Overview of data formats for biomedical signals.

Alois Schlögl <a.schloegl@ieee.org>

4. Nov. 2008

At the "18th meeting of: ISO/TC215 — HEALTH INFORMATICS WG7 — DEVICES, Jointly with 33rd meeting of CEN TC251 — HEALTH INFORMATICS WGIV — DEVICES at the Joint TC Working Meetings" was suggested to make "an open resource listing of the biosignal data formats currently available be compiled – together with some analysis of scopes, technical characteristics and usage". The following work is a draft to compile the relevant information.

The information in this document was compiled based on the experience of implementing import and export filters and data converters [26] for a number (more than 50) of different biomedical signal file formats within BioSig – the open source software library for biomedical signal processing [1,2]. The work consists of two parts, the first part compares the different data formats on eleven technical features, the second part discusses soft factors like usage and distribution of the data format. Open resources for biomedical signal formats are also compiled at wikipedia http://en.wikipedia.org/wiki/List_of_file_formats#Biomedical_Signals_.28Time_Series.29

This first draft might be incomplete, contain errors and be biased. In order to enable a correction of such flaws, this work is licensed with the "Creative Commons Attribution share-alike" (by-sa) license.



PART 1: Comparison of technical features (hard factors)

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.

- 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

- 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.

- Sensor positions

Standardized 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 (MEG). 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.

- Open Source software converters

are useful to address several purposes (i) it demonstrates the feasibility of the format, (ii) it can be used to demonstrate compatibility, or incompatibility between different formats, (iii) it increases the acceptance of a new format, and easies the migration to a new format. Therefore, it is am important criterion whether a format can be converted to some other formats.

There are a number of open source projects. However, BioSig is the highest ranked open source software project on SourceForge in the field of biomedical signal processing (search terms: "ECG, EEG"). And BioSig provides import and export filters as well as converters for a number of (about 50) different data formats. And BioSig is also used within other software projects (e.g. EEGLab, Fieldtrip, etc.) Therefore, one criterion is whether a specific format is supported by BioSig or not.

Vendor formats

Many vendors 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.

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 -2 or lower, indicating that at least two important features are missing. Only GDF gets higher scores, and only the latest version (GDF v2.1) gets the highest possible score of 0, because it addresses all issues listed above.

Table 1: Features and properties of several advanced data formats. Supported features are indicated by \checkmark ; insufficient support is indicated by \checkmark . 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 (7) muriple inc	Multiple sampling rates and scaling factors	Multiple binary data types	Supports automated overflow detection	Represetation 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]		Х						Х		X		-3/ X
BCI2000 [3,4]	X (4)	1	X	X	X	1	X	1	X	r✔	1	-5/ X
BDF [25]	•	Х	X	X (5)	X	1	Х	1	X	rw✓	1	-6/ X
BKR [5]	X	Х	X	Х	X	Х	Х	•	X	rw✓	1	-8/ X
DICOM- Waveform								×		Х		-2/ X
EBS	Х	Х			Х			1		X	1	-4/ X
EDF [6]	1	Х	Х	X (5)	Х	Х	Х	1	Х	rw✓	1	-7 X
EDF+ [7]	1	Х	Х	X (5)	X	1	Х	1	Х	r✔	1	-6/ X
FEF [20]	1	1	1	1	1			Х		Х	1	-2/ X
GDFv1 [8]	1	1	1	X (5)	Х	1	1	1	Х	rw✓	1	-3/ X
GDFv2.0 [9]	1	1	1	1	Х	1	1	1	1	rw✓	1	-1/ X
GDFv2.1 [9]	1	1	1	1	1	1	1	1	1	rw✓	1	0/•
HL7aECG [14]	1	Х	Х	1	1			Х		rw✓	1	-3/ X
MFER[15]	1	1	Х	X (6)	1	Х	Х		Х	r✔	1	-5/ X
OpenXDF [27]	1	1	1	1	1	1	Х	X (7)		Х	(7)	-3/ X
Physio-bank [10]	•	X (2)	X	X	X	✓ (3)	✓ (3)	•	×	r✔	Х	-6/ X
SCP-ECG [16]	Х	1	X	Х	1	X (3)		1	X	rw✓	1	-5/ X
SIGIF [11]	1	1	X	1	Х		Х	1	Х	r✔	1	-4/ X
Unisens [22]	1	1	Х					Х		X	Х	-3/ X

PART 2: Comparison of usage, supporting organization and other soft factors

This part addresses the purpose, scope, usage and supporting organization.

- ACQ AcqKnowledge File Format for Windows/PC is a vendor-defined format of Biopac Systems Inc., Goleta, CA, USA.
 - The recording systems of biopac seem to store the information natively in the ACQ format. The the specification of the ACQ format is publicly available, the format is used within the Biopac systems. It is not known whether other vendors support this format. The biopac product line seems to aim at education and research applications. There are no specific means to support quality control (patient identification, recording technician).
- ASTM-E1467 Standard Specification for Transferring Digital Neurophysiological Data Between Independent Computer Systems.
 is an ASCII-based format. This standard is not used and was withdrawn in 2004.
- BCI2000 was defined by the BCI2000 research project, Albany, NY, USA.
 The the specification of the BCI2000 format is publicly available, the format is used within the BCI2000 framework.
- BDF BioSemi data format from BioSemi B.V. Amsterdam, Netherlands. BDF is the native data format of the BioSemi recording systems. The BDF format is basically a 24bit extension of the EDF(1992) format. The only differences to EDF are:
 - the first 8 bytes have a different signature ("BIOSIG\xff") instead of "0".
 - the data samples are stored in int24 values instead of int16
 - markers are stored as bitmasks in the so-called status channel, which is different than the annotation channel of EDF+

The format is used within the BioSemi systems. Obviously, the need for BDF has arisen because EDF was unable to support anything else than int16 values.

BKR

it is a 16- data format for EEG data (single sampling rate, single scaling factor for all channels) and was developed in the 1990ties, at the University of Technology Graz, Austria. Nowadays, the format is not been used anymore but replaced by the GDF format.

CFWB

is the Chart Data File Format from ADInstruments Pty Ltd, Bella Vista NSW, Australia. It is used by the Chart software for data export and data interchange (The internal format ADICHT is proprietary). It is not known whether other vendors support this format.

DICOM — Waveform

is an attempt to store biosignal data within the Dicom framework. The current specification considers only ECG. The use of this solution is limited. The overhead of the DICOM framework is in general not accepted by the vendors of biosignal recording devices.

ecgML — A markup language for electrocardiogram data acquisition and analysis.
 has been proposed by a research group. It is an XML-based format for storing ECG and related data. It is different to the FDA-XML/HL7aECG and its not clear whether the format is used outside their research group.

• EDF — European Data Format.

EDF is a 16 bit data format supporting multiple sampling rates and multiple scaling factors. The original EDF was defined by a group of European researchers (Kemp et al 1992.) and was rather simple. This was considered one of the main advantages of EDF. Later, EDF+ was defined by Kemp and Olivan (2003), moreover a number of "implementation recommendations" are provided on the EDF web site. EDF is widely accepted in the research community for data exchange and many vendors provide import and export filters. However one must note that information is lost during conversion into the EDF format often losses some information available in the original (vendor-specified) file format. E.g. Impedances values (like in the brain vision format) are not maintained, or 22bit data (BioSemi) is truncated to 16 bits, manufacturer information (MFER, SCP) is not maintained.

Critical remark: The policy of EDF maintenance is based on the principle that the first EDF software should always be able to read newer data (This became clear in 1998 by discussion at the public EDF mailing list and was also emphasized by Alpo Varri at the Gothenborg meeting in May 2008). This principle is different to the approach that newer software should be able to read older data and makes it impossible to address certain new requirements in future (e.g. more than 16 bit, extended header information, support for automated overflow detection, etc.).

• FEF — File Exchange Format for Vital signs, CEN TS 14271.

The aim of FEF was to provide a general purpose format for biomedical signals. Its definition was based an ASN.1. FEF was not accepted by the industry, as far as I know only one vendor implemented the format.

• GDF — The General Data Format for biomedical signals was defined because the existing data formats were missing some important features (E.g. support for automated overflow detection, support for more than 16 bits, eventtable, predefined event codes, etc.), and because it was not possible to improve existing formats (see EDF). GDF was developed with the aim to incorporate the best features from all available data formats. In order to address that goal, fields for patient information (like in MFER and SCP) are included, Electrode position and Electrode impedances can be stored (like in the BrainVision format from BrainProducts), different data types (from int8 to float28) are supported, an eventtable and predefined event codes are used; up to 255 user-specified event types are possible; the physical units are encoded according CEN14271 (FEF). In order to support quality control, saturation detection, information about recording device (manufacturer, model serial number), and recording technician, software version are supported. And there is a mechanism in place to incorporate new features when new

GDF is routinely used within the Graz BCI lab for its Brain-Computer-Interface research; GDF is implemented within the BCI2000 framework. GDF is improved as new needs and requirements appear (e.g. manufacturer information was just recently added, because it is important for quality control of the data). Almost any biomedical signal data can be converted into GDF without loss of any information.

Open source Read/Write support is implemented in C/C++ as well as in Matlab/Octave.

HL7aECG — Health Level 7 annotated ECG.

requirements appear.

With version 3.0, HL7 will move to an XML-based standard. Within this HL7 v3.0, annotated ECG will be supported. Such an approach was recommend earlier by the FDA. One the OpenECG mailing list was stated that "the FDA XML format was developed so that pharmaceutical companies could lodge their drug trials ECG data in the FDA ECG Data Warehouse." The development of this approach is also fostered by Mortara (Italy), who claims to be the world leader in ECG technology. Based on some non-representative information, FDA-XML/HL7aECG is picked up by other vendors as well. The approach seems to address short-time ECG data (up to 40 s).

An open source implementation for this file format was encouraged by the OpenECG consortium which is also maintaining the SCP-ECG format.

MFER Medical waveform Encoding Rules ISO/TS 11073/92001. Format MFER has been developed in Japan. It is used to store various data formats including short-term, and long term ECG, EEG, and other signals. Presumably, the format is supported by Nihon-Kohden, although Nihon Kohden is also using the dataformat of the EEG1100 series. The use of MFER by other vendors is not clear. For example, the ETG4000 NIRS system from Hitachi (another Japanese vendor) is using an ASCII-based data format. MFER is dedicated to clinical application as it provides information about the recording device (manufacturer, model, serial number), patient identification (including age, weight, height, birthday, etc) and the recording technician, which are important information for quality control. A problem is that no all fields a fully specified. For example the remark on the field messages (tag=134,0x86) says "Message will be defined for each purpose"; neither is there a definition of "purpose" nor a list of definitions.

OpenXDF - Open Exchange Data Format

is an XML-based data format defined by Neurotronics, Inc. Gainesville, FL, USA. It claims that it is an extension of EDF. This is correct in the sense that there software can handle both formats. Basically, it is converting the header information into an XML-based format and uses a binary data section. Because of this transformation, other EDF-compatible software can not read the data. It is not clear whether any other vendor is supporting this format. No open source software is known that supports this format.

 SCP-ECG — Standard Communication Protocol for Computer assisted electrocardiography EN1064:2007, ISO/DIS 11073/91064.

SCP is a data format for storing short-term ECG data, and supports also the analysis of various ECG features and diagnostic information derived from the raw data. SCP supports also data compression, and automated ECG analysis. SCP is dedicated to clinical application includes fields for the information about the recording device (manufacturer, model, serial number), patient information, and the recording technician, which are important information for quality control.

Limitations: The specification contains a limit that no more than 65536 samples per channels can be stored. Also the channel identification is limited to ECG channels only. For this reason, SCP is a suitable for short-term ECG data.

- SIGIF A digital SIGnal Interchange Format with application in neurophysiology. was proposed by a group of Portuguese researches. It contains an ASCII header of variable length and a binary data section. It is not clear whether SIGIF is used outside their research lab. Biosig has implemented reading support for SIGIF data.
- Unisens was suggested [22] and provides a web site at http://unisens.org/. The fle format seems to consist

of multiple files (which can have different formats) and a common XML header. Some open source software is provided by the group. It is not clear whether the file format is used outside their lab.

- WFDB Format of Physiobank is the dataformat of Physiobank. It is mostly used for ECG data, WFDB supports different bit depth (8,10,12,16 bits). Each recording consists of an ASCII-based header file, and a binary data file. Optionally, there is an additional annotation file. BioSig supports reading the WFDB format, other open source tools are available from Physiobank.
- Vendor-specified formats.
 There is a huge number of vendor-specific format. Almost every equipment provider has developed their own data format e.g. ADInstruments Bella Vista, NSW, Australia (ADICHT, CFWB); ANT Enschede, The Netherlands (EEProbe); Best Medical Systems, Korneuburg, Austria (alpha-trace); Biopac, Goleta, CA USA (ACQ); BioSemi, Amsterdam, The Netherlands (BDF); BrainProducts, Gilching, Germany (BVA); Delsys, Boston, MA, USA (DEMG); Electrical Geodesics, Eugene, OR, USA (EGI); Flaga, Reykjavik, Iceland (Embla); Hitachi Medical Corporation Tokyo, Japan (ETG4000); EBNeuro, Firenze, Italy (GTF); Compumedics, Charlotte, NC, USA (Neuroscan); Micromed, s.r.l. Treviso, Italy (TRC), Neurotronics Inc. Gainesville, FL, USA (OpenXDF), CardinalHealth-NeuroCare, Madison, WI, USA (Nicolet); Nihon Kohden, Tokyo, Japan; Philips Healthcare, Best, The Netherlands (SierraECG); SIGMA Medizin-Technik GmbH, Thum, Germany (SIGMA); TMS International, Enschede, The Netherlands (TMS32, Portilab); Walter Graphtek GmbH, Lübeck, Germany (WG1); XLTEK, Oakville, Ontario, Canada; and many more.

Some vendors make the specification public available, several vendors provide the information upon request, but there are also other vendors who do not provide the specification and look-in the patient data to there own proprietary format. So far there is no agreement between the vendors for a common data format.

Summary:

A motto of ISO is "do it once, do it right, do it internationally" (see [28] p.3]). Currently several biomedical signal formats are under consideration as ISO standard (SCP-ECG, MFER, FEF/EDF [15,16,20]). Moreover, most companies use there own vendor-specific data format – a possible reason is that existing standards are not sufficient to address the needs in the field. Therefore, a single file format that addresses the needs of biomedical signal recording and processing systems is needed.

Such a data format should be able to address the various technical needs and requirements (see part 1). The acceptance of such a common data format will be certainly increased when open source implementations provide converters for smooth migration to the new format, and can can be used to validate the format specification.

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

- [27] J. Smith, J. Johnson, J. Schubert, R. Widell, A New File Format for Polysomnography Data. *Journal Sleep*, 28(11), 2005.
- [28] My ISO Job Guidance for delegates and experts". Available online http://www.iso.org/iso/my iso job.pdf