31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Extensible Biosignal Metadata
A Model for Physiological Time-series Data

David Brooks, *Student Member, IEEE*

***Abstract*— The domain specific nature of biosignal storage formats, along with the lack of support for metadata in general-purpose biosignal libraries, has hampered the easy interchange of biosignals between disciplines and their integration with physiological modelling software.**

Extensible Biosignal Metadata (XBM) is introduced as a standard framework to facilitate the sharing of information between and within research groups for both experimentalists and modellers; to help establish more web-accessible biosignal repositories; and, by using semantic web technologies, to result in the discovery of knowledge by automated reasoning systems.

I. Introduction

Most biosignal storage formats are domain specific. Metadata is specified in a variety of ways with no common standard to link information. This has disadvantaged multidisciplinary research and the integration of biosignals with physiological modelling software.

The users of biomedical and physiological biosignals have traditionally been clinicians and experimental physiologists. Different recording and storage formats have been developed, with standards evolving to allow groups to share information amongst themselves and use equipment from different manufacturers [1]. Standardisation has largely been driven by the requirements of cardiology (electrocardiogram, ECG), neurology (electroencephalography, EEG), and polysomnography (which uses a large range of biosignals, including EEG and ECG). Each discipline has developed their own standards and software tools for analysis, visualisation and annotation.

The systems biology and physiological modelling communities have developed standard modelling languages such as Systems Biology Markup Language (SBML) [2] and CellML [3] to facilitate the development and sharing of biological models. They have an ongoing programme to improve the sharing of data, models, software and knowledge [4]. This work includes the linking of models to experimental data including biosignals, encoded in markup-language data standards [5]. The lack of suitable standards has meant that comma-separated-value files have at times been used to store and exchange signal data [6], with the loss of any stored metadata.

A. Existing Biosignal Formats

Many different file formats exist to store and exchange biosignal data – the BioSig Project has a list of over 90 formats [7]. Information about some of the more common formats is summarised in Table I.

D. Brooks is with the Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand. d.brooks@auckland.ac.nz

Signals are generally stored as binary data; in a few cases Extensible Markup Language (XML) is used. Metadata and annotation are usually kept in predefined fields as an integral part of the signal file, with little, if any, ability to be extended. Metadata in XML-based formats is stored as freeform text using XML's tag and attribute structure — the more general Resource Description Framework (RDF) is not used.

Besides core biosignal metadata, the different formats can usually store metadata relevant to the area of study. If a general purpose format is to be adopted then this additional information must *not* be discarded — an extensible mechanism is required for domain specific metadata.

The biosignal format that appears to be the most domain neutral is ISO 11073-92001:2007 (MFER), which has been designed as a general specification for medical waveform data and to "*be used with other relevant protocols, such as HL7, DICOM, ISO/IEEE 11073, and database management systems for each purpose.*" [8]. The intent is for experts in specialised fields to develop relevant standards and use MFER to store waveform data.

PhysioBank [9] is an online archive that currently contains around 220 gigabytes of physiological signals and related data. Each recording in the archive usually consists of at least three files — the actual signal data in WaveForm DataBase (WFDB) format; a header file with metadata (text); and a binary annotation file. A set of software tools (PhysioToolkit) is available to work with the archive and includes a function library for users developing their own code.

Even though it is a widely used web resource, Physiobank does not provide a web-accessible metadata view of its contents.

There are two publicly available software libraries for accessing biosignal files: libRASCH [10], and The BioSig Project [11].

Both of these libraries enable biosignals to be used independently of the major formats. However they do not provide any consistent view of, nor structure to, metadata. While signal annotation and file headers can be read (and written), the meaning given to this information is specific to the particular format and not part of any generic model.

B. Metadata, Ontologies and the Semantic Web

Annotations and other metadata are often used to describe a biosignal and its features. This may be in natural language, with a controlled vocabulary used for consistency. More formally, terms and relations can be defined in an ontology, ensuring that the same things are referred to in the

TABLE I

COMMON BIOSIGNAL FORMATS

Name	*Format*	*Description*
DICOM Supplement 30 Waveform Standard (DICOM 30) [12].	Binary	Uses the DICOM imaging and communications standard for the interchange of clinical waveforms.
European Data Format (EDF and EDF+) [13], [14].	Binary	Originally designed for storage of polysomnograms. Widely used, particularly in Europe.
FDA format for ECG signals (FDA-ECG) [15].	XML	Part of the Health Level 7 (HL7) standards. Developed so that ECGs could be included in drug trial reports.
GDF - A general format for BIOSIGNALS [16].	Binary	Developed to meet the needs of research in Brain-Computer interfaces and to unify other biomedical signal processing fields. Similar to EDF and EDF+ in record structure.
Medical waveform Format Encoding Rules (MFER) [8].	Binary	ISO 11073-92001:2007 – designed to encode medical waveforms separately to metadata. Widely supported by Japanese equipment manufacturers.
Open eXchange Data Format (OpenXDF) [17], [18].	Binary	Signal data is contained within XML formatted metadata. Developed for polysomnography.
Philips ECG format [19].	XML	Philips have made their implementation of the FDA-ECG format available to users.
Standard Communications Protocol for ECG (SCP-ECG) [20].	Binary	ISO 11073-91064:2008 – an ECG format that is an ISO Standard.
WaveForm Data Base (WFDB) [9], [21].	Binary	The default format used by Physiobank, a major physiological signal repository. A range of open-source software tools are available.

same manner, and allowing for knowledge to be processed computationally in a comparable way to numeric data [22].

The World Wide Web Consortium's Semantic Web initiative [23] includes a layered set of specifications that allow data from disparate sources to be combined and integrated and defines languages to relate data to real world objects. The core components are the Resource Description Framework (RDF) [24] and the Web Ontology Language (OWL) [25]. OWL is based on a Description Logic (DL) which allows an automated reasoning system to infer knowledge using metadata annotation and rules.

Various organisations have defined standard terms and properties for describing resources, some of which are well established international standards; others are at different stages of development. Those that are relevant for biosignal annotation include:

- Dublin Core Terms [26] is a formal definition of properties and classes applicable to a wide range of resources.
- OWL-Time [27] defines temporal concepts such as an *Instant*, *Interval* and *DurationDescription* and relationships between them, such as *before*, *hasBeginning* and *intervalOverlaps*.
- The Timeline Ontology [28] defines a *TimeLine* as a coherent backbone for addressing temporal information.
- Open Biomedical Ontologies (OBO) [29] have a growing number of reference ontologies for biomedical domains, including the Foundational Model of Anatomy.
- The CellMLBiophysical/OWL Ontology [30] provides biophysical meaning to CellML entities, and can be used to visualise the biological model represented by a CellML model.
- The Physiology Reference Ontology is being developed to provide a symbolic representation of biological functions [31].
- The Cardiovascular Research Grid ECG Ontology is being developed for describing all aspects of ECG data collection, ECG waveform features, and ECG data analysis [32].

Disciplines such as sleep medicine have a well defined set of annotation terms specified in a reference manual [33]. It is expected that terms and concepts such as these will become increasingly available in a format suitable for automated processing and reasoning. Using RDF for biosignal metadata provides an extensible way to *future-proof* a general purpose format so that new ontologies can be used for annotation terms, as they become available.

II. ABSTRACT MODEL

Biosignals are used for a wide variety of purposes ranging from individual healthcare and clinical medicine through to experimental research and physiological modelling; they are collected, processed, displayed, analysed, annotated and stored in many different ways; metadata provides the *glue* that holds everything together. This is illustrated in Fig. 1

In a very general sense, a *biosignal* is any kind of measurable time-varying quantity that is the direct result of a biological process. Here we consider a biosignal to consist of a sequence of time-varying data points (i.e. a time-series) which has been obtained from a biological signal by sampling; a biosignal may consist of multiple channels or signals (e.g an EEG). Intrinsic to a signal is the notion of time; expressing temporal relationships is an important function of biosignal metadata. Fig. 2 shows some common metadata terms usually associated with a biosignal. This model is the basis of a Biosignal ontology which also imports concepts from Owl-Time, Timeline and other ontologies referred to above.

It is vital when working with metadata that everyone has a common understanding of not only terms (e.g. ECG,

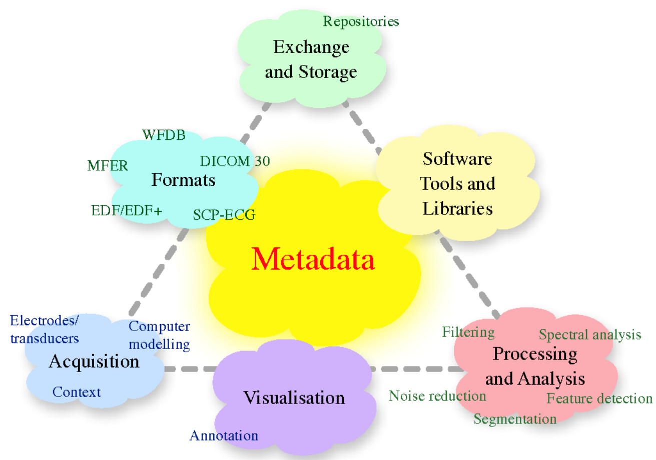

Fig. 1. Biosignals are collected, processed, displayed, analysed, annotated, stored and used in many different ways.

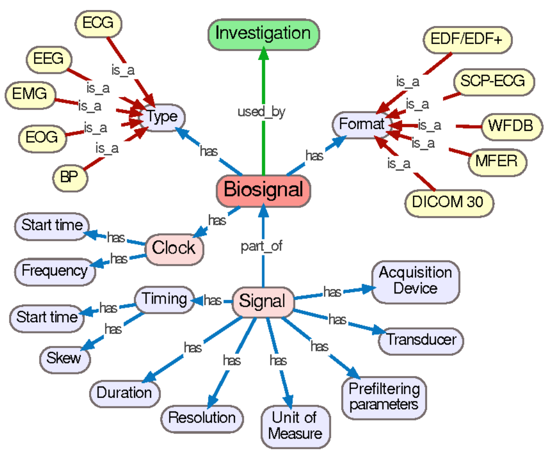

Fig. 2. Some common metadata associated with a biosignal.

frequency, REM sleep), but also of relationships — what is meant by *has*, *part_of* or *used_by*? The Open Biomedical Ontologies consortium includes a Relation Ontology [34] that specifies a methodology to be used when defining relations in biomedical ontologies. This will be used by the Biosignal Ontology when defining relationships for biosignal metadata.

As an example, Fig. 3 displays RDF metadata (serialised as RDF/XML) stating that the object identified by *biosig:SIG1234* has a format of *xbm:FormatMFER* and is used by *abi:VF1257* — this could be the identifier of a research study.

III. Extensible Biosignal Metadata – XBM

A common standard is required for effectively working with metadata across multiple biosignal formats and a diverse range of applications. In a similar way to Adobe's Extensible Metadata Platform [35], Extensible Biosignal Metadata (XBM) is being developed to standardise the definition, creation, and processing of biosignal metadata. It will:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xbm="http://biosignalml.org/xbm/1.0/"
  xmlns:abi="http://bioeng.auckland.ac.nz/projects/1.0/"
  xmlns:biosig="http://biosignalml.org/biosig/1.0/">
  <rdf:Description rdf:about="biosig:SIG1234">
    <dc:date>2009-04-02T16:51:23.561Z</dc:date>
    <dc:format rdf:resource="xbm:FormatMFER"/>
    <xbm:used_by rdf:resource="abi:VF1257"/>
  </rdf:Description>
</rdf:RDF>
```

Fig. 3. Metadata about the *biosig:SIG1234* object.

- Define standard terms for describing biosignals using formal ontologies.
- Use the Resource Description Framework model with metadata serialised as RDF/XML.
- Where possible, hold metadata in existing biosignal files.
- Reconcile XBM metadata with existing metadata components of current file formats.

Some biosignal file formats allow for general text to be stored as part of the file. e.g. MFER can contain an arbitrary number of 1024 byte *message* blocks; EDF+ supports *EDF Annotations* signals that can contain general text. This gives us a way to embed RDF/XML metadata into a biosignal file without conflicting with existing applications that process the file.

An important component of the XBM framework is the specification and implementation of an open, multi-platform, software library. A key function provided by the library will be the ability to set and retrieve temporal metadata relationships and properties — *"What other signals have unusual behaviour around the time this signal is ...?"* — and to make such queries against a database or repository.

In order for existing, non-XBM aware, applications use current values, metadata carried *natively* in a file format must always be seen as definitive. The software library will ensure that those metadata elements able to be held natively are kept consistent with RDF/XML metadata.

This library will also have functions to retrieve and store signal data held in common biosignal file formats. The use of such a library by domain specific annotation and analysis tools would result in biosignal files containing metadata in a standard format, which when stored in suitable repositories, would allow both data and metadata to be easily incorporated in other investigations.

One aspect of biological systems is their multi-scale nature. Biosignals reflect this by having regions of interest ranging from very short time scales to much longer ones. The biosignal software library needs to reflect this by providing *views* of signals at different temporal resolutions, both of signal data and for metadata annotation.

IV. Discussion and Conclusion

This paper has focused on a standard way to work with biosignal metadata rather than suggest a standard format for biosignal data. Of the several existing data formats, MFER stands out — it is domain neutral; an ISO standard; and has a brief, easily implemented, specification. A new format could be defined around a scientific data format such as Hierarchical Data Format (HDF5) [36] or be based on a binary serialisations of XML such as Fast Infoset [37] or Efficient XML Interchange (EXI) [38]. However, why reinvent the wheel? Especially given one group of medical device manufacturers have adopted MFER.

What is noticeable about existing biosignal formats and general-purpose libraries is their fragmented approach to metadata. Extensible Biosignal Metadata (XBM) is proposed as a standard means to work with and extend biosignal metadata that is compatible with existing signal formats and toolsets. This will allow the integration of biosignals into modelling environments such as CellML, not only for input data but also to store results, and also to automatically update metadata in the biosignal.

An XBM framework would facilitate the sharing of information between and within research groups for both experimentalists and modellers; would help establish more web-accessible biosignal repositories; and, by using semantic web technologies, result in the discovery of knowledge by automated reasoning systems.

V. Acknowledgements

This work is supported by the Auckland Bioengineering Institute through a studentship and conference funds. I am grateful for the fruitful discussions and critical reading of this paper by Dr. Mark Titchener and Dr. Bruce Smaill of the Auckland Bioengineering Institute.

References

[1] A. Varri, B. Kemp, T. Penzel, and A. Schlogl, “Standards for Biomedical Signal Databases,” *Engineering in Medicine and Biology Magazine, IEEE*, vol. 20, no. 3, pp. 33–37, May/Jun 2001.

[2] M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden *et al.*, “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models,” pp. 524–531, 2003.

[3] C. Lloyd, M. Halstead, and P. Nielsen, “CellML: its future, present and past,” *Progress in Biophysics and Molecular Biology*, vol. 85, no. 2-3, pp. 433–450, 2004.

[4] H. Sauro, D. Harel, M. Kwiatkowska, C. Shaffer, A. Uhrmacher, M. Hucka, P. Mendes, L. Stromback, and J. Tyson, “Challenges for Modeling and Simulation Methods in Systems Biology,” *Winter Simulation Conference, 2006. WSC 06. Proceedings of the*, pp. 1720–1730, 2006.

[5] P. Hunter, W. Li, A. Mcculloch, and D. Noble, “Multiscale Modeling: Physiome Project Standards, Tools, and Databases,” *Computer*, vol. 39, no. 11, pp. 48–54, 2006.

[6] B. Caldwell, in conversation with author, June 2008.

[7] A. Schlögl. Scientific data formats. [Online]. Available: http://www.dpmi.tu-graz.ac.at/schloegl/matlab/eeg/

[8] ISO/TS 11073-92001:2007, *Health informatics – Medical waveform format – Part 92001: Encoding rules*. Geneva, Switzerland: International Standards Organisation, 2007.

[9] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.” *Circulation*, vol. 101, no. 23, pp. e215–e220, June 2000.

[10] R. Schneider. About libRASCH. [Online]. Available: http://www.librasch.org/librasch/

[11] A. Schlögl. The BioSig Project. [Online]. Available: http://biosig.sourceforge.net/

[12] DICOM Standards Committee, Working Group 1, *Supplement 30: Waveform Interchange*. National Electrical Manufacturers Association, September 2000. [Online]. Available: ftp://medical.nema.org/medical/dicom/final/sup30_f2.pdf

[13] B. Kemp, A. Värri, A. Rosa, K. Nielsen, and J. Gade, “A simple format for exchange of digitized polygraphic recordings,” *Electroencephalography and Clinical Neurophysiology*, vol. 82, pp. 391–393, 1992.

[14] B. Kemp and J. Olivan, “European data format ‘plus’(EDF+), an EDF alike standard format for the exchange of physiological data,” *Clinical Neurophysiology*, vol. 114, no. 9, pp. 1755–1761, 2003.

[15] U.S. Food and Drug Administration. Providing Digital Electrocardiogram (ECG) Data. [Online]. Available: http://www.fda.gov/cder/regulatory/ersr/default.htm#ECG

[16] A. Schlögl, “GDF - A general dataformat for BIOSIGNALS,” *CoRR*, vol. abs/cs/0608052, 2006.

[17] J. Smith, J. Johnson, J. Schubert, and R. Widell, “A New Format for Polysomnography Data,” *Sleep*, vol. 28, no. 11, p. 1473, 2005.

[18] OpenXDF Consortium, *Open eXchange Data Format Specification, Version 1.1*, December 2008. [Online]. Available: http://www.openxdf.org/spec.php

[19] S. Zhou, G. Guillemette, R. Antinoro, and F. Fulton, “New approaches in Philips ECG database management system design,” *Computers in Cardiology, 2003*, pp. 267–270, 2003.

[20] ISO 11073-91064:2008, *Health informatics – Standard communication protocol – Part 91064: Computer-assisted electrocardiography*. Geneva, Switzerland: International Standards Organisation, 2008.

[21] M. Costa, G. B. Moody, I. Henry, and A. L. Goldberger, “PhysioNet: An NIH Research Resource for Complex Signals,” *J Electrocardiol*, vol. 36 Suppl, pp. 139–144, 2003.

[22] O. Bodenreider and R. Stevens, “Bio-ontologies: current trends and future directions,” *Briefings in Bioinformatics*, vol. 7, no. 3, p. 256, 2006.

[23] World Wide Web Consortium. W3C Semantic Web Activity. [Online]. Available: http://www.w3.org/2001/sw/

[24] ——. (2004, February) Resource Description Framework (RDF): Concepts and Abstract Syntax. [Online]. Available: http://www.w3.org/TR/rdf-concepts/

[25] ——. (2004, February) OWL Web Ontology Language Overview. [Online]. Available: http://www.w3.org/TR/owl-features/

[26] DCMI Usage Board. (2008, January) DCMI Metadata Terms. [Online]. Available: http://dublincore.org/documents/dcmi-terms/

[27] World Wide Web Consortium. (2006, September) Time Ontology in OWL. [Online]. Available: http://www.w3.org/TR/owl-time/

[28] Y. Raimond and S. Abdallah. (2007, October) The Timeline Ontology. [Online]. Available: http://purl.org/NET/c4dm/timeline.owl

[29] The OBO Foundry. The Open Biomedical Ontologies. [Online]. Available: http://www.obofoundry.org/

[30] P. Hunter and P. Nielsen, “Distributing and maintaining models in CellML,” in *SIAM Conferences on the Life Sciences*, Montreal, 2008.

[31] D. Cook, J. Mejino, and C. Rosse, “Evolution of a Foundational Model of Physiology: symbolic representation for functional bioinformatics,” *Proceedings MedInfo 2004*, pp. 336–340, 2004.

[32] CardioVascular Research Grid. Ontologies. [Online]. Available: http://www.cvrgrid.org/?q=Ontologies

[33] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. Westchester, IL.: American Academy of Sleep Medicine, 2007.

[34] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse, “Relations in biomedical ontologies,” *Genome Biology*, vol. 6, p. R46, 2005.

[35] Adobe Systems Incorporated. Extensible Metadata Platform (XMP). [Online]. Available: http://www.adobe.com/products/xmp/

[36] The HDF5 Group. HDF5 Home Page. [Online]. Available: http://www.hdfgroup.org/HDF5/

[37] ISO/IEC 24824-2:2006, *Information technology – Generic applications of ASN.1: Fast Web Services*. Geneva, Switzerland: International Standards Organisation, 2006.

[38] World Wide Web Consortium. Efficient XML Interchange Working Group. [Online]. Available: http://www.w3.org/XML/EXI/