An overview on data formats for biomedical signals.

A. Schlögl[1]

[1] Institute for Human Computer Interfaces, University of Technology, Graz, Austria

[2] Institution/Department, Affiliation, City, Country

***Abstract*— Biomedical signals are stored in a variety of different data formats. This obstructs true interoperability, increases costs for software development and maintenance of software. Therefore, it is would be desirable to have a single data format for biomedical signals.**

In order to address this issue, (I) essential properties of biosignal data formats are identified, and (ii) an overview of various data formats is provided. The advantages and disadvantages of 19 data formats are compared in detail.

***Keywords*— data formats, biomedical signals, standardization**

INTRODUCTION

There is not a lack of standards for biomedical signals but to many different standards. Almost every equipment vendor is storing the data in a different format. There are a few international standards like EN1064 and ISO 91064 (SCP-ECG) [16], and MFER ISO 92001 (MFER) [15], CEN 14271 (FEF) [20]. There are some more or less widely accepted quasi-standards like EDF/EDF+ [6,7] and HL7 v3.0 annotated ECG [14]. Moreover, several equipment provider made the documentation of there data format public available. This is often the case for vendors with a large market share in there respective area. Many companies provide the specification of the data format to their costumers upon request. Several vendors do not provide the specification of the data format, either because the recording equipment is not supported anymore, or by the policy to keep the data format closed.

Within the Free and Open Source Software library for biomedical signal processing (BioSig) [1], about 50 different data formats are supported. The large number of different data formats make interoperability very difficult. Data from a some recording equipment can often be analyzed only by the corresponding software. This can result in a so-called "vendor-lockin", making the customer dependent on a single provider. In the following, an overview of a number of different data formats, as well as its advantages and disadvantages are provided.

A complete analysis of all data format is not only prevented by the large number of formats and unavailable specifications. In some cases, the legal situation is not clear whether the specification can be discussed in public without an additional legal conformation. Fortunately, this affects only minor and less important data formats. Nevertheless, for these reasons, the following analysis of different data formats will include only data formats, where the specification is available in public. A list of public information on biomedical signalb is available at http://en.wikipedia.org/wiki/List_of_file_formats#Biomedical_Signals_.28Time_Series.29.

METHOD

In order to evaluate the different formats, the needs and requirements of biomedical signal processing are compared with the provided features of the formats.

Multiple sampling rates and scaling factors

Polysomnograms (PSG) contain different types of biomedical signals including EEG, ECG, EMG, EOG, oxygen saturation, respiration, etc. In order to store PSG efficiently, support for different sampling rates and scaling factors is required.

Support of automated overflow/saturation detection

Biomedical signals are often contaminated by artifacts. For example, mechanical movement of electrodes or sweating can cause saturation in EEG recordings. It is important to identify such overflow artifacts. If the information about the dynamic range is available, one could easily detect these artifacts automatically. Unfortunately, this information is often not available [23].

Multiple data types, dynamic range

Recording devices use a variety of analog-to-digital (A/D) converter types. There is a general trend to increased sampling depth. In earlier days, 8 bits were very common, over the last ten years a shift from 12 bit to 16 bits can be observed, and today devices with more the 16 bits are already available.

Physical units

It is crucial to know the physical unit of a recorded signal, i.e. whether the sample values represent millivolts (mV) or microvolts (uV). ISO11073:10101 defines about 190 different physical units, each can be combined with 21 prefixes [24]. Most data formats use a fixed unit (e.g. mV or uV), MFER [15] supports 23 predefined units, and several formats use an ASCII-string of fixed length. E.g. EDF/EDF+ [6,7] provide 8 byte ASCII for this field; however, about 20 to 30 % of the physical units listed in [24] require more than 8 bytes.

O. Dössel and W.C. Schlegel (Eds.): WC 2009, IFMBE Proceedings 25/IV, pp. 1557–1560, 2009.
www.springerlink.com

Events, Annotations and Markers

are stored in a variety of different approaches. Markers are stored in a status channel (BCI2000 [3,4], BDF [25]) or in annotation channels (EDF+ [7]), or in event tables (CNT-Neuroscan), or trigger files (e.g. BrainVision). The information is highly dependent on the experiment and the encoding used by the researcher or investigator. If data are exchanged, additional specifications must be added, otherwise one can not identify the meaning of this marker information. Free text annotations have the disadvantage that the same type of event can be described in different ways, or some description has ambiguous meanings. Therefore, a unique and unambiguous encoding of events and markers is desirable.

Archives and databases, demographic information

For the analysis of large databases and archives, it is important to have information about patient demographics, recording equipment, researcher/investigator etc. available. Most data formats do not provide means to store this information. Retrieving this information from other sources (if available) is often prohibiting important meta-analysis of biosignal data.

Random data access and streaming

Most data formats store the data in binary format. The main reasons for this are: (i) it is memory efficient, (ii) random data access is available (iii) a lean software environment is sufficient. However, several data formats store the data in ASCII-encoded text format (like ASTM-E1467 [12], Hitachi-ETG4000, Vital-FEF, ...). More recently XML-based data formats [13,14, 21, 22] have been proposed. Data compression is often suggested to overcome the limitation of Text and XML formats. However, random data access is not supported, and data streaming (pipelining blocks of data) is not possible if the samples from each channel are stored together. This causes "*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*" according to [5] discussing ASTM-E1467. Therefore, random data access is an important aspect for biosignal data formats.

Electrode positions

*S*tandardized schemes for the electrode positions have been described for EEG and ECG. In these cases, the electrode label is sufficient to obtain the absolute position. However, for other biomedical signals or if no standardized montage is used, coordinates of electrode positions must be also stored. Moreover, some recording techniques may require additional information about recording sensors, such as sensor orientation in the case of the magnetoencephalogram. Rarely any format supports storing electrode or sensor positions.

Single file

Some data formats store the information in different files (e.g. separate header file, marker file and data file). This causes problems when the filenames are changed or some files are moved into a different directory, or if one of the files is accidentally replaced. This can lead to an inconsistent or even invalid database, and there is no mechanism for detecting such a state. These problems can be avoided if all related information is stored in a single file.

Vendor formats

Many vendors already provide the specification for their data formats. However, these data formats are often tightly locked to their recording equipment, and thus the format might not be suitable for recordings from other recording equipment. And even if the technical properties of the format would be suitable, a competing company might choose not to use the format because the specification of the format could be changed by the competitor at any time. Possible legal issues (e.g. copyright) may also prevent the wide-spread acceptance of vendor-specific formats. The vendor-specific formats are not included in the evaluation because there is no guarantee that the specification does not change and will be available in the future.

RESULTS

Table 1 shows the extent to which data formats fulfill the requirements discussed in the previous section, excluding vendor-specific formats. Most data formats have a score of six or lower, indicating that some important features are missing. Only GDF gets higher scores, and only the latest version (GDF v2.1) gets the highest possible score of 11, because it addresses all issues listed above. GDF supports multiple sampling rates and multiple scaling factors, it supports different data types ranging from int8 to float128, it fosters automated saturation detection, the physical units are encoded according to [24], events are stored in a dedicated event table, predefined codes for a number of events are used and up to 255 user-specified events can be defined. To facilitate dissemination, we implemented support for the GDF format in M-code (BioSig for Octave and Matlab), in C/C++ (BioSig for C/C++), provide an interface to Python, and support GDF in the SigViewer data viewing and scoring utility. All these tools are available from BioSig, the free and open source software library for biomedical signal processing [1,2]. The BioSig project provides bidirectional converters (save2gdf) between different file formats [26], and it can be used to test the compatibility between the formats. Currently, the converter supports 25 formats for input, and 7 formats for output.

Table 1 Features and properties of several advanced data formats. Supported features are indicated by ✔; insufficient support is indicated by ✗. Only formats from standardization organizations or vendor-independent researchers are listed shown. In case were some criterion is not fulfilled, the evaluation of other criteria might not be completed (empty fields). (1) insufficient; (2) multiple data types but none with more than 16 bits; (3) ecg only; (4) multiple scaling factors are supported; (5) fixed to char[8], therefore about 25% of the units listed in [24] cannot be correctly represented; (6) 23 pre-defined units, (7) multiple files within a ZIP-container. (8) read (r)/write(w) support of BioSig .

Format	Multiple sampling rates and scaling factors	Multiple binary data types	Supports automated overflow detection	Representation of all Physical units from [24]	Patient info, recording equipment, investigator	Events, markers, annotations /	predefined event codes	random data access, streaming	Sensor position / orientation	Open Source Converter (8)	Single file	Score / OK
E1467[12]		✗						✗		✗		-3/✗
BCI2000 [3,4]	✗(4)	✔	✗	✗	✗	✔	✗	✔	✗	r✔	✔	-5/✗
BDF [25]	✔	✗	✗	✗(5)	✗	✔	✗	✔	✗	rw✔	✔	-6/✗
BKR [5]	✗	✗	✗	✗	✗	✗	✗	✔	✗	rw✔	✔	-8/✗
DICOM-Waveform								✗		✗		-2/✗
EBS	✗	✗			✗			✔		✗	✔	-4/✗
EDF [6]	✔	✗	✗	✗(5)	✗	✗	✗	✔	✗	rw✔	✔	-7✗
EDF+ [7]	✔	✗	✗	✗(5)	✗	✔	✗	✔	✗	r✔	✔	-6/✗
FEF [20]	✔	✔	✔	✔	✔			✗		✗	✔	-2/✗
GDFv1 [8]	✔	✔	✔	✗(5)	✗	✔	✔	✔	✗	rw✔	✔	-3/✗
GDFv2.0 [9]	✔	✔	✔	✔	✗	✔	✔	✔	✔	rw✔	✔	-1/✗
GDFv2.1 [9]	✔	✔	✔	✔	✔	✔	✔	✔	✔	rw✔	✔	0/✔
HL7aECG [14]	✔	✗	✗	✔	✔			✗		rw✔	✔	-3/✗
MFER [15]	✔	✔	✗	✗(6)	✔	✗	✗		✗	r✔	✔	-5/✗
OpenXDF [27]	✔	✔	✔	✔	✔	✔	✗	✗(7)		✗	(7)	-3/✗
Physio-bank [10]	✔	✗(2)	✗	✗	✗	✔ (3)	✔ (3)	✔	✗	r✔	✗	-6/✗
SCP-ECG [16]	✗	✔	✗	✗	✔	✗(3)		✔	✗	rw✔	✔	-5/✗
SIGIF [11]	✔	✔	✗	✔	✗		✗	✔	✗	r✔	✔	-4/✗
Unisens [22]	✔	✔	✗					✗		✗	✗	-3/✗

The GDF format has been in use for several years by the Graz Brain-Computer Interface Laboratory. More recently, support for GDF v2.1 has been implemented in the BCI2000 framework [3,4]. The BCI competitions 2005 (http://ida.first.fraunhofer.de/projects/bci/competition_iii/) and 2008 (http://ida.first.fraunhofer.de/projects/bci/competition_iv/) provided EEG data in GDF format. Recently, also a client-server archiving system based on GDF has been implemented. Further applications are investigated.

CONCLUSIONS

Due to varying requirements in biomedical signal processing, a variety of data formats for biomedical signals has been described. GDF has been developed with the aim to combine the best features from existing data formats. The present evaluation shows that this goal has been reached. GDF is also designed for possible extension to accommodate future needs. Several free software tools for accessing GDF data are available.

REFERENCES

1. BioSig – a free and open source software library for biomedical signal processing (2003-2008). Available online from http://biosig.sf.net
2. Schlögl A., Brunner C. BioSig: A Free and Open Source Software Library for BCI Research, *Computer*, vol. 41, no. 10, pp. 44-50, October, 2008.
3. G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw: BCI2000: A General-Purpose Brain-Computer Interface (BCI) System, *IEEE Trans Biomed Eng*, 51(6), June 2004. BCI2000 is available at http://www.bci2000.org/
4. J. Mellinger, G. Schalk: BCI2000: A General-Purpose Software Platform for BCI Research, In: G. Dornhege, J. del R. Millán, T. Hinterberger, D.J. McFarland, K.-R. Müller (eds.), Toward Brain-Computer Interfacing, *MIT Press*, 2007.
5. The BKR-format v2.07: Specification is available from: http://hci.tu-graz.at/%7Eschloegl/matlab/eeg/bkr.html
6. B. Kemp, A. Värri, A.C. Rosa, K.D. Nielsen and J. Gade. A simple format for exchange of digitized polygraphic recordings. *Electroencephalography and Clinical Neurophysiology*, 82(5):391-393, May 1992.
7. B. Kemp and J. Olivan . European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data . *Clinical Neurophysiology*, 114(9):1755-1761, Sep 2003.
8. A. Schlögl, O. Filz, H. Ramoser, G. Pfurtscheller, GDF - A general data format for biomedical signals Version 1.25 (GDF v1), 1999. Online available from http://hci.tugraz.at/%7Eschloegl/matlab/eeg/gdf4/TR_GDF.pdf
9. A. Schlögl, GDF – A general data format for biomedical signals Version 2.10. Specification available online http://arxiv.org/abs/cs.DB/0608052
10. http://www.physionet.org/physiotools
11. Cunha MB; Cunha JP; Oliveira e Silva T. SIGIF: a digital signal interchange format with application in neurophysiology. *IEEE Trans Biomed Eng*, 44(5):413-8, 1997.
12. ASTM E1467-94:2000. Standard Specification for Transferring Digital Neurophysiological Data Between Independent Computer Systems (Withdrawn 2004).
13. N. Stockbridge, B. Brown. Annotated ECG waveform data at FDA. *Journal of Electrocardiology*, 37(Suppl 1):63 – 64, Oct 2004.
14. HL7 annotated ECG: http://www.mortara.com/research_HL7%20Documents.htm
15. Medical waveform format (MFER) ISO/TS 11073/92001:2007.
16. Standard communications protocol for computer-aided electrocardiography (SCP-ECG). EN1064:2007 and ISO/DIS 11073/91064.
17. CEN/TC251/PT-40 (FEF) Public web site online available from http://www.cs.tut.fi/~varri/tc251/pt40/index.html
18. CEN/TC251/WGIV N98-11. New WI proposal, File Exchange Format for Vital Signs, available online at: http://www.tc251wgiv.nhs.uk/pages/pdf/iv811_1.pdf
19. Varri A, Kemp B, Penzel T, Schlögl A. Standards for biomedical signal databases. *IEEE Engineering in Medicine and Biology Magazine* 20(3): 33-37, 2001.
20. File exchange format for vital signs (*FEF) – CEN/TS* 14271:2003.
21. H. Wang, F. Azuaje, B. Jung and N. Black. A markup language for electrocardiogram data acquisition and analysis (ecgML). *BMC Medical Informatics and Decision Making* 3:4, 2003.
22. M. Kirst, J. Ottenbacher, R. Nedkov. UniSens – ein universelles Datenformat for Multisensordaten. Workshop Biosignalverarbeitung, Universität Potsdam, Germany, p.106-108, 16-18. July 2008.
23. A. Schlögl, B. Kemp, T. Penzel, D. Kunz, S.-L. Himanen, A. Värri, G. Dorffner, G. Pfurtscheller. Quality Control of polysomnographic Sleep Data by Histogram and Entropy Analysis. *Clin. Neurophysiol.* 110(12): 2165 – 2170, Dec. 1999.
24. ISO 11073-10101:2004 Health Informatics - Point-of-care medical device communications - Part 10101: Nomenclature p.62-75. Table A.6.3: Vital signs units of measurements
25. The Biosemi data format (BDF), 2008. Specification is available from: http://www.biosemi.com/faq/file_format.htm
26. A. Schlögl, F. Chiarugi, E. Cervesato, E. Apostolopoulos, C. Chronaki, Two-Way Converter between the HL7 aECG and SCP-ECG Data Formats Using BioSig. *Computers in Cardiology Conference*; 34:253 – 256, 2007. available online: http://www.cinc.org/Proceedings/2007/pdf/0253.pdf

Author: Doz. Dr. Alois Schlögl
Institute: Institute for Human Computer Interfaces
Street: Krenngasse 37/EG
City: Graz
Country: Austria
Email: alois.schloegl@ieee.org