HDF5 based data format for archiving complex neuro-monitoring data in Traumatic Brain Injury patients

Manuel Cabeleira [1], *Ari Ercole* [2], *Peter Smielewski* [1]

[1] *Brain Physics Lab, Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK*

[2] *Division of Anaesthesia, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK*

Address: Neurosurgery Unit, Level 4, A Block Addenbrooke's Hospital Cambridge biomedical Campus, Cambridge, CB2 0QQ
Email: mc916@cam.ac.uk
Tel: +441223 336946

Abstract

Modern neurocritical care units generate high volumes of data. This data is originates from a multitude of devices in various formats and levels of granularity. We present a new data format intended to store this data in an ordered and homogenous way.

The adopted data format was built on the hierarchical model based on the HDF5 format capable of dealing with a mixture of small and very large data sets with equal ease. It is possible to access and manipulate individual data elements directly within a single file and is extensibile and versatile. Excellent compression is possible and comprehensive support exists for platforms including MATLAB, R and Python.

The new file format was implemented in ICM+ software and validated as part of the CENTER-TBI project where it has been used to storage all acquired ICU monitoring data from the participating centres across Europe.

Keywords: **HDF5, Multimodal monitoring, Data storage, Data format**

Introduction

High volumes of data are generated during a patient's stay in any modern ICU and neurointensive care in particular. Data comes not only from a multitude of bed-side monitors (ABP, ICP, ECG) but also, lab data, manual measurements/ observations and event annotation. With the introduction of electronic medical record systems (EMRs) this data is captured and archived electronically. Currently, a notable exception to this is the routine storage of high-resolution, full waveform data. These are generally displayed at the bed-side but only summary values are archived. Such a situation is problematic for research and clinical care where as important clinical information can be extracted from multimodal waveform data.

Despite numerous attempts to address these interoperability issues with unifying medical communication standards, notably the ISO standard IEEE11073 (1), none have managed to truly fulfil the needs for physiological data, and thus have not been widely adopted. A number of proprietary formats has emerged but these all suffer from limitations making the creation of multi-centre databases difficult. Developing a well-annotated standard for data archiving is of paramount importance to facilitate further advances in computer supported individualised management of patients.

We propose a prototype format that that can be used as the foundation for the creation of a standard that covers all the needs of modern critical care environment. This format has been implemented for the CENTER-TBI study, (2) where it will be used as the main data storage format for high resolution data generated across Europe.

Materials and Methods

This project builds on the hierarchical model developed by the open source HDF5 file format specification (3). This file format is already extensively used in other areas of science where extensive amounts of data need to be stored(4–6). In particular it has found application as a standard for experimental neuroinformatics (7).

HDF5 is attractive for its flexible format offering a self-describing hierarchical, tree like structure capable of holding almost any kind of well annotated data objects in a single file. To add to its versatility, most of the most widely used scientific tools already provide libraries to handle this type of files (MATLAB, Python, Java, C), with MATLAB even adopting the file format as its primary storage format. An HDF5 file is composed of 2 main building blocks: Groups and Datasets (Figure 1). The groups form the tree's stem and branches and are therefore responsible for the hierarchical organization of the data. Every dataset is a uniform multidimensional array of data objects or elements, represented by one of the predefined simple data-types or compound elements composed of a mixture of data-types. Groups and datasets can be further annotated with metadata contained in associated 'Attributes'.

. HDF5 may be compressed for storage efficiency and the format reduces data access time and facilitates access and extraction of only data subsets.

The following requirements for the file internal structure design were defined:

- Accommodate all data types from bedside monitors at full temporal resolution.
- Each data object fully described by its attributes.
- Accomodation of structured clinical annotation. Provision for metadata.
- Completely self-described.

Results

HDF5 data file structure

Our file structure is presented in Figure 1. The data is divided in different groups (categories) (Table 1). Each group contains a specific data type associated with it and a set of attributes.

Waveform and numeric data sets include one data set per modality or, in case of composite recordings such as EEG, one group of datasets (individual channels) per modality (see Figure 1). Each data set in turns include a series of uninterrupted (continuous) data streams characterised by its position in the data set, its actual start time, and its sampling frequency detailed in the dataset 'Index Table' attributes (Table 3, Figure 2). In addition each of those data

sets also has an attribute data quality tracking (Table 3, Figure 2). Episodic (Table 3) and annotation data (Table 3) contain individual data samples (simple, or composite), each with its own time stamp, and a quality flag where relevant.

In addition, the file structure also contains: Summaries, Definitions and Clinical annotations groups as well as two data sets, PatientInfo and Presentation, all of which are described in Table 1 and Table 3.

Finally, a set of attributes was chosen for the root group containing the most important metadata describing the dataset

Data Compression and performance

Data compression intrinsic to the HDF5 standard is used to minimize overall file size. For time series datasets (numeric, waveform and electrophysiology) the Scale Offset algorithm was used. For the summaries group, the GZip algorithm was used with a power level of 6 for its handling of NAN values (ensuring seamless Excel compatibility). This data compression was shown to be effective (Table 2) and capable of processing an average of 1.7Mb HDF5 data/second. Within-file data access time was negligible.

This format has already been used to archive and transfer to a central repository more than 100 recordings from 22 different centres across Europe. The average file size was 512 Mb containing an average of 7.4 days of data and at least ABP, ICP and ECG full resolution waveforms.

Discussion

The HDF5 data platform proved to be robust, extensible, scalable, self-describing and easy to handle. One important advantage of our format is the capacity for easy manipulation of its contents. In particular it allows post-creation removal, addition or transformation of datasets. This feature allows supplementing the ICU recordings with external data. Unlike other formats

it is possible to extract subsets of data without parsing the whole file making handling very large files practical.

The compression ratios observed in this file format were very good compared to others allowing the creation of lightweight databases and facilitating the overall file handling, yet still retaining excellent performance of the data archiving process.

Conclusions

The developed format can be used to archive multi-centre data in a homogenous, robust, self-described and accessible way. We propose our format as a prototype for a standard for clinical physiological data from intensive care. The creation of such standard is of paramount importance not only to ease inter-centre collaborations, but also to simplify planning and execution of future multi-centre studies.

Acknowledgements

This work was supported by a European Commission FP7 award (CENTER-TBI).

References

1. Franklin DF, Ostler DV. Proposed standard IEEE P1073 Medical Information Bus: medical device to host computer interface network overview and architecture. In: Eighth Annual International Phoenix Conference on Computers and Communications 1989 Conference Proceedings [Internet]. IEEE Comput. Soc. Press; [cited 2016 Oct 26]. p. 574–8. Available from: http://ieeexplore.ieee.org/document/37448/

2. Maas AIR, Menon DK, Steyerberg EW, Citerio G, Lecky F, Manley GT, et al. Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): a prospective longitudinal observational study. Neurosurgery [Internet]. 2015 Jan [cited 2016 Oct 31];76(1):67–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25525693

3. HDF5 File Format Specification Version 3.0 [Internet]. Available from: https://support.hdfgroup.org/HDF5/doc/H5.format.html

4. Dougherty MT, Folk MJ, Zadok E, Bernstein HJ, Bernstein FC, Eliceiri KW, et al. Unifying Biological Image Formats with HDF5. Commun AMC. 2009;52(10):42–7.

5. Rees N, Billich HR, Koziol Q, Wintersberger E, Götz A, Pourmal E, et al. Developing HDF5 for the Synchrotron Community. 2015;WEPGF063.

6. Rübel O, Prabhat M, Denes P, Conant D, Chang E, Bouchard K. BRAINformat: A Data Standardization Framework for Neuroscience Data. bioRxiv. 2015;

7. Eglen SJ, Weeks M, Jessop M, Simonotto J, Jackson T, Sernagor E, et al. A data repository and analysis framework for spontaneous neural activity recordings in developing retina. Gigascience [Internet]. 2014 Dec 26 [cited 2016 Oct 31];3(1):3. Available from: http://gigascience.biomedcentral.com/articles/10.1186/2047-217X-3-3

Table 1 - Division of data into logical categories – HDF5 groups *EEG/ECoG are part of a group that is inside the waveforms group **Patient.info and presentation are not associated to any group * Composite types are explained in Table 3**

Data Type	Modalities	HDF5 Dataset Type	Attributes	Description
Numerics	• ETCO2 • SPO2 • PBTO2 • HR • Temperature	1-dimentional Array of 32-bit Float Numbers	• Index Table (Composite Array) • Quality Table (Composite Array) • Units (String) • Location (String) • Metric (String) • Modality (String) • Source (String)	• For all the low temporal resolution data (sampled <= 1Hz)
Waveforms	• ABP • ICP • ECG • CVP • EEG* • ECoG*			• For high temporal resolution, waveforms
*EEG/ECoG (inside waveform group)	• EEG Channels			• Group inside the Waveforms used to store EEG data.
Summaries	(Processed data) • Minute-by-minute • Hourly	1-dimentional Array of a composite type***	• Index Table (Composite Array) • Column labels (String) • Title (String)	• Group which stores synchronised Excel spread sheet type, time synchronised, representation of the numerics and waveforms data sets, produced by averaging values with granularity of 1 minute in one dataset, and 1 hour in the other.
Episodic	• Microdialysis • Clinical observations		• Columns Labels (String) • Title (String) • Textual description (String)	• For irregularly sampled data and manual measurements
Clinical Annotations	• Nursery events • Spontaneous notes			• Including data sets with labelled events and textual notes (ex. Nursing events)
Definitions	• eventTypes • indexStruct • qualityRef • qualityStruct			• Group containing details, with descriptions, of any complex data formats used for datasets or attributes. The group is also meant to contain physiological variables' and events' labels nomenclature.
**Patient.info	(Patient Demographics)		• Columns Labels (string) • Title (string) Textual description (string)	• A data set in the root group that contains name=value pairs of patient demographics, if such patient identifier are to be stored in the files, otherwise may be omitted
**Presentation	(Clinical information)			• A data set in the root group that contains name=value pairs of any auxiliary clinical data fields that are useful to keep together with the physiological data, and which do not de-anonymise the data (example, GCS at admission, type of trauma, etc).

Table 2 - Acheived compression rates

Original file type	Average compression rates achieved
ICM+ Raw	3.37
LabChart	4.99
Draeger proprietary binary	1.62
Moberg Binary	2.8

Table 3 - Table describing the composite data types.*Depending of the specific needs of the definition table these 2 composite arrangements can be found

Composite Type	Data arrangement	Descriptions
Summaries	• [**Time-stamp** (64-bit float), **1xN** 32-bit float]	• **Time-Stamp** – Excel date time format (number of days since 1/1/1990 • **1xN** – 1 Dimensional array with one value for each (**N**) variable collected in the file
Episodic	• [**Code** (String), **Time-stamp** (64-bit Float), **Duration** (32-bit Float), **Comment** (String), **Value** (32-bit Float)]	• **Code –** Textual code for the variable/observation being inserted • **Time-Stamp** - UNIX format time stamp (milliseconds since 1/1/1970) • **Duration** – Duration of the duration/effect of the variable being inserted (to use only if applicable) • **Comment** – Textual comment on the variable/observation being inserted (to use only if applicable) • **Value** – Numerical value of the variable/observation being inserted (to use only if applicable)
Clinical Annotations	• [**Code** (String), **UNIX Time-stamp** (64-bit Float), **Duration** (32-bit Float), **Comment** (String)]	• **Code –** Textual code for the Clinical Annotations being inserted • **Time-Stamp** - UNIX format time stamp (milliseconds since 1/1/1970) • **Duration** – Duration of the duration/effect of the Clinical Annotations being inserted (to use only if applicable) • **Comment** – Textual comment on the Clinical Annotations being inserted (to use only if applicable)
*Definitions	• [**Name** (String), **Description** (String)] **or** • [**Value** (32-bit unsigned integer), **Description** (String)]	• **Name** – Name of the element being described • **Description** – Textual description of the element. **or** • **Value** – Numeric code of the element being described • **Description** – Textual description of the element.
Index table	• **Start index**(64-bit integer), **Start time** (64-bit Unsigned integer), **Duration** (64-bit integer), **Sampling frequency**(64-bit float)	• **Start index** - index of the first sample in this continuous data block • **Start time** - modified UNIX format time stamp (microseconds since 1/1/1970) • **Duration** - number of data samples in this data block • **Sampling frequency** - data sampling frequency [Hz]
Quality Table	• [**TimeStamp**(64-bit unsigned integer),**Code**(32-bit unsigned integer)] • [**TimeStamp Code**] • [64-bit unsigned integer 32-bit unsigned integer]	• **Time-Stamp** - UNIX format time stamp (milliseconds since 1/1/1970) • **Code** - (bit)set of quality indicators valid for data starting at this time point until the next quality indicator or until the end of data
Patient.info (Patient Demographics) Presentation (Clinical information)	• [**Field** (String), **Value** (String)]	• **Filed** – Textual name of the patient or clinical information • **Value** – Textual value of the

Figure 1 – (Left) HDF5 hierarchical data storage concept (Right) structure of the proposed neurocritical care physiological data file

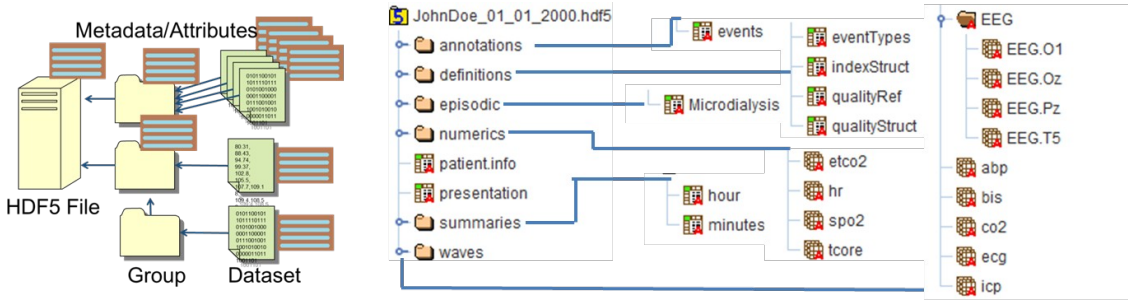

Figure 2 - The concepts of the Index Table (Left), and the Quality Table (Right)