33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

BioSignalML – a meta-model for biosignals

David J. Brooks, Peter J. Hunter, Bruce H. Smaill and Mark R. Titchener

Abstract— **The multitude of biosignal file formats used in research has hampered the easy exchange of biosignals and their use with physiological modelling software. We describe an abstract data model that accommodates the diversity of formats, along with a software implementation which links biosignal data into the Semantic Web, using existing data formats. Initial application of our work is to sleep study research.**

I. Introduction

Many different file formats exist to store and exchange biosignal data – well over 100 different formats are listed at http://pub.ist.ac.at/~schloegl/biosig/TESTED [1]. Several formats are likely to be used for a given project, not by design, but because manufacturers, software vendors and in-house developers are unlikely to be using a common format, with archived recordings in yet other formats; the numbers compound when collaborating with other research teams. Although a number of conversion tools exist (e.g. libRASCH [2] and the BioSig Project [3]) they do not provide any standard view of metadata contained in a signal file. The lack of a standard, domain-neutral framework for working with biosignals has hampered their easy interchange between disciplines and integration with physiological modelling software [4]. The BioSignalML data model described in this paper is intended to address this issue.

As part of the Semantic Web initiative [5], the World Wide Web is evolving into a Web of Linked Data [6]. This is providing a framework in which raw data is linked to arbitrary things (including other data sets, standard vocabularies, data producers and consumers), and is enabling automated reasoning to be applied to vast heterogeneous data sets. Standardising and sharing the meaning of metadata is crucial to this effort.

In order to address problems arising from the multitude of formats and lack of standardised metadata, the BioSignalML project [7] is developing a standard framework for the exchange and storage of physiological time-series data (biosignals). This framework:

- encapsulates common features of biosignal file formats in an abstraction layer;
- uses ontologies to define biosignal terms and attributes;
- can be extended to incorporate domain-specific concepts and terms;
- provides software tools and libraries that allow the use of disparate signal file formats;
- includes a repository component that allows recordings to be accessed using standard web software as a Linked Data resource.

We are initially working with biosignal recordings that are associated with sleep research. Polysomnograms contain a comprehensive range of physiological signals, including those measuring brain, heart, breath, eye and muscle activity, often at different sampling rates. Recordings are usually over several hours, with possibly several recordings for a single patient. A variety of signal storage formats are used, from proprietary to European Data Format (EDF and EDF+) [8], [9], Stanford Data Format (SDF) [10] and WaveForm DataBase (WFDB) [11].

Our work is also being used to facilitate the use of biosignals with physiological modelling languages such as CellML [12]. A HDF5 [13] based file format is being specified for biosignal exchange and storage along with a streaming format for real-time connections between simulation components.

The remainder of this paper is structured as follows: Section II describes a simple model for working with biosignals; Section III outlines the design of a biosignal repository based on the model; and Section IV discusses our work and looks at related research before concluding.

II. BioSignalML Abstract Model

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. Intrinsic to a signal is the notion of time; expressing temporal relationships is an important function of biosignal metadata.

In order to be able to work with the wide range of biosignal formats a number of concepts are defined in an *Abstract Model*:

- A **Recording** is the set of signals from a single recording session. Storage formats usually keep these multiple signals in a single file (e.g. EDF) or group of files held together (e.g. SDF, WFDB).
- A **Signal** is the set of time-varying values of some measurable quantity. The physiological signals we are dealing with are usually oversampled with a regular sampling period.
- Within the context of the abstract model, an **Annotation** has a time associated with it, and may be thought of as a signal – of textual comments or of values from

D. Brooks, P. Hunter, B. Smaill and M. Titchener are with the Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand. (e-mail: d.brooks@auckland.ac.nz, p.hunter@auckland.ac.nz, b.smaill@auckland.ac.nz, mark@tcode.auckland.ac.nz)

a small, predefined set (e.g. sleep-stage, scored every 30-seconds; an ECG beat annotation).
- An **Event** is something that happens in time and has a time of occurrence and usually a duration (e.g. an obstructive sleep apnea).

Annotations and events may be directly linked to a particular signal or to a recording as a whole.

Data points, annotations and events all have one or more times associated with them; these are temporal positions in some coordinate system that represents time, a **Timeline**. Using concepts from The Timeline Ontology [14], timelines are continuous or discrete, with related timelines linked together using mapping functions.

A recording has a continuous timeline with the zero point being the start of the recording; this timeline is usually embedded in a local or universal timeline (e.g. Coordinated Universal Time (UTC)). A sampled signal has a discrete timeline and mapping that relates sample number to elapsed time, which we call a **Clock**. Clocks may be classified as either uniform or irregular; several signals may share a single clock.

A. Identifying Signals and Recordings

In order to make metadata statements and assertions about biosignals we need a general way of identifying them. Uniform Resource Identifiers (URIs) are compact sequences of characters that identify an abstract or physical resource [15] and are widely used to identify resources on the Internet.

BioSignalML specifies that all objects are identified by a URI. More than one URI may be used to identify a given object, in which case a metadata statement must be made stating that the URIs identify the same resource.

B. Biosignal Metadata

What are general attributes of biosignals? What does a particular signal represent? When was it recorded? How and by whom? What processing has been applied? What is the purpose of a recording besides being a collection of signals? Do other people know our meaning of the terms we have used to describe properties?

Current biosignal file formats usually have a limited number of fields for metadata and these are usually only pertinent to the domain the format was designed to be used in. These fields often contain free-format text without a controlled vocabulary to specify content, leading to possible future ambiguity.

We use ontologies to provide meaning to terms and relations – an ontology can be defined as a formal way of specifying concepts [16] to ensure that the same thing is referred to in the same manner. They allow for knowledge to be computationally processed in a similar way to numeric data [17].

BioSignalML specifies a core set of terms and relationships for metadata. Some of these are defined in the BioSignalML ontology [18]; others are taken from existing standard ontologies; additional domain-specific concepts and terms can easily be added.

Metadata statements about biosignals are made using the Resource Description Framework (RDF) [19], with the Web Ontology Language (OWL) [20] being used to specify ontologies. Both RDF and OWL allow for biosignal data to be integrated into the broader context of the Semantic Web without restricting future applications and extensions.

Some ontologies applicable to biosignal annotation are well established international standards; others are at different stages of development. Both the OBO Foundry [21] and NCBO BioPortal [22] provide repositories of publicly available biological and biomedical ontologies. Ontologies directly relevant to our work include: Dublin Core Terms [23]; the Timeline Ontology [14]; the Foundational Model of Anatomy [24]; the Relation Ontology [25]; the CellML-Biophysical/OWL Ontology [26]; the Physiology Reference Ontology [27]; the Cardiovascular Research Grid ECG Ontology [28]; and the Sleep Domain Ontology [29].

The abstract model's concepts have been realised as a set of objects and methods as part of developing an Application Programming Interface (API) and software library. This library allows signals and their metadata to be created and accessed in a format independent way and forms the basis of a web-accessible biosignal repository.

III. BioSignalML Repository

Biosignal recordings are usually stored as files on a computer system in whatever format they were recorded in. Exchanging recordings with colleagues will often involve file copying and possibly format conversion. The emphasis is usually on obtaining signal data for processing and analysis without accompanying metadata, this being kept and exchanged in the form of laboratory notes.

An alternative to simply working with recordings as computer files, BioSignalML provides a repository application in which metadata is treated as a first-class component of a biosignal. The repository has been designed as an easy-to-use, extensible, cross-platform resource that integrates with existing signal processing workflows.

The repository stores BioSignalML objects or resources – Recordings, Signals, Annotations and Events. Each instance of an object has its own URI; a request for an object returns a representation of the object. This could be actual signal data in some format; a HTML web page describing the object; or a RDF description of the object complete with metadata links to other resources, with the particular type of representation depending upon the request. The repository provides a HTTP interface and may be accessed using standard web-browsers.

Signal recordings submitted to the repository are kept in their original format. Metadata about BioSignalML objects contained in the recording is extracted into a RDF triple-store. Each recording has a Named Graph [30] holding its metadata, so that all metadata from the recording is treated as a single resource, which allows statements to be made for provenance and access control. Domain-specific metadata is mapped to ontological terms via separate user-editable mapping statements, allowing future extension as ontologies are developed and refined.

In order to manage collections of recordings, the BioSignalML abstract model is extended with two additional concepts – a **Collection** is a group of Recordings that share a common research project or investigation, and an **Archive** is a group of Collections that all pertain to the same domain or area of study.

Signal recordings are either exchanged with the repository as files or as a telemetry stream, for realtime input/output with processing and simulation environments. File formats currently supported include EDF/EDF+, WFDB, SDF, a HDF5 based format, and a proprietary one; adding new formats requires a software module for conversion into the BioSignalML abstract model along with mapping statements for domain-specific metadata.

Fundamental to the repository and its interfaces is the use of URIs to identify biosignals. In line with Linked Data guidelines [31], the BioSignalML repository uses the "http" scheme for URIs, except for local filesystem resources. Because full "http" URIs can be unwieldy for users, the API and user tools allow base prefixes to be given and used to construct relative URIs. When a new recording is submitted, any URIs required for Signals are formed by appending "/signal/N" to the Recording's URI, where "N" is the numerical index of the signal in the recording.

Temporal segments of a Recording or Signal can be requested by using a comma separated list of intervals as the query component of the object's URI (i.e. following a "?")[1]; an interval can be either in the form of "start-end" or "start:duration". Actual time values are normally expressed in seconds; sample indices can also be used. Signal data returned in response will always span the requested interval(s); data outside of the request may also be returned, depending on the actual format requested.

IV. Discussion and Conclusion

Our abstract-model based repository has recently been deployed in a modelling and simulation environment to hold both reference patient recordings and simulation results. Amongst other things, recordings will be used to validate processing algorithms and test equipment, a role in which provenance is important. User feedback will help determine extensions and revisions to the repository and associated tools.

In a modelling environment such as CellML it is imperative that actual signals used as parameters have the correct units specified in a model. As an example, many signals are measured in microvolts. Different biosignal file formats represent this in various ways – EDF has a free-text field without restriction, and gives an example of "μV" [8]; EDF+ restricts this field to standard texts and uses "uV" for microvolts [9]; we found an SDF file generated by proprietary software using "μV"; the Medical waveform Format Encoding Rules (ISO 11073-92001:2007) specify that "V" be used with an exponent of 10^{-6} [32].

[1]The fragment component (following a "#") can not be used because fragment information is intended solely for user agents and is not passed to a server [15].

And what about the "μ" symbol? Unicode [33] has two different characters that look the same in some fonts – a "micro" sign, which is encoded as U+00B5, and the Greek letter "mu", which is encoded at U+03BC. Any naive check based on string comparison will not find the different character strings to be representations of the same thing.

The RDF statements:

```
:signal-uri bsml:units
    <http://www.w3.org/2007/ont/unit#V> ;
  bsml:unitsExponent "-6"^^xsd:integer .
```

unequivocally state that the signal's physical units are in microvolts; this could be defined using other ontologies provided statements are made relating the definitions – an automated reasoner would then know their equivalence.

Archived sets of reference biosignals are invaluable for comparing different models and processing algorithms, and it is important that these datasets contain as much metadata as possible, even information that may appear inconsequential at recording time. As an example, the PhysioBank [34] repository includes arterial pressure signals which could be used to validate an anatomically correct blood flow model. A sample recording selected from this repository however did not state from what artery the pressure was recorded, nor the position of the cannula needle along the artery, making the particular signal of little value to our proposed use.

A. Related Work

FieldML [35] is a meta-language for describing fields, with initial application in field visualisation and large-scale field computation. Given that a signal can be considered to be a field on a timeline, a simple FieldML description can be made for it. However, rather than seeing signals as general fields, experimentalists are used to working with biosignals as time-series and using signal processing tools, which is where our work is directed. An extension to our work would be to create conversion tools between FieldML and BioSignalML.

PhysioBank [34] is an online archive that currently contains around 700 gigabytes of physiological signals and related data. Even though it is a widely used Internet resource, PhysioBank does not use RDF nor ontologies for describing metadata; instead information about the recordings is presented as textual web pages with signal annotations held in binary files. Our work can provide a wrapper around PhysioBank so that it becomes a Semantic Web resource, simplifying the process of selecting suitable signals (e.g. the above example of validating a blood flow model).

Physio-MIMI [36], [37] is a data integration project being developed for clinical researchers, combining data across institutions without requiring a common data model. Its initial application is for sleep research data, but the system is able to be generalised. While Physio-MIMI provides a federated view of SQL databases, our repository provides a SPARQL view of RDF triplestores. Our work with metadata is aimed at linking biosignals into the emerging Semantic

Web (Web 3.0), in contrast to Physio-MIMI which appears to be applying Web 2.0 technology to institutional databases.

B. Conclusion

This paper has provided an overview of BioSignalML and presented a standard model for working with biosignals. The mapping of signal attributes into an abstract model enables applications to be format neutral, without requiring signals to be converted to a new format; the use of URIs, backed by ontologies, links signal repositories to the Semantic Web, allowing web-based tools to enquire and reason over information; using a general purpose format (RDF) for metadata future-proofs the model, enabling new ontologies to be used as they become available.

The future worth of biosignal reference sets will be enhanced by collecting as much metadata as possible at the time of recording, using standard vocabularies and ontologies.

V. Acknowledgements

D. Brooks is grateful for financial support from the Auckland Bioengineering Institute.

References

[1] A. Schlögl. Dataformats supported by BioSig http://biosig.sf.net. [Online]. Available: http://pub.ist.ac.at/~schloegl/biosig/TESTED

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

[3] A. Schlögl, "An overview on data formats for biomedical signals." in *Image Processing, Biosignal Processing, Modelling and Simulation, Biomechanics*, ser. IFMBE Proceedings, O. Dössel and A. Schlegel, Eds., vol. 25/4, World Congress on Medical Physics and Biomedical Engineering. Springer, September 2009, pp. 1557 – 1560.

[4] D. Brooks, "Extensible Biosignal Metadata – A Model for Physiological Time-series Data," in *Engineering in Medicine and Biology Society, 2009. IEEE 31st Annual Conference*, 2009.

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

[6] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, no. 1, pp. 1–136, 2011.

[7] IEEE Engineering in Medicine and Biology Society. (2008) BiosignalML. [Online]. Available: http://www.embs.org/techcomm/tc-cbap/biosignal.html

[8] 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.

[9] 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.

[10] Z. Sahul, O. Carrillo, and J. Black. (2006, January) SDF: Stanford Data Format. [Online]. Available: http://www.physionet.org/standards/npsg/Sahul.pdf

[11] G. B. Moody, *WFDB Applications Guide*, 10th ed. MIT Room E25-505A, Cambridge, MA 02139, USA: Harvard-MIT Division of Health Sciences and Technology, March 2011. [Online]. Available: http://physionet.org/physiotools/wag/wag.htm

[12] 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.

[13] M. Folk, A. Cheng, and K. Yates, "HDF5: A file format and I/O library for high performance computing applications," in *Proceedings of Supercomputing'99 (CD-ROM)*, 1999.

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

[15] T. Berners-Lee, R. Fielding, and L. Masinter. (2005, January) RFC 3986 Uniform Resource Identifier (URI): Generic Syntax. [Online]. Available: http://tools.ietf.org/html/rfc3986

[16] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International Journal of Human Computer Studies*, vol. 43, no. 5, pp. 907–928, 1995.

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

[18] D. Brooks. (2011, April) The BioSignalML Ontology. [Online]. Available: http://www.biosgnalml.org/ontologies/2011/04/biosignalml

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

[20] ——. (2004, February) OWL Web Ontology Language Overview. [Online]. Available: http://www.w3.org/TR/owl-features/

[21] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.

[22] N. Noy, N. Shah, P. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. Rubin, M. Storey, C. Chute *et al.*, "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, no. suppl 2, p. W170, 2009.

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

[24] C. Rosse and J. Mejino, "A reference ontology for biomedical informatics: the Foundational Model of Anatomy," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 478–500, 2003.

[25] 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.

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

[27] 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.

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

[29] S. Arabandi, "Developing a Sleep Domain Ontology," in *Signs, Symptoms and Findings: Towards an Ontology of Clinical Phenotypes*, Milan, Italy, September 2009. [Online]. Available: http://bimib.disco.unimib.it/images/7/7c/SSFW09_Arabandi_SleepOntology.pdf

[30] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named graphs, provenance and trust," in *Proceedings of the 14th international conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 613–622. [Online]. Available: http://doi.acm.org.ezproxy.auckland.ac.nz/10.1145/1060745.1060835

[31] T. Berners-Lee. (2006, July) Linked Data. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

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

[33] The Unicode Consortium, *The Unicode Standard, Version 6.0.0*. Mountain View, CA: The Unicode Consortium, 2011. [Online]. Available: http://www.unicode.org/versions/Unicode6.0.0/

[34] 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.

[35] G. Christie, P. Nielsen, S. Blackett, C. Bradley, and P. Hunter, "FieldML: concepts and implementation," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1895, p. 1869, 2009.

[36] G. Zhang, "Physio-MIMI," in *Sleep Research Network Conference, October 30, 2009*. Washington, DC, USA: Sleep Research Network, October 2009. [Online]. Available: http://www.wpic.pitt.edu/research/srn/PhysioMIMI-SRN-Zhang.pdf

[37] G. Zhang, T. Siegler, P. Saxman, N. Sandberg, R. Mueller, N. Johnson, D. Hunscher, and S. Arabandi, "VISAGE: A Query Interface for Clinical Research," *AMIA Summits on Translational Science Proceedings*, vol. 2010, p. 76, 2010.