Standards for biomedical signal databases

Article *in* IEEE Engineering in Medicine and Biology Magazine · May 2001

DOI: 10.1109/51.932722 · Source: PubMed

4 authors, including:

Alpo Värri
Tampere University of Technology
98 PUBLICATIONS **1,682** CITATIONS

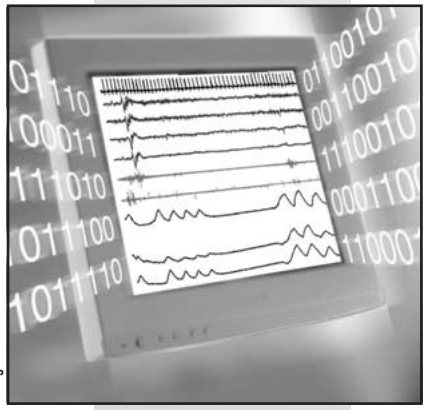
SEE PROFILE

Bob Kemp
95 PUBLICATIONS **2,466** CITATIONS

SEE PROFILE

Thomas Penzel
Charité Universitätsmedizin Berlin
671 PUBLICATIONS **9,781** CITATIONS

SEE PROFILE



©Digital Vision

Standards for Biomedical Signal Databases

Factors to Consider when Choosing the Proper Data Format for Biosignal Exchange and Archiving

Alpo Värri[1], Bob Kemp[2], Thomas Penzel[3], Alois Schlögl[4]
[1]Tampere University of Technology, Tampere
[2]Sleep Center, MCH - Westeinde Hospital, Den Haag
[3]Policlinic of Philipps University, Marburg
[4]University of Technology, Graz

Biomedical signal databases are used in engineering for the testing and design of software, in scientific research, in drug trials, and in healthcare. Standardization of databases strongly facilitates international and multicenter collaboration and multiple use of the same data. This, in general, improves the quality of the developed software, scientific publications, outcome of drug trials, and healthcare. It also reduces costs and in some cases leads to results that could not have been obtained without such collaboration.

Telemedicine requires the use of standards in information interchange when the equipment of several manufacturers needs to work together. Initial trials can be done with devices of a single manufacturer [1], but the foreseeable market of third-party biosignal analysis services requires adequate standards supported by many manufacturers. In medical imaging, particularly in radiology, the DICOM/MEDICOM standard accepted both by American and European manufacturers has already solved the interoperability problem, but in the biomedical signal domain various alternatives still exist.

This article presents some requirements imposed on data format specifications derived from different biosignal recording environments. This is followed by a review of four particularly interesting data formats and some notes about other specifications. Finally, the merits of different formats are discussed and some views about future developments are given.

Requirements for Biosignal Formats

Biosignal databases are collected with different goals and requisites. The applicability of a format specification needs to be judged with respect to these requisites but also to the intended or potential user community of the biosignal database.

One measure to judge the applicability of a data format to a specific purpose is to examine what data types it supports. Due to various reasons beginning from the imprecise placement of the measurement sensor or electrode, the accuracy of a biomedical signal measurement is often not higher than three decimals. Therefore, in most cases the dynamic range of such a measurement can be covered by 16-bit integer values when the amplification of the signal has been arranged properly. In some measurements, 8 bits are enough, and if the format supports 8-bit integers, 50 percent of the storage space is saved. As some analog-to-digital converters used in EEG measurements have 22-bit resolution, the manufacturer would perhaps like to see support for a 32-bit integer, too.

Signals are hardly ever sampled as floating-point numbers, and the lack of support for this data type is not a serious flaw to a storage format. Sometimes there may, however, arise a need to store intermediate analysis results (e.g., autoregressive model parameters of an EEG signal segment) along with the original signals, and in that case the use of the floating-point data type is necessary.

Especially in long-term measurements, the signals form the bulk of the data. When the signals have been analyzed either by a human or a computer, the need to store the results of the analysis emerges. Human analyzers mostly produce their results in the form of annotations. The annotations can be in textual form or in the form of diagnostic or other codes defined in a local or more widely used coding scheme. The annotations may be linked to certain time instances in the recording or they may form a summary annotation (e.g., a diagnostic statement)

 0739-5175/01/$10.00©2001IEEE

of the whole recording. In the first case the time stamp of the annotation needs to be stored, as well. Other types of time-stamped observations are various events that are related either to the signals (e.g., computer identifies the beginning of an artifact) or external conditions (e.g., the lights are switched off in the sleep laboratory).

Alarms are a special type of event that sometimes occurs in intensive care unit and anaesthesia recordings. Storing them is sometimes important also for legal reasons when an archival copy of the anaesthesia recording is made.

It is naturally possible to store the analysis results and events into separate files, but storing them together with the signals into the same file offers the advantage of keeping all data of the recording in one place. This solution requires less file management and less prudence from the operators. This is advantageous especially when the data is sent to another institution for whatever purpose.

Another type of vital sign that may need to be stored are numeric measurements taking place during the biosignal recording. Characteristic to these measurements is that they are made at irregular intervals and they cannot therefore be stored as a separate biosignal channel. These measurements are usually linked to time, and therefore a time stamp is needed to accompany them, as well.

Digital video is becoming more and more commonplace in the intensive monitoring laboratories of epilepsy. A mechanism to link the video and the signals together in time is necessary to analyze the origin of the symptoms seen on the video. Voice annotations are not that hard to imagine to accompany biosignal recordings in some circumstances. Sometimes a simple image of, for example, the electrode positions on the head in a special EEG recording may come in useful. It is easy to foresee that the biosignal format could also provide some mechanisms to either embed these multimedia data into the biosignal file or to have links to point to those data in other files.

Apart from the supported data types there are also other factors that are decisive with respect to the suitability of the format to a particular application. In some applications it is not economical to use the same sampling frequency for all channels because their rate of change is different. The format should therefore support several sampling frequencies in the same file.

The size of the resulting file is of rather little importance in short recordings, such as an evoked potentials study in which only a second or two of the averaged waveform is stored for 20 channels. In long-term recordings it is very useful that the data are stored efficiently because, for example, in a sleep recording of 16 channels, many of them sampled at 200 Hz produces often around 100 MB of data per night. Compression of such data is an obvious solution to suggest, and it has been used successfully in normal ECG measurements (the SCP-ECG prestandard *CEN/ENV 1064*) [2]. In many other biosignal measurements, compression is not equally beneficial if the whole dynamic range of the analog-to-digital converter has been utilized almost uniformly and the signal does not contain periodicities or other predictable elements. If the signal is not uniformly distributed, entropy coding can provide gains in storage space. The benefits of compression come also with a price: it requires time both to compress the data and to decompress it for viewing and analysis. Compression adds complexity to the format specification as well. This can be prohibitive for researchers who are not specialists in programming but who would like to test their own methods on the biosignal data. For this group of people, the simplicity of the format specification is also an important parameter, if supporting tools for compression and decompression are not available.

If the format specification is suited for online recording of biosignal data (i.e., it supports streaming) and additionally the data can be presented on a display from this format in a rapid fashion, a manufacturer might decide to use it as a native recording format for his devices. This property would therefore be very useful in order for the specification to achieve acceptance among manufacturers.

Finally, the popularity of the storage format is a factor not to be forgotten. The user of a perfect format might be happy with it, but if no one else is using that format, the user has not reached the goal of using a standard format—the easy exchange of biosignal information.

Some Format Specifications

Ideally, the user-friendly data format should be simple to understand and implement and still be able to store all the information in an efficient and unambiguous way. We will now give an impression about the required effort and applicability of a few formats. These formats have been chosen with the last criterion of the previous chapter—the support of more than one institution in more than one country.

CEN/ENV 1064 (SCP-ECG) Specification

One standard data format has been established as European prestandard *CEN ENV 1064*, and it is widely available for ECG recordings. This prestandard is also known as the SCP-ECG standard [2]. This standard can handle binary signals and annotations in a number of defined sections as they are obtained in different tests with ECG recordings. These include short-term ECG recordings, long-term ECG recordings, stress tests with ECG, and also angiography with ECG. In addition it can also handle data compression using known algorithms. At this instance the handling of images is also mentioned in the SCP-ECG specification. Data compression is possible within the format, which means that compressed and uncompressed data sections can reside in the same file. The signal characteristics of the data format in terms of filter specifications, sampling rates, and annotation codes are limited to ECG recording and thus this format cannot be used for other purposes. SCP-ECG is supported by many major manufacturers of ECG equipment. Therefore, it is a recommendable alternative for ECG databases.

ASTM 1467 Specification

The only standard for neurophysiology thus far that is supported by an official standardization body is the *ASTM E-1467-94* [3]. The specification has been prepared in cooperation with HL-7 and actually defines a message using the HL-7 methodology [4, 5]. It offers support for practically all neurophysiological measurements such as electroencephalograms (EEGs), polysomnograms (PSGs), evoked potentials (EPs), electromyograms (EMGs), etc., but it is not limited to neurophysiology. It is also suitable for ECG and vascular/intracranial pressure monitoring or gastrointestinal motility studies.

The specification has provisions for the storing of both primary and derived data. In addition to digitized biosignal waveforms of multiple channels, the primary data includes channel identifications, sensitivities, filter settings and sampling frequencies, averaging parameters (when necessary), time stamps, elec-

trode or transducer locations and attributes, measured distances, stimulation parameters, calibration data, annotations, medications administered, special procedures performed, and instruments used. Optional derived data can consist of measured feature or peak latencies, amplitudes and other derived measures entered by a technician or calculated by the device, spectral analyses, and other quantitative and qualitative results.

The implementations are facilitated by defining different levels of implementation depending on the needs of the system. Level 1 is the most basic level of implementation, and it contains only the biosignal waveforms and their channel labeling. A level 1 receiver implementation does not need to accept level 2 or 3 data, but it shall not generate an error either if it receives this kind of data. In level 2 both waveforms and procedure annotations can be conveyed. In level 3, standard alphanumeric codes are defined for several of the textual data elements such as diagnoses.

The *ASTM-1467* messages consist of a sequential series of segments, with each segment conveying one aspect of a message. There are header segments, segments that contain the order of an investigation, and segments of the result. Segments consist of fields that define one attribute of the segment. A field may also be a hierarchical structure containing subfields.

ASTM-1467 defines a number of data types which, however, are special types of text strings as all segments and fields are represented in ASCII characters. Special delimiter characters are used to separate the fields and segments from each other. The use of ASCII is a stumbling block to many implementers as the ASCII encoding of 16-bit waveform values may require 7 bytes (including the delimiter), which can triple the required storage space. Another difficulty with ASCII encoding of samples is the long search time of a particular point of time in the recording as the sample of interest cannot be located with a quick calculation but requires a read through all the preceding samples. The specification mentions that binary encoding can be used in special cases but does not lay out a single interoperable format of doing so.

Despite the extensive coverage of various biosignal measurement parameters (or perhaps due to it), this specification, which was finalized in 1994, has hardly been implemented and has been mainly used for records of procedures and accounting issues. The Mayo Foundation (Rochester, Minnesota, USA) has provided software support to the *ASTM-1467* user group including format conversions.

EDF Specification

EDF (European Data Format) is a simple format for exchange and archiving of multilingual (not necessarily biological) time series. The signals can have any (and different) physical dimensions and sampling frequencies. EDF was developed in 1991 and published by European engineers who had their occupation in medical environments [6]. The work was coordinated by the task group on Signal Analysis, which operated within the European Concerted Action on "Methodology for the Analysis of the Sleep-Wakefulness Continuum." The publication contains the full specification of EDF, which is only one page.

An EDF file has an ASCII header containing mainly patient and time identification, the number of signals, and the technical characteristics (mainly dimension, calibration values, sampling frequency) of each signal. The header is followed by subsequent data records, each of the same duration, that contain the recorded signals in 2-byte integer values. The duration of the data records is specified in the ASCII header.

EDF can also accommodate annotations, markers, and events. A simple quarter-page scheme for this has been described in section 2.2 in [7]. Another half-page scheme describes how EDF can be used for floating-point time series, be it with limited accuracy [8].

Studying the specification and then implementing an EDF import/export unit typically takes a few days of work. Many research groups and companies, particularly in Europe, Australia, and the USA, have implemented it. It has been, and is being, applied in neurophysiological equipment and software, scientific research projects including multicenter drug trials, the development of signal analysis software (the European IMPROVE, IBIS and SIESTA projects), and in educating and consulting medical specialists (the European ENN project). All information about EDF including its full specification, a list of EDF-supporting companies, many links to downloads of software and data, the floating-point scheme, and a FAQ-list including the annotations scheme is available on the Internet (www.hsr.nl/edf).

CEN-TC251/FEF Specification

The design of the CEN/TC251 biosignal format specification [file exchange format (FEF) for vital signs] [9] is based on the CEN/TC251/WGIV model and extensive nomenclature of biomedical measurements (vital signs information representation [10]). The model was designed to incorporate data items found in intensive care units, anaesthesia departments, and clinical laboratories including neurology. Therefore, it supports all the requirements stated above. The side-effect of this is naturally increased complexity of implementation with respect to the EDF specification.

A biosignal file in FEF format consists of sections. Each section has a tag in the beginning for the receiving application to identify each section. The tag is followed by a length field that indicates the length of the section. If the receiving application does not need the information the section contains, it can skip the section by reading the file forward the amount of bytes the length field indicates. The actual data follows the length field. Some sections can contain subsections that also follow the tag-length data structure.

The demographics section contains information about the recorded patient. A healthcare provider section is provided to store basic textual data of the healthcare institution and personnel that collected the data. The medical device system presentation section contains a structured description of the devices (one or many) that participated in the collection of the biosignals and other data in the file. With the appropriate use of links ("handles" in the specification) it is possible to identify which device generated any particular piece of information. An optional manufacturer-specific section is provided for such data for which there is not yet a structured presentation in the format specification. A manufacturer may choose to save custom device setting information into this section. Additionally, there are also other places where it is possible to store manufacturer-specific information in order to attract manufacturers to use the format as the native format for their instruments. An optional multimedia section has been defined for those purposes in which the biosignal file needs to embed or provide a link to audio, image, or video data. The acquired biosignal, event,

time-stamped measurement, and alarm data are stored into a structure called the session archive. It may consist of one or more session tests. Session tests are further divided into one or more session phases during which the channel configuration, signal amplification, etc., each remains the same. Each session phase contains descriptive data about the measurement channels (sampling frequencies, sensor types and locations, amplifications, etc.) and the measured data in digital form.

A long list of 16-bit codes exists to encode the measured biosignal, body site locations, events, and units of measurements. These code tables are a copy of those in the European prestandard *ENV 13734*, *vital signs information representation*, which was prepared in cooperation of CEN/TC251 and the IEEE 1073 committee on the medical information bus. This code table is also on its way to becoming an ISO standard through the ISO committee for health informatics (ISO/TC215). This effort, as well as the FEF definition process, is a serious attempt to unify biosignal and related measurement offline storage needs for both the various electrophysiological laboratories and intensive care/anaesthesia departments. Free-of-charge support software is planned to promote the popularity and ease the implementation of FEF.

Alternative Specifications

Generally, if one has the possibility to choose the format specification to be used in a data interchange project, the use of one of the specifications above is recommended. If none of the above seems to fit for the purpose adequately, and the price of not being interoperable is not too high, there are also other available specifications to study. If one of the less popular alternatives is suitable for the purpose, one usually has the advantage of some existing software to support the format to begin with. Sometimes the support for the format has ceased to exist due to the fact that the authors of the format have left the field and the user community has been too small to carry out support.

Although the purpose of the DICOM committee is to standardize image communication in medicine, one of its side products is a definition of an exchange format for biosignals. The DICOM Supplement 30, Waveform Interchange, was created due to popular demand (from the ECG community) for purposes where biosignals are collected in connection with a medical imaging procedure [11]. The implementation of this format requires understanding of the DICOM philosophy, which is not possible by reading *Supplement 30* alone.

The Extensible Biosignal Format (EBS) was designed by the epilepsy research group of the University of Erlangen [12]. The format was designed to support EEG and MEG signals. Due to the tag-length data architecture, additional data items such as image or audio data could be included as well. Due to the EEG background (high-pass filters are usually used in signal recording) a serious flaw took place in the design of the format by the omission of a dc-offset field in the description of the biosignals and thus the calibrated recording of dc types of signals such as blood pressure is not possible. Free software was also provided through the WWW for potential users of the format, but the most active support for the format has ceased to exist.

The SIGIF format was designed with neurophysiological biosignal exchanges in mind at the University of Aveiro, Portugal [13,14]. A SIGIF file contains a header describing the measurement set-up followed by the biosignal data sampled at an uniform sampling rate. The header is extensible and designed in such a way that the following versions of the format do not break old software trying to access information using the access method of the older version of the specification. The data can be stored as single- or double-length integers, as floating point values, or as an ASCII representation of any of these. SIGIF also supports the compression of biosignals with a method defined by the authors. A need for a coding scheme for biosignals was recognized and a small one was created. Free software was provided to help implementers of the format.

In addition to the above, other format specifications have been published as well (MIT arrhythmia database format, IFFPHYS from Australia, etc.) and it is difficult to draw a line to describing them here. Some user groups of a particular brand of data acquisition system have standardized the format specified by the manufacturer, but a format of any manufacturer is not general enough for universal use since few manufacturers have even intended that to happen. Some of them even prefer to keep their formats secret in order to be able to protect their intellectual property or in order to be able to make changes without protests from their customers.

Discussion

In most cases, the biomedical signals are sampled by a single device and their dynamic range can be covered by 16-bit (or less) integer values. In those cases, the EDF (Kemp et al. 1992) has proven its usefulness. The specification of this format takes only one page, and it is simple and many companies and researchers are already exchanging data using EDF. The format specification does not give defined signal labels, nor sampling rates nor other restrictions. Thus, the format can be used for many signals using free signal definitions without violation. On the other hand, when no signal labels have been defined, different users can use different labels for the same signal thus making it difficult for analysis programs to choose the channels correctly without operator intervention. This is avoided in the more complex *ASTM 1467-94* and CEN/FEF formats. The general issue in designing data formats is the balance between free handling of the data format and giving as few definitions as possible. EDF is a good example) and on the other extreme define codes that have to be filled in each single place in the file (ASTM and SCP-ECG being examples). The first extreme is the ideal format for developers because their acquisition software will almost every time work right due to the few restrictions. As labor costs form the majority of costs in health care, the trend is toward fully interoperable automatic systems (with extensive coding schemes) that do not require human intervention.

Conclusions

A general-purpose, effective, user-friendly, and universally accepted data format is not available at the moment. The user must make a trade-off between simple implementation and scope of data types the format covers. The EDF specification takes one or two pages, the CEN/FEF specification takes approximately 100 pages, and the ASTM specification takes 144 pages (see the ASTM website). If the simple format is not adequate, support software for the more complex specifications may ease the pain of their adoption.

Although a unified file format for biosignals is a long-term goal of many system integrators, the choice of a format still depends on the field of application.

SCP-ECG is relatively well established in the routine ECG recordings. EDF has proven to be applicable in research and routine work in neurophysiological and sleep laboratories. ASTM offers many codes for all different kinds of neurophysiological signal recording but has only reached acceptance in the United States because it is so close to the American reimbursement system. The CEN/FEF format can be applied in more situations at the cost of greater complexity, but the future will show its popularity as implementations are still rare.

Although several specifications are available, the designers of innovative recording protocols may still find them inadequate in a particular need related to their new type of measurement setting. In this case, yet another specification will be designed. The designers could, however, learn from the above formats and adopt their best characteristics, thus avoiding reinventing the wheel completely. As viable alternatives already exist, any such new definition has, however, a difficult time becoming a standard or even a quasi-standard.

Alpo Värri received the M.Sc. in electrical engineering in 1986 and the Dr.Tech. degree in signal processing in 1992, both from the Tampere University of Technology, Finland. Currently, he is a senior researcher and the vice head of the Signal Processing Laboratory of Tampere University of Technology. Since 1994 he has participated in health informatics standardization within CEN/TC251. His research interests include biomedical signal processing, particularly in sleep research and computer programming.

Bastiaan (Bob) Kemp received the M.Sc. in electrical engineering and a Ph.D. on model-based monitoring of sleep stages from Twente University of Technology, Enschede, The Netherlands, in 1977 and 1987, respectively. He is a medical physicist in the Department of Neurology at Leiden University Medical Centre, The Netherlands. He is also a clinical physicist and director of the Center for Sleep and Wake Disorders in MCH-Westeinde Hospital in Den Haag, The Netherlands. His interests are in sensors, models, and algorithms for the study of sleep and movement disorders.

Thomas Penzel studied physics and mathematics in Göttingen, Berlin, and Marburg. He received his diploma in theoretical physics in 1986. He then studied human biology and obtained a doctorate thesis in 1991. In 1995 he completed his habilitation in physiology at the University of Marburg. He also studied medical informatics and completed this in 1997 with a certificate from the German Society on Medical Informatics (GMDS). Since 1982 he has worked at the sleep laboratory of the University in Marburg. Since 1987 he has been leading the task group "Methodology" of the German Sleep Society (DGSM). Since 1992 has been the chair of the "commission for the accreditation of sleep laboratories" in Germany. In 1993 he was elected as an extended board member of the German Sleep Society. Since 1997, he has been a member of the board of the International Society on Biotelemetry (ISOB). He has written more than 80 national and international papers and more than 50 book chapters.

Alois Schlögl received his Dr.Tech. degree for his Ph.D. thesis on "the electroencephalogram and the adaptive autoregressive model" in 2000 from the University of Technology Graz, Austria. During his Ph.D. study, he was a research assistant at the Institute of Biomedical Engineering. He was working for a project on an EEG-based Brain Computer Interface and for the SIESTA project on multicenter sleep analysis.

Address for Correspondence: Alpo Värri, Signal Processing Laboratory, Tampere University of Technology, P.O. Box 553, FIN-33101, Tampere, Finland. E-mail: varri@cs.tut.fi.

References

[1] P. Loula, E. Rauhala, M. Erkinjuntti, E. Räty, K. Hirvonen, and V. Häkkinen, "Distributed clinical neurophysiology," *J. Telemed. Telecare*, vol. 3, no. 2, pp. 89-95, 1997.

[2] *ENV 1064 Standard communications protocol for computer-assisted electrocardiography*. European Committee for Standardisation (CEN), Brussels, Belgium, 1996.

[3] *Standard specification for Transferring Digital Neurophysiological Data Between Independent Computer Systems*, ASTM E1467-94, American Society for Testing and Materials, ASTM, (www.astm.org), 1916 Race St., Philadelphia, PA 19103, USA, 1994.

[4] E.C. Jacobs, T.D. Lagerlund, T.F. Collura, and R.C. Burgess, "A data interchange standard for clinical neurophysiology," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1993, pp. 813-817.

[5] E.C. Jacobs, T.D. Lagerlund, T.F. Collura, and R.C. Burgess, "Data interchange for clinical neurophysiology," *Stud. Health Technol. Inform.*, vol. 6, pp. 195-202, 1993.

[6] B. Kemp, A. Värri, A.C. Rosa, K.D. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings," *Electroencephalogr. Clin. Neurophysiol.*, vol. 82, pp. 391-393, 1992.

[7] M. van de Velde, M.M.C. van den Berg-Lenssen, G.J.M. van Boxtel, P.J.M. Cluitmans, B. Kemp, J. Gade, C.E. Thomsen, and A. Värri, "Digital archival and exchange of events in a simple format for polygraphic recordings with application in event related potential studies," *Electroencephalogr. Clin. Neurophysiol.,* vol. 106, no. 6, pp. 547-551, 1998.

[8] B. Kemp, T. Penzel, A. Värri, P. Sykacek, S.J. Roberts, and K.D. Nielsen, "EDF: a simple format for graphical analysis results from polygraphic Siesta recordings," *J. Sleep Res.*, vol. 7 (suppl. 2), p. 132, 1998.

[9] "File Exchange Format for Vital Signs, Interim Report, Revision 2," TC251 Secretariat, Stockholm, Sweden, CEN/TC251/PT-40, 2000.

[10] "Vital Signs Information Representation," European Committee for Standardisation (CEN), Brussels, Belgium, ENV 13734, 2000.

[11] *DICOM Suppl. 30, Waveform interchange*, Nat. Elect. Manufacturers Assoc.: ARC-NEMA, Digital Imaging and Communications, NEMA, Washington D.C., 1999.

[12] G. Hellman, M.G. Kuhn, M. Prosch, M. Spreng, "Extensible biosignal (EBS) file format: simple method for EEG data exchange," *Electroencephalogr. Clin. Neurophysiol.*, vol. 99, no. 5, pp. 426-431, Nov. 1996.

[13] M.B. Cunha, J.P. Cunha, and T. Oliveira e Silva, "SIGIF: A Digital Signal Interchange Format with Application in Neurophysiology," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 5, pp. 413-418, May 1997.

[14] M.B. Cunha, J.P. Cunha, and T. Oliveira e Silva, "SIGIF: A Digital Signal Interchange Format for Biological Signals," in *Proc. 15th Annu. Int. Conf. IEEE Engineering and Biology Society*, 1993, Part 2, pp. 644-645.