Increased government regulation is also likely, requiring an organized approach to the location of PHI within the format and options for encryption of all or part of individual datasets. An ideal format would be hierarchical (i.e. contain multiple files for each recording), to mitigate damage if a file is corrupted and facilitate recording different channel types and different sampling frequencies. But the format should not be so hierarchical as to create a byzantine collection of thousands of files for a multi-day recording. Multiple data stream recording and storage, a feature of

many new high-resolution EEG amplifiers, should be supported in the format, allowing clinicians to easily view downsampled EEG data. Any future format should be nonproprietary with freely available and accurately defined specifications. The format should also support international language specifications and international standard nomenclature for electrode positions and common annotations (Herman).

Whereas broad implementation of an appropriate standard neurophysiology format would obviously improve data sharing between individuals and institutions, the format must also provide excellent support for both clinical and research applications. Because of the expansion of web-based tools for EEG review (Halford et al., 2011), the format should provide support for both offline and online review of data. Full and accurate documentation should be provided as well as open source freely available development tools. The format should be maintained by an independent organization which provides updates to the format at specified time intervals and which is responsive to any changes that are needed in the neurophysiology community. Creating such a standard format will require one or more sources of funding and require broad industry and academic support.

MULTISCALE ELECTROPHYSIOLOGY FORMAT, VERSION 3.0 (MEF3)

The Multiscale Electrophysiology Format, version 3.0 (MEF3) is the only neurophysiology format which fulfills all of the requirements described above, except as yet items 10 and 11 from Table 1. This format was designed to accommodate large amounts of high-resolution data produced by multiple federal grant projects. For example, in the International Epilepsy Electrophysiology Database (NINDS/NIH U24 grant #NS63930), over 200 terabytes of EEG data have been acquired at a 32kHz sampling resolution. The remainder of this article provides an overview of the features of MEF3 and how they fulfill the requirements listed in Table 1.

1. Hierarchical file structure with flexible segmentation

The hierarchical structure of MEF3 is designed to be intuitively accessible to human readers. Naming in the MEF3 hierarchy is specified via filename extensions, and designed to provide consistency, human readability, and ease of use in file system level operations. Figure 1 displays the MEF3 file type hierarchy and naming conventions. Channels are stored within a Session Directory. A session is defined as the collection of all channels associated with a recording. All levels of the MEF3 hierarchy are assigned their own universally unique identifiers (UUIDs), which are shared by all the files in that level. A UUID is an identifier standard used in software construction which is simply a 128-bit value, the meaning which is defined at each bit by any of several variants.

MEF3 stores each channel in its own directory. Each time-series sample is stored as a 32 bit integer, allowing adequate bit depth for most electrophysiology applications. Data for each channel are stored in an individual Time Series Segment folder within the hierarchy in a compressed format (RED) discussed below. Data are stored in independent contiguous blocks that are indexed in a separate file (Time Series Indices file), allowing rapid random access. Identification of discontinuities is often important in data analysis and review so recording discontinuities are also permitted and marked *within* a data segment. Video stream

data are stored in any of the standard video formats (e.g. MPEG) and indexed according to frame number and time, in the Video Indices file.

Clinical EEG recording systems generally segment their prolonged recordings by interrupting recording briefly, closing all files and re-initiating acquisition on new files. Doing so allows acquisition systems to be rebooted at scheduled times and facilitates the movement of completed recording segments to other file systems to conserve space or for backup. MEF3 explicitly supports segmentation of channel data. However, segments can be of any length, so an entire channel's data can be stored in a single segment. Segments are stored as directories within the channel directories.

Storing each channel in its own set of files (channel independence) has two significant advantages. First, channel independence allows for storage of signals in an efficient manner if some channels have different sampling rates than others. Variability in sampling rates among channels is not uncommon in modern neurophysiology recordings since different channels sometimes contain signals with different frequency content, and therefore different optimal sampling rates. Secondly, channel independence makes parallel processing easier since multiple processing cores can act on multiple channel data files at the same time. Since parallel processing is one of the most important trends in computing over the last decade, this feature is critical to the analysis of modern data sets. In keeping with this requirement, care has been taken with the MEF3 library code to maintain thread safety, meaning that the code only manipulates the format data structures in a manner that guarantees safe execution by multiple threads at the same time.

Real-time viewing or analysis software can read the file structure using the same software as those used for reading completed files. A potential weakness of this approach is that the number of open files during data acquisition may be quite large. However, system open-file limits are soft limits and can be adjusted in software. Modern operating systems routinely handle many thousands of open files without difficulty. If the large number of open files becomes a problem in the future, an alternative approach could be implemented such that some or all of the files could be opened, written to, and re-closed at block boundaries (as data is recorded block-by-block) by independent program threads with no need for acquisition interruption. The number of samples in a block can vary, but is most often a fixed user-selected number.

2. Extensible and flexible support for big data applications

There is no limit to the number of channels which can be included in a recording. Currently only two channel types are supported, time series channels and video stream channels, but the format is designed to accommodate other channel types as required in the future without format revision. Many other channel types, such as image stacks, may be desired in the future, and can be easily incorporated into the format. The MEF3 format is highly customizable and extensible. The format contains large regions of reserved "discretionary" space in all of the major file types. This space is reserved for end users to add their own data of various types to the MEF3 files. There is no expectation that the format of the data stored in these regions will be shared with the community at large. Additionally, the session, channel, and segment levels of the hierarchy contain optional indexed record data files. Each

record in a record data file begins with a prescribed record header that allows a record to be skipped if its format is unknown to the current software. The body of the record is end-user defined and can contain any type of data. Examples include online spike detections, system logs, seizures marks, behavioral state, miscellaneous notes, and images, but there is no restriction on the type of data that can be stored in a record. Generally useful record formats will be shared with the community and incorporated into the standard. Extensibility is provided through a large “protected region” (not available for general use) in all MEF3 files, provided in parallel with each discretionary file region mentioned above, which is reserved for potential future additions to the format.

3. Efficient lossless or lossy data compression

Time series data are efficiently compressed in a lossless, or optionally lossy, manner. Data compression can significantly reduce storage requirements and the time required to read/write and transfer data files. Additionally, by speeding up the read/write process, time required for data analysis and trending can be substantially reduced. This is particularly important for very large datasets such as those found in intracranial monitoring and/or research applications. Video streams are already efficiently compressed with their native format (e.g. MPEG), and so further compression is not required.

The time series compression algorithm used in MEF3 is called “Range Encoded Differences” (RED). RED is an adaptive algorithm and compresses in rough proportion to the local information content. The algorithm generates time series differences from a data block. It models the frequencies of these differences and stores this information in the block header. The difference model is used to range encode the differences; a procedure similar to arithmetic compression that consumes encoding space in inverse proportion to difference frequencies. There is no token assignment in range encoding, rather a numerical range is consumed according the frequencies of the sequential differences, accomplishing fractional bit compression. On typical EEG data RED achieves 80-90% compression losslessly (see Brinkmann et al., 2009a for more details). Prior to generating differences, two optional steps can be performed. The first is detrending of the data in the block. This procedure is lossless and can be useful in either lossless or lossy compression, but generally is more useful in lossy compression. The block data can also be scaled, and this is a lossy operation. If detrending is opted for, this is performed prior to scaling. The scaling of the data reduces the distribution of the differences, allowing higher compression during the range encoding step. The steps in RED compression are graphically demonstrated in Figure 2. Decompression is simply the reverse of the compression process.

By default, RED compression is lossless, but three lossy compression modes are also currently supported, though others could be easily accommodated. All compression is performed block-wise, and is therefore adaptive. Lossy compression is useful if the sampling rate or bit depth vastly exceeds the storage space necessary to encode the true information content of the signal, such as in conditions of high baseline noise or unnecessarily large dynamic range. Lossy compression can also be useful in generating temporary lower-fidelity versions of the data in situations where full resolution is not critical (e.g. the transmission of

data from a server to a viewer where performance needs outweigh the need for the fine structure of the signal).

Lossy compression is accomplished by (optionally) detrending the data in a block and applying a floating point scaling factor to the data prior to compression. The loss is introduced by rounding the scaled data to the nearest integer value prior to applying the RED algorithm. The reverse sequence is performed in decompression. Detrending itself is an inherently lossless transform, however it substantially improves the performance of lossy compression in signals with large offsets, or trends by reducing the range that the scale factor will need to span to achieve a desired compression level. The three lossy compression modes supported by MEF3 at this time are: (1) use of a fixed scale factor, specified by the user; (2) automatic selection of the minimum scale factor to achieve a user-specified compression ratio; and (3) automatic selection of a scale factor that maintains a user-specified minimum data fidelity. Lossy compression can introduce significantly higher compression ratios with little compromise of data fidelity (see Figure 3). Some lossy compression modes may increase the time necessary for compression, but add negligible time to decompression. The independence of blocks and channels in the MEF3 format lends itself to parallel processing and so speed losses are easily compensated for by multithreaded applications which take advantage of decreased disk write or network transfer latencies (Brinkmann et al., 2009a, Brinkmann et al., 2009b).

4. Flexible and secure multi-layer encryption

MEF3 implements a flexible dual-tiered encryption schema allowing selective access to either biological data (neurophysiologic signals and video) or metadata (subject information which may contain PHI). A mechanism for time-obfuscation is also provided, allowing the user to remove information about when the recording was made while preserving information about the true time of day of the recording. Neither encryption nor time-obfuscation are required by the format; each recording's encryption scheme selected by the file creator. De-identification procedures to remove PHI, often required when sharing human data, can be completely avoided via this mechanism.

In MEF3, since the potentially subject identifying data is segregated from the biological data, one or both can be encrypted at various tiers, at the user's discretion. Encryption can be applied selectively to four different 'regions' of data: (1) technical recording data such as sampling frequency & filter settings; (2) data which contains potentially subject identifying information; (3) individual Data Blocks of the segment data files, and (4) individual records of the record data files. Encryption in MEF3 is performed using the 128-bit Advanced Encryption Standard (AES-128) algorithm which exceeds the 112-bit requirement of HIPAA for symmetric encryption of human data (AES Standard, NIST, 2001). A hash (type SHA-256) of the each password is stored in each MEF3 file for password validation.

The two encryption tiers used in MEF3 are referred to as 'Level 1' or 'Level 2'. Each of the regions described above can be encrypted with Level 1, Level 2, or no encryption (Level 0). For the reader of the data, Level 2 access guarantees Level 1 access, but not the converse. A typical encryption strategy would be to encrypt section 2 of the metadata files (technical recording data) with level 1 encryption, and to encrypt section 3 of the metadata files

(containing potentially subject-identifying PHI) with Level 2 encryption. Time series data blocks can also be encrypted efficiently (by encrypting only the statistical data model in the RED block header), but typically are not, as encryption of the technical metadata is generally adequate to prevent interpretation of the raw signal data. Users with Level 2 access, such as those directly involved in a patient's care, would have access to all details of the recording session including PHI. Data encrypted in this way could be shared, intact, with a research collaborator, who would otherwise need to be given a specifically de-identified version of the data, simply by providing him or her with only Level 1 access. This model of differential access reduces the burden of creating de-identified data sets and/or the storage of multiple versions of single data sets with and without PHI.

Recording times are not currently required to be obscured by HIPAA, but they are potentially subject identifying, and so their obscuration is increasingly being required by clinical institutional review boards (IRBs). Time in MEF3 is stored in micro Coordinated Universal Time (μ UTC) time. This is the time, in microseconds, since midnight on January 1, 1970 in Greenwich, England. This date and time, in Unix parlance is known as "The Epoch", and is supported by all Posix compliant Unices. Recording times in MEF3 are optionally obscured by offsetting them such that the true time of day of the recording is preserved, but if translated into a human readable date, appears as if the recording began on 1 January 1970, in Greenwich. All subsequent times in the recording are likewise offset. The recording time offset is stored in Section 3 of the metadata files, and therefore only accessible to a user with access to that section.

5. Time-synchronized customizable event and annotation records

MEF3 supports binary event and annotation records at all levels of the file hierarchy. The record types are user-extensible and the records are indexed, individually encryptable, and time-synchronized.

6. Robustness against file corruption and interruptions in data acquisition

Large data files acquired in a research setting may be kept and reanalyzed innumerable times. For these very large data files, backup can be delayed or at times impractical. Despite the incredible fidelity of digital storage media and network transmission protocols, data in these large files can become corrupted in small ways over time. Independence of the data blocks in MEF3 restricts the extent to which local damage can spread. In the time series channel files, data are blocked. Each block of the time series data files begins with a cyclically redundant checksum (CRC) so that data corruption can be readily identified and localized to a particular block, corruption is restricted to affected blocks only. Additionally, every file in the MEF3 hierarchy begins with a universal header, which itself begins with a file specific CRC. File damage can be detected simply by checking a file's CRC. Recovery from catastrophic damage is facilitated, to some extent, by the extensive alignment requirements of the MEF3 format, which are imposed largely for purposes of computational efficiency. There is some minimal but intentional redundancy in the format as well, which can be used for data recovery, but is also present for convenience and computational efficiency. Failures during data acquisition, such as power failure, or system crashes, leave a

fully intact session with the sole possible loss of terminal data blocks. Therefore, the MEF3 format easily accommodates both planned and unplanned interruptions in data acquisition.

7. Facilitate both online and offline review and analysis

Each data and event file is associated with a small index file facilitating rapid searches and random access. The format provides multiple technical data fields, facilitating programming for data analysis and review (which is critical to efficient handling of very large data files). Data files are acquired in their final format so online viewing or analysis tools need not be adapted from offline versions these tools.

8. Fully documented, open source, and with freely available development tools

At this time, there is fully-developed open-source and freely-available C library of source code supporting all the features of MEF3. The library includes all of the encryption, compression, and UTF-8 functions required by the format. It also contains integrated filtering functions, as this is a very common need in time series processing. The full specification and library are available at msel.mayo.edu.

9. Support for international adoption

All text fields and passwords in MEF3 support use of the 8-bit Universal Coded Character Set and Transformation format (UTF-8), providing international character support and allowing representation of text in any language. As discussed above, time is represented in μ UTC Time, which is the primary time standard by which the world regulates clocks and financial transactions and therefore provides an absolute global reference time frame.

10. Broad industry and academic community support

Since this is a problem that has not been discussed to a significant degree for at least the last decade, much work needs to be done to generate broad industry and academic community support for the adoption of a standard format. The process for format adoption should attempt to follow the successful path of DICOM. It should therefore begin with a partnership between academic societies such as the American Clinical Neurophysiology Society, the Society for Neuroscience, the American Board of Registration of Electroencephalographic and Evoked Potential Technologists (ABRET), and a society of electrical equipment manufacturers which promotes standards, such as the National Electrical Manufacturers Association (NEMA). Discussions and planning should involve representatives from OEMs from the very beginning, since their commitment to the project is critical. Leaders from other relevant physician organizations such as the American Epilepsy Society and the American Academy of Neurology also need to be involved in the process to help move things forward.

11. Format maintenance provided by an independent standards organization

Once a format has been agreed upon by leaders in academics and industry, the format should be submitted for formal designation as a standard by a standards developing organization (SDO). In the US, the main SDO is the American National Standards Institute (ANSI). Chosen by ANSI, the National Information Standards Organization (NISO) represents US

interests on the Technical Committee of the largest international SDO, the International Organization for Standardization (ISO). Although the format will be open source and freely available, OEMs may be asked to pay a small licensing fee to use the standard, proportional to their market share, in order to help pay for functions of the standards organization, such as committee meetings and web site maintenance. Any plan to adopt standards should involve a slow roll-out over a decade or more. Export and import tools using the new standard could be created by manufacturers in the first few years, followed eventually by adoption of the new standard as the native format for neurophysiology recording.

DISCUSSION

In 2003, the Institute of Medicine released a report “Patient Safety: Achieving a New Standard of Care” which described the importance of creating data standards in order to improve healthcare data interchange and improve patient safety (Aspden et al., 2004). As the number of medical devices increases and the medical record system becomes fully electronic, the emphasis on creating data standards will continue to grow. Establishing reliable methods for storing neurophysiology data, as an important area of neurophysiology research, has been under-emphasized. This area of research can be described as “neurophysiology informatics”. Unlike in the fields of electrocardiography and imaging, there are few scientific articles about informatics in neurophysiology. Increased attention needs to be given to the need for data standards in the neurophysiology field.

The adoption of standards in industry is often driven by economic forces. When railways were first built in the US, different railroad companies built train tracks with different track widths (called “gauges”), preventing railcars from riding on all tracks. The decision to use different gauges occurred due to differing engineering traditions. But growing demand for interregional rail traffic led to inefficiencies as cargos had to be exchanged between rail cars of various sizes. Eventually, companies realized that standardization would lead to greater productivity and considerable resources were spent on the standardization of track gauges throughout the North American continent (Puffert, 2000). The difference between this railroad example and clinical neurophysiology is that it is not the neurophysiology OEMs who are shouldered with the burden of “transferring the freight between railcars”. It is neurophysiology clinicians and researchers who have to do the work of translating formats and installing additional software programs to allow exchange of data. So the OEMs have little short-term financial incentive to help solve the problem. In fact, in the short term, the manufacturers have quite a bit of incentive not to change their recording format. The change would require considerable investment of engineering and programming resources and the transition period might cause instability in their systems, which could lead to mission-critical data loss and increased customer support calls. But in the long term, a common neurophysiology recording standard would benefit the OEMs in two important ways. First, a common format would make research collaboration easier, speeding the discovery of new applications of neurophysiologic recording technology to patient care. This could lead to greater demand for equipment and services. Second, new features of a modern standard format such as improved data compression and encryption of PHI could make data storage and transmission easier (requiring less space and less transmission bandwidth) and more compliant with future federal regulations. So not only would a common standard format

improve patient care and facilitate collaborative research endeavors, but it would provide long-term benefit to industry as well.

CONCLUSIONS

This review describes the importance of standardization of data formats in healthcare, past attempts at creating data standards in neurophysiology data storage, and proposes the use of the MEF3 format as standard format for neurophysiological data. A standard neurophysiology format would enhance development of new techniques for the use of neurophysiologic data in both the clinical and research arenas. The creation of a standard neurophysiology format is an ambitious but necessary project to move the field forward.

ACKNOWLEDGEMENTS

Special thanks to Ben Brinkmann and Dan Crepeau of Mayo Systems Electrophysiology Lab who have contributed significantly to the concepts and code base supporting MEF3 through its current and prior versions.

SUPPORT

NIH/NINDS U24 NS63930 The International Epilepsy Electrophysiology Database

NIH/NINDS R01 NS78136 Microscale EEG interictal dynamics & transition into seizure in human & animals

REFERENCES

- Aspden, P.; Corrigan, JM.; Wolcott, J.; Erickson, SM. Patient Safety: Achieving a New Standard of Care. National Academies Press; Washington DC: 2004.
- Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M. Large-scale Electrophysiology: Acquisition, Compression, Encryption, and Storage of Big Data. *Journal of Neuroscience Methods*. 2009a; 180:185–92. [PubMed: 19427545]
- Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M. Multiscale electrophysiology format: an open-source electrophysiology format using data compression, encryption, and cyclic redundancy check. *Conf Proc IEEE Eng Med Biol Soc*. 2009b:7083–6. [PubMed: 19963940]
- Chronaki CE, Chiarugi F, Lees PJ, et al. Open ECG: a European Project to Promote the SCP ECG Standard, A Further Step towards Interoperability in Electrocardiography. *Computers in Cardiology*. 2002; 29:285–8.
- Halford JJ, Pressly WB, Benbadis SR, et al. Web-based Collection of Expert Opinion on Routine Scalp EEG: Software Development and Inter-rater Reliability. *Journal of Clinical Neurophysiology*. 2011; 28:174–8.
- Hellmann G, Kuhn M, Prosch M, Spreng M. Extensible biosignal (EBS) file format: simple method for EEG data exchange. *Electroencephalography & Clinical Neurophysiology*. 1996; 99:426–31. [PubMed: 9020801]
- Herman, ST. Data Exchange Standards for Clinical Neurophysiology. Online Powerpoint Presentation: <http://www.physionet.org/standards/npsg/Herman.pdf>
- Jacobs EC, Lagerlund TD, Collura TF, Burgess RC. Data Interchange for Clinical Neurophysiology. *Studies in Health Technology and Informatics*. 1993; 6:195–202. [PubMed: 10163815]
- Kemp B, Olivan J. European data format ‘plus’ (EDF+), an EDF alike standard format for the exchange of physiological data. *Clinical Neurophysiology*. 2003; 114:1755–61. [PubMed: 12948806]
- Kush R, Goldman M. Fostering Resonsible Data Sharing through Standards. *New England Journal of Medicine*. 2014; 370:2163–5. [PubMed: 24897080]
- Pianykh, OS. *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide*. Springer-Verlag; Berlin: 2008.

- Puffert DJ. The Standardization of Track Gauges on North American Railways, 1830-1890. *The Journal of Economic History*. 2000; 60:933–60.
- Rath D. Trend: privacy: the days of relaxed privacy provisions under HIPAA are coming to a close, as HITECH brings with it changes that will rock unprepared hospitals. *Healthcare Informatics*. 2010; 27:20–3.
- Schloegl A, Chiarugi F, Cervesato E, Apostolopoulos E, Chronaki CE. Two-Way Converter between the HL7 aECG and SCP-ECG Data Formats Using BioSig. *Computers in Cardiology*. 2007; 34:253–6.
- Trigo JD, Alesanco A, Martinez I, Garcia J. A review on digital ECG formats and the relationships between them. *IEEE Transactions on Information Technology in Biomedicine*. 2012; 16:432–44. [PubMed: 22128009]
- Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The Value Of Health Care Information Exchange And Interoperability. *Health Affairs*. 2005; W5:10–8.
- Willems JL, Campbell G, Bailey JJ. Progress on the CSE diagnostic study. Application of McNemar's test revisited. *Journal of Electrocardiology*. 1989; 22(Suppl):135–40. [PubMed: 2533232]
- Zywietz C. SCP-ECG and Vital Signs Information Representation - two examples of successful transcontinental cooperation in medical informatics standardization. *International Journal of Medical Informatics*. 1998; 48:195–9. [PubMed: 9600420]

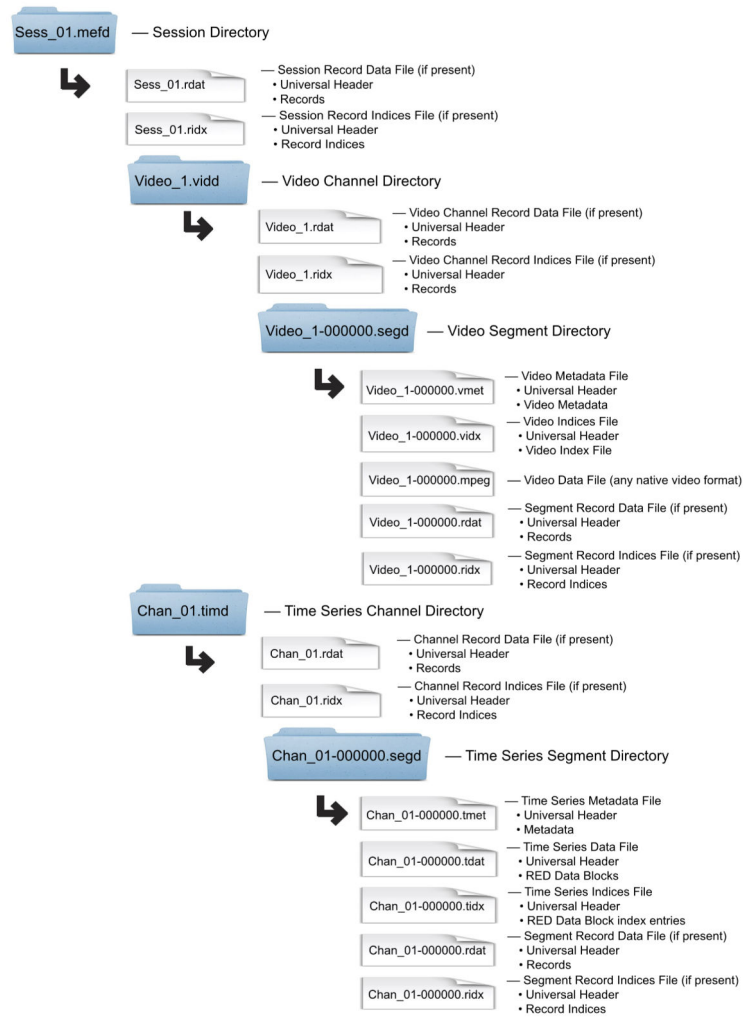


Figure 1.
MEF3 File Hierarchy

RED Compression

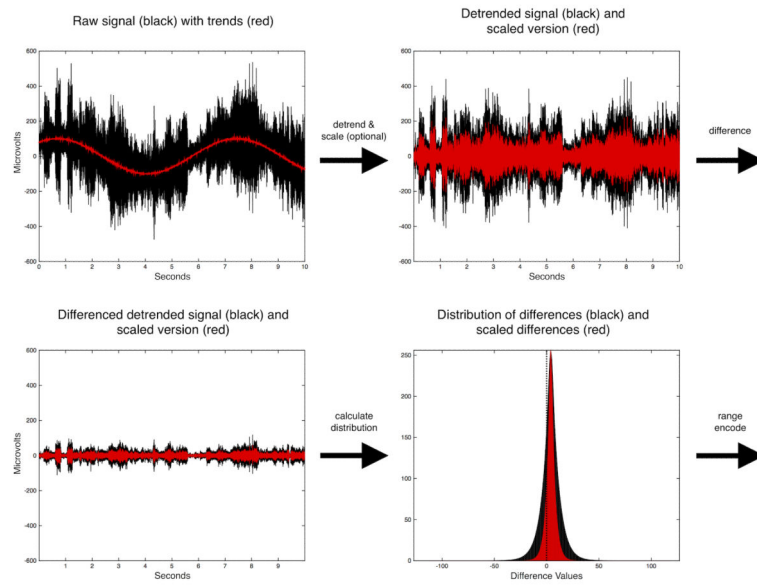


Figure 2.
Steps in Compression Using RED Algorithm

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

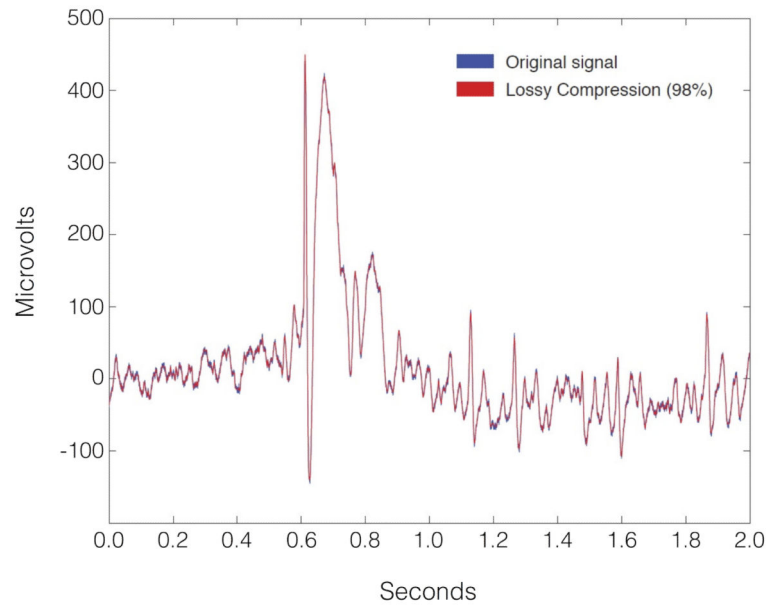


Figure 3.
Lossy Compression of an EEG Segment Using the RED Algorithm

TABLE 1

Required features/processes for a standard neurophysiology format

| |
|---|
| 1. Hierarchical file structure with flexible segmentation |
| 2. Extensible and flexible support for big data applications |
| 3. Efficient lossless or lossy data compression |
| 4. Flexible and secure multi-layer encryption |
| 5. Time-synchronized customizable event and annotation records |
| 6. Robust against file corruption and interruptions in data acquisition |
| 7. Facilitates both online and offline review and analysis |
| 8. Fully documented, open source, and with freely available development tools |
| 9. Support for international adoption |
| 10. Broad industry and academic community support |
| 11. Format maintenance provided by an independent standards organization |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript