On data and medatata formats for electrophysiological experiments

Jiří Vaněk, Roman Mouček
Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia
Univerzitní 8, 306 14, Pilsen, Czech Republic
Email: vanek2@kiv.zcu.cz, moucek@kiv.zcu.cz

***Abstract*—Despite standardization efforts there is still no widely used standard format for storing electrophysiological data. This standard is necessary for effective collaboration between scientists. This work deals with adjustments of existing general data/metadata model for electrophysiological experiments and proposal/implementation of HDF data format for their storage.**

I. Introduction

Brain research has been very popular recently. Experimental electrophysiological procedures produce large data and it is very important to store them with associated metadata. It was common that experiments were designed, measured and analyzed and after evaluation this recorded data were deleted. However, during recent years it has become more obvious to store these data and metadata for later use and share them with other researchers to provide subsequent independent analyses.

To facilitate data and metadata sharing it is important to identify and develop an independent standard for storing and exporting experimental electrophysiological data and metadata. If this standard is accepted by a larger community, it will allow easy sharing and better understanding of experimental results.

Currently there exist many models and formats for storing electrophysiology data and some terminologies and ontologies for storing electrophysiology metadata. Although at our neuroinformatics laboratory we are performing electrophysiology experiments. Experimental data are first stored in of open proprietary formats (BrainVision format) and finally with associated metadata (organized in custom metadata structures) stored in the web portal application - the EEGBase Portal [5].

International Neuroinformatics Coordinating Facility (INCF) Electrophysiology Task Force for data sharing [16] works on providing standards for these data and metadata. The mapping of the data and metadata model used in our EEGBase Portal (including extension of the standard metadata terminology) to the standards introduced within INCF Electrophysiology Task Force are the focus of this article.

The paper is organized in the following way. First data sharing culture and INCF Program for Data Sharing are introduced. Then two selected formats for storing electrophysiology data, the NIX format and BrainVision format, are described. Moreover, HDF data model and odML initiative for metadata description are introduced. Then the mapping of BrainVision format to the NIX format and mapping of EEGBase metadata terminology to general metadata terminology are described. The next part deals with design, implementation, and testing of HDFExport Program that converts existing data and associated metadata stored in the EEGBase Portal into HDF5 container in which the standardized structures are defined. Conclusion section sums up and discusses the work results.

II. State of the Art

A. Data Sharing

A trend towards sharing of neuroinformatics data has emerged in recent years. Nevertheless, a number of barriers continue to impede easy sharing of experiment's data. Many researchers and institutions remain uncertain about how to share data or lack the tools and expertise to participate in data sharing. The motivation for sharing is:

- **to accelerate progress in understanding of the brain**
  Several researchers claim that more rapid scientific discoveries are possible with shared data [2] [1].
- **to improve data quality**
  The sharing data helps uncover mistakes as missing data, noise, errors, etc. and improves the quality of the data in future experiments [3].
- **to reduce cost of research**
  For example, neuroimaging research is costly both in terms of the data acquisition costs and the time spent in data documentation. A significant amount of money could be saved from redundant data acquisition if data were shared with appropriate metadata descriptions [3].

B. Program on Standards for Data Sharing

INCF is an international non-profit organization devoted to advancing the field of neuroinformatics. The INCF Program on Standards for Data Sharing was established to specify a standard for storing neuroinformatics data. This Program also aims to develop generic standard and tools to facilitate the recording, sharing, and reporting of neuroscience metadata in order to improve practices for the archiving and sharing of neuroscience data. Metadata define the methods and conditions of data acquisition and subsequent analytical processing, Metadata also describe conditions under which the actual raw-data were acquired [4].

The current focus of the Program on Standards for Data Sharing is in two areas: neuroimaging and electrophysiology.

The most important requirement of such a standard is to accommodate common types of data used in electrophysiology or neuroimaging and also the metadata required to describe them [4].

C. Current Formats for Storing Electrophysiology Data

Many of the existing formats for storing electrophysiology data are proprietary and even though some of them are well documented, they are complicated to use or edit due to their licensing policy. Focusing on the open ones, the most known and used formats are Ovation [22], NeXus Format [20], NEO [18], NeuroHDF [19], EDF+ [17], and NIX (Pandora) [21]. Most of them use the HDF format for storing electrophysiology data. Also Electrophysiology Task Force of the INCF Program on Standards for Data Sharing in requirements for a standard recommends basing a standard on HDF5 [8]. In the following sections only the NIX format and BrainVision format (used in our lab) are described.

1) NIX format: The NIX project (previously called Pandora) started in the context of the Electrophysiology Task Force. The NIX format specification closely defines an inner structure of file, especially the data part. The meta data part is defined by odML (described in Section II-E).

NIX uses highly generic models for data as well as for metadata and defines standard schemata for HDF5 files representing these models. The design principle of the data model used by NIX was to create a rather minimalistic, generic, yet expressive model that is able to represent data stored in other widely used formats or models like Neuroshare or NEO without any loss of information. Due to its generic approach, the data model is also able to represent other kinds of data used in the field e.g. image data or image stacks [21]. NIX provides a convenient C++ library to simplify the access to the defined format.

The NIX format's scheme is shown in Figure 1. Experimental data are stored in blocks. Each block identifies an experiment and related metadata section. Raw data (signals, stimuli) are stored in DataArrays and specified by the attributes Dimension, Sample, Set, Representation, and Range (Figure 1) and could be specified by the section DataTag. The stimuli and artifacts are stored in the section SimpleTag (one stimulus) or MultiTag (more stimuli). The source for the sections DataArrays and/or Tags could be specified by the section Source. Each section could contain a link to the metadata part that contains information about experiment.

The NIX format's scheme serves as a base to which EEGBase data model (based on the BrainVision file format) is mapped.

2) BrainVision Format: Electrophysiology data in our neuroinformatics laboratory are recorded using the BrainVision Recorder [11] and stored into three files. The format of these files (BrainVision file forma)is described in the BrainVision Recorder User Manual [10]:

- **data file**
  This is binary file which contains recorded values from a recording device. The data are stored as double numbers.
- **vhdr file**
  This text file consists of various named sections containing key values. It includes basic information about measuring: coding, data file name, marker file name, number of channels, sampling interval in microseconds, information about binary format (IEEE_FLOAT_32), and information about channels (channel number, channel name, unit, resolution of unit).
- **vmrk file**
  This marker file is based upon the same principle of sections and key values as the header file is. This text file contains information about stimuli: stimuli number, type of the stimuli, stimuli description, position, size and their channel numbers.

These files are stored (together with associated matadata) in the EEGBase Portal [5].

D. Hierarchical Data Format

HDF is a data model, file format and library for storing extremely large and complex data collections. This technology is able to store any kind of data and is used all over the world in research centers and government agencies. For example, the format HDF5 (specific HDF implementation)is used by Cardiff University for resolving their problem with grid computing, Deutsche Bank for financial engineering, Diamond Light Source in synchrotron science, or Laboratory for Neural Computation for bio-engineering.

A lot of formats for storing electrophysiology data use HDF5. "The grouping structure in HDF5 enables applications to organize data objects in HDF5 to reflect complex relationships among objects. The rich collection of HDF5 datatypes, including datatypes that can point to data in other objects, and including the ability for users to define their own types, lets applications build sophisticated structures that match well with complex data. The HDF5 library has a correspondingly rich set of operations that enables applications to access just those components that are important." [14]

HDF is similar to XML documents, self-describing HDF files allow users to specify complex data relationships and dependencies. Several APIs for programming languages C, C++, Fortran 90, Java and others are available for this format. HDF is open-source (BSD license). Stored data are human readable and the metadata model is easily customized.

E. Open Metadata Markup Language

The metadata in electrophysiology domain are indispensable for the analysis and management of experimental data within a lab. However, only rarely are metadata available in a structured, comprehensive, and machine-readable form [13]. odML defines the metadata format, not the content, it means that it is inherently extensible and can be adapted to the specific requirements of any laboratory. For data sharing a correct understanding of metadata and data is only possible if the same terminology is used or if mappings between terminologies are provided. For this purpose terminologies with definitions of commonly used terms were assembled. [9]

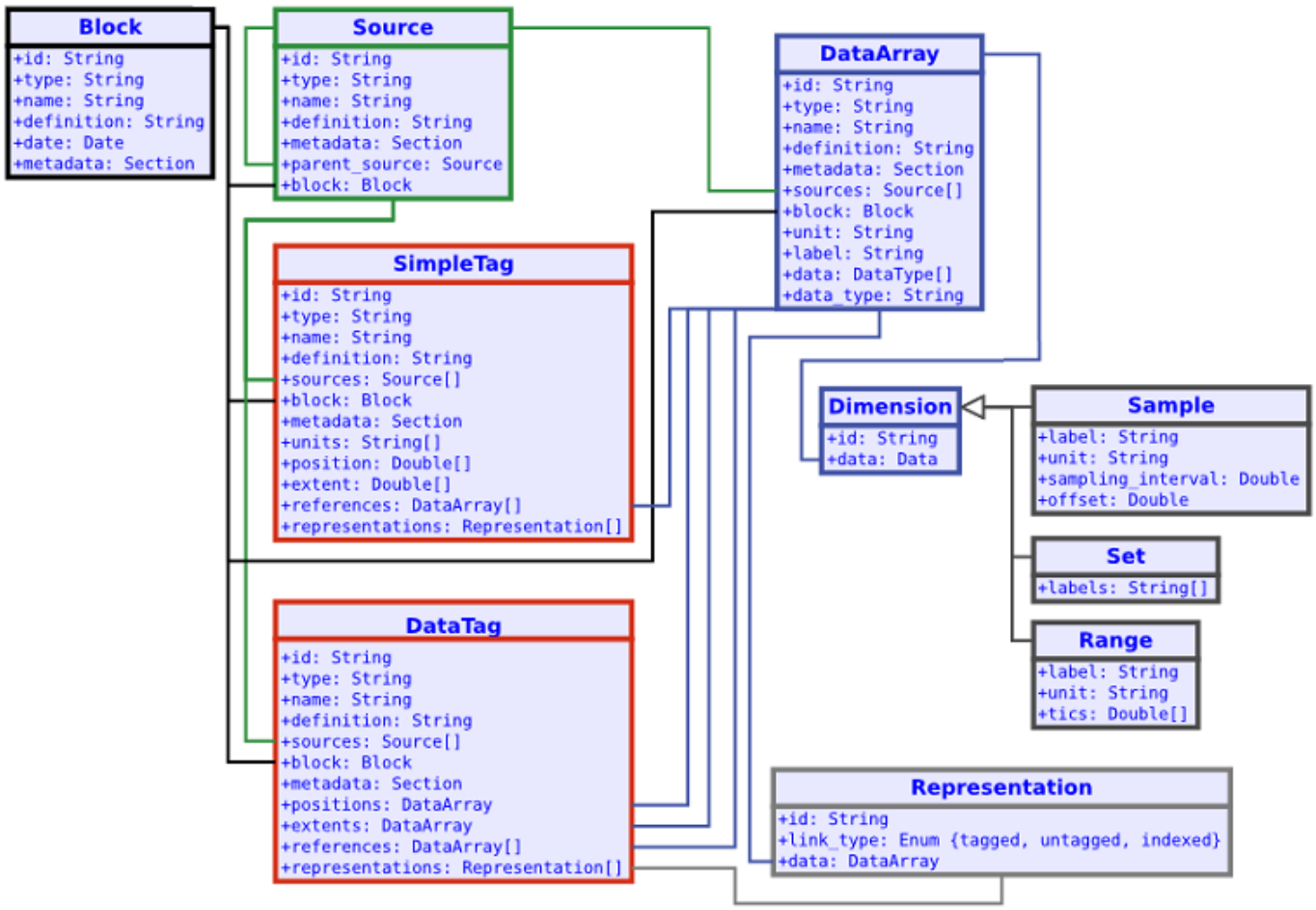

Figure 1. NIX data scheme. [21]

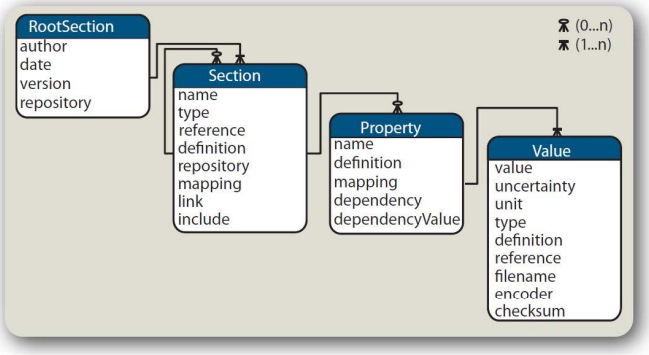

Figure 2. Open metadata Markup Language Entity-Relation diagram. [13]

III. Mapping of Data Models and Metadata Terminologies

This section describes the mapping of the EEGBase data model to the NIX data model and comparison of odML metadata terminology with the EEGBase metadata terminology. The EEGBase model is divided into two autonomous parts DATA and METADATA, which relate to each other, but could be read or written separately.

A. Data Model

The EEGBase data model is based on the NIX data model (Figure 1 and Section II-C1). Because the NIX model is able to save data from any electrophysiology experiment, it is too general for our lab purposes. As a result we used only the necessary parts of the model while other sections were omitted. All the omitted parts are in the NIX model optional, in this way the EEGBase data model is compatible with the NIX definition. The EEGBase format's scheme is described in Figure 3. It uses the NIX scheme of sections Block, DataArray, MultiTag, DataTag and SimpleTag. The section Block is used to divide measurement, the section DataArray is used for storing raw signal data and stimuli, the section MultiTag is used for storing stimuli information, and the section DataTag contains EEG channel information. DataArrays are divided into SIGNAL and MARKER parts for better transparency. The names of DataArrays also correspond to the names of channels. These adjustments allow better human readability of the file and do not influence information content at all. Metadata necessary for description of raw data form a part of the EEGBase format's scheme (they are also included in the

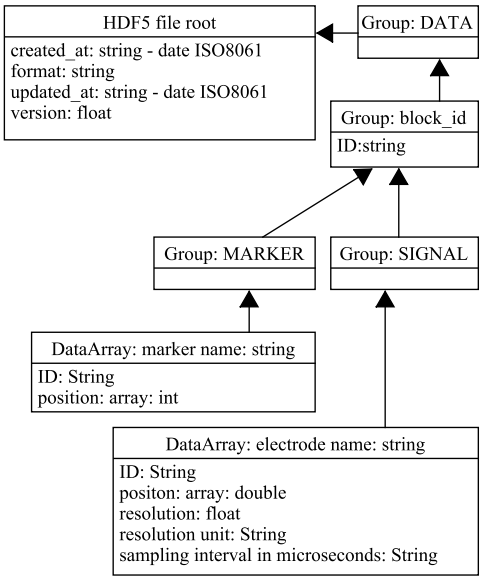

Figure 3. The EEGBase data model. The data are stored in a tree structure using fixed terminology. Each Group MARKER and SIGNAL could contain zero or more DataArrays sections containing raw data.

NIX model).

B. Metadata Model

EEGbase metadata scheme and its terminology are organized according to odML general metadata scheme and its terminology. Since the EEGBase metadata scheme used terms that had not been included or not had an alternative term in the original odML terminology, the odML terminology was extended with these terms. These changes and adjustments are described in Section III-C.

C. Metadata Terminology Extensions

In order to store all our metadata we extended the odML terminology. These extensions were committed to G-Node INCF GitHub repository [7]. New sections **Environment**, **Protocol** and **Software** and several attributes to the existing sections **Person** and **Electrode** were added. All suggested changes were accepted by odML developers. All modifications are listed in Table I.

IV. HDFExport Program

This section deals with design and implementation of HDFExport program that converts the BrainVision file format data and EEGBase portal metadata (stored locally or in the EEGBase Portal in sql and/or no-sql repositories) into the HDF5 container.

[1]Added terminology that describes usage of electrode.
[2]For floating licenses

Table I. Modifications of the odML model.

Name	**Property**	**Value**	**Definition**
Electrode	Usage	Ground	Usage of electrode.[1]
Electrode	Usage	Reference	Usage of electrode.[1]
Electrode	Usage	Channel	Usage of electrode.[1]
Electrode	Description	String	
Environment	Weather	String	
Environment	RoomTemperature	String	
Environment	AirHumidity	float	The air humidity in %.
Environment	Description	String	
Protocol	Description	String	Description of the experiment
Protocol	Author	person	The persons who create this protocol.
Protocol	ProtocolFile	binary	Protocol File.
Protocol	ProtocolFileURL	URL	URL of protocol file.
Protocol	Version	String	Version of the protocol.
Person	Education level	String	Highest archived education level of the person.
Person	Role	Subject	The role of this person.
Person	Email	String	Person's e-mail.
Person	PhoneNumber	String	Person's phone number.
Person	Laterality	String	Handedness - The dominant hand of the subject.
Software	Name	String	The software name.
Software	Owner	String	The owner of software.
Software	Developer	String	Developer or developers firm of the software.
Software	Version	String	Version of the software.
Software	License	String	License type.
Software	LicenceStart	date	The start date of time limited license.
Software	LicenceExpiration	date	The end date of time limited license.
Software	LicenceDuration	String	Duration of the license for the software.
Software	LicenceCount	int	Number of the software license.[2]
Software	Distribution	String	Distribution type.
Software	Description	String	
Software	LicenceDuration	String	Duration of the license for the software.

A. Use Cases

The HDFExport program is designed for several use cases that could be identified by data and metadata location (Figure 4):

- data and metadata from locally stored BrainVision files are only exported,
- data and metadata from locally stored BrainVision files and metadata from the EEGBase Portal are exported,
- data and basic metadata from BrainVision files stored in the EEGBase Portal are exported, most experimental metadata are not exported,
- all data and metadata stored in the EEGBase Portal are exported.

We can also identify use cases according to the type of exported data (Figure 5):

- Only raw EEG data and basic metadata are exported. Only data and metadata from the BrainVision files are exported, stimuli are not exported (not all measurements use stimuli).
- Raw EEG data, basic metadata and stimuli are exported. It means that all data, metadata and stimuli from BrainVision files are exported.

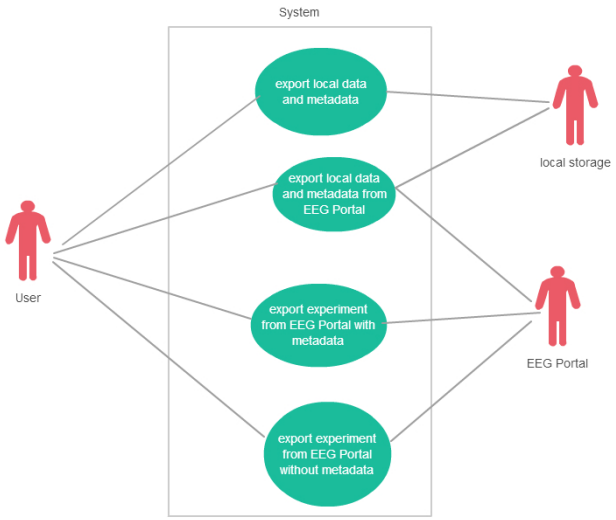

Figure 4. Use cases of HDFExport Program.

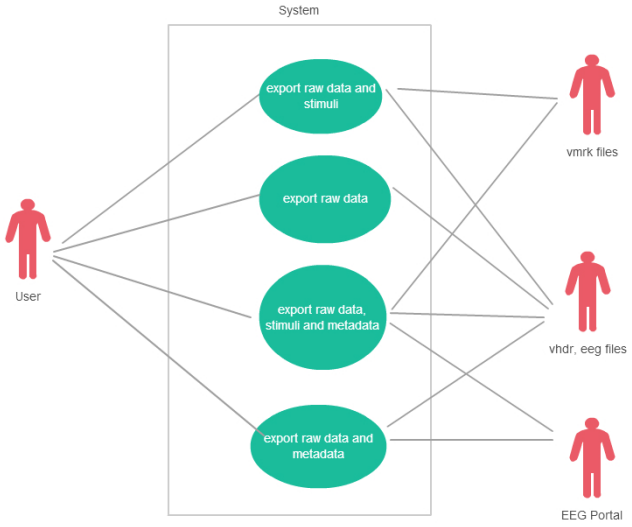

Figure 5. Use cases of HDFExport Program.

- Raw EEG data, basic metadata, stimuli and experiments metadata are exported. It means that all data, metadata and stimuli from the BrainVision files and experiments metadata from the EEGBase Portal are exported.
- Raw EEG data, basic metadata and experiments metadata are exported.

B. Implementation

Java was selected as a programming language for program development, since a parser for BrainVision files had been already developed. Moreover, the EEGBase Portal is written in Java. There is an effort to integrate the export program into the EEGBase Portal.

V. Tests

The program was manually tested for several use cases and all created HDF5 files were verified manually (they were opened and the contents was checked) using HDFView [6] in version 2.11, the software tool provided by the HDF Group.

Table II. Performance tests with GZip compression. Each file was saved seven times.

File size of BrainVision files	Time needed for conversion in ms			
	Best	Worst	Average	With compression
21,3 MB	3010	3501	3132	7529
51,2 MB	6160	7906	6700	26051
58 MB	7668	8262	7947	29032
87,2 MB	31364	35957	32820	44877
197,8 MB	13060	14625	13502	40168
221,3 MB	14589	16321	15197	42679
221,6 MB	14347	16108	14871	43586

Table III. File size tests.

File size of BrainVision files	HDF5 file size	HDF5 file size with compression	Ratio
21,3 MB	80,9 MB	26,92 MB	1,26x
51,2 MB	194,2 MB	112,98 MB	2,21x
58 MB	221,1 MB	102,79 MB	1,77x
87,2 MB	329,7 MB	124,51 MB	1,43x
197,8 MB	370,9 MB	305,19 MB	1,54x
221,3 MB	414,9 MB	355,58 MB	1,61x
221,6 MB	415,5 MB	334,46 MB	1,51x

A. Performance Tests

The performance tests were conducted to determine time consumption of export. The tests were run on a standard desktop computer (Intel Core i7 at 3,4 GHz, 8 GB of DDR3 RAM, standard HDD with 7200 rpm). An unusually high memory consumption was detected during the test. The size of occupied memory by the EEGExport program for big (220 MB) experimental data reached up to 4 GB of memory. Further testing showed that Java Virtual Machine, in attempt to speed up export, did not free allocated memory. However, when the program was paused, the size of allocated memory became lower. Finally, the disk writing speed was a limited factor (it means that using any other implementation of HDF libraries would not help). The conversion times with GZip compression that was used to reduce file size are shown in Table II. (The GZip compression is integrated in HDF libraries and it is supported natively, reading of the data does not require any special actions.)

B. File Size Tests

Several file size tests were conducted to determine the resulting file size of the HDF5 container storing all data. Data from real experiments were used for these tests. The test showed that the size of the HDF5 container was influenced by exported data and the original file size was not the only decisive factor. The size of the HDF5 container was always bigger than original BrainVision files, four times bigger at the worst case, two times at the best case. The results are shown in Table III and Table IV.

VI. Conclusion

This article described standardization efforts in description of experimental electrophysiology data and metadata for our neuroinformatics lab purposes. We examined current file formats and standardizing efforts dealing with description and storage of electrophysiology data. Many of them use HDF as a data container. Then the most suitable format (the NIX project)

Table IV. FILE SIZE DEPENDENCY ON COMPRESSION CHUNK SIZE.

Original file size	HDF5 file size in MB			
	chunk 64	chunk 256	chunk 512	chunk 1024
21,3 MB	39,76	26,92	23,30	20,46
51,2 MB	155,47	112,98	99,31	87,85
58 MB	151,31	102,79	88,60	77,96
87,2 MB	176,53	124,51	112,43	104,15
197,8 MB	365,63	305,19	292,59	285,48
221,3 MB	422,83	355,58	341,72	333,83
221,6 MB	402,81	334,46	320,22	312,10

was selected and INCF recommendations were taken into account. The local EEGBase data model was mapped to the NIX data model and several terms from the EEGBase metadata set extended (some of them were mapped to) odML metadata terminology. These data and metadata models were fully reflected when the inner structure of the HDF5 container was proposed. Finally, the program HDFExport was designed and developed. This software tool transfers both our experimental data and metadata stored within or outside the EEGBase portal repositories to the HDF5 container. The resulting HDF5 container was tested for its contents and the process of creating HDF5 container was tested for its performance and size of the resulting file.

The proposed EEGBase file format in the HDF5 container is capable of storing a large variety of electrophysiology data and metadata. Exporting experimental data and metadata stored in the EEGBase to the HDF5 container improves sharing capabilities of the EEGBase portal and overall attractiveness of stored experiments.

The software tool HDFExport currently uses EEGBase Portal web services for metadata download, the usage of Portal web services for data download is currently prepared. The next step is to integrate this yet standalone application in the EEGBase Portal.

ACKNOWLEDGEMENTS

This work was supported by the UWB grant SGS-2013-039 Methods and Applications of Bio- and Medical Informatics.

REFERENCES

[1] POLDRACK, R. A. The future of fMRI in cognitive neuroscience. *NeuroImage*. aug 2012, 62, 2, s. 1216–1220. 10.1016/j.neuroimage.2011.08.007. Available from: http://dx.doi.org/10.1016/j.neuroimage.2011.08.007.

[2] MILHAM, M. P. Open Neuroscience Solutions for the Connectome-wide Association Era. *Neuron*. jan 2012, 73, 2, s. 214–218. 10.1016/j.neuron.2011.11.004. Available from: http://dx.doi.org/10.1016/j.neuron.2011.11.004.

[3] POLINE, J.-B. et al. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*. 2012, 6. 10.3389/fninf.2012.00009. Available from: http://dx.doi.org/10.3389/fninf.2012.00009.

[4] FRIEDRICH, S. et al. Mission and activities of the INCF Electrophysiology Data Sharing Task Force. *Frontiers in Neuroinformatics*. 2014, 8. 10.3389/conf.fninf.2014.08.00088. Available from: http://dx.doi.org/10.3389/conf.fninf.2014.08.00088.

[5] *EEGbase* [online]. 2015. [cit. 10.5.2015]. RRID:nif-0000-08190. Available from: https://eegdatabase.kiv.zcu.cz/.

[6] *HDF Java Products* [online]. 2014. [cit. 1.5.2015]. HDF Group. Available from: https://www.hdfgroup.org/products/java/.

[7] *INCF GitHub odML repository* [online]. [cit. 12.5.2015]. Available from: https://github.com/INCF/odml-terminologies.

[8] Requirements for storing electrophysiology data, Version 0.72. Available from: https://goo.gl/QbClkE. Electrophysiology Task Force of the INCF Program on Standards for Data Sharing, 2014.

[9] BENDA, J. From recording to sharing of data - embedding metadata handling into the laboratory workflow using odML. *Frontiers in Neuroinformatics*. 2011, 5. 10.3389/conf.fninf.2011.08.00100. Available from: http://dx.doi.org/10.3389/conf.fninf.2011.08.00100.

[10] Brain Products GmbH. *BrainVision Recorder User Manual*. Brain Products GmbH, 011 edition, December 2014.

[11] *Brain Products GmbH* [online]. 2014. [cit. 03.07.2014]. Brain Products GmbH. Available from: http://www.brainproducts.com/index.php.

[12] BRINKMAN, R. R. et al. Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*. 2010, 1, Suppl 1, s. S7. 10.1186/2041-1480-1-s1-s7. Available from: http://dx.doi.org/10.1186/2041-1480-1-S1-S7.

[13] GREWE, J. – WACHTLER, T. – BENDA, J. A Bottom-up Approach to Data Annotation in Neurophysiology. *Frontiers in Neuroinformatics*. 2011, 5. 10.3389/fninf.2011.00016. Available from: http://dx.doi.org/10.3389/fninf.2011.00016.

[14] *HDF Technologies* [online]. 2013. [cit. 03.07.2014]. The HDF Group. Available from: http://www.hdfgroup.org/.

[15] *INCF* [online]. 2015. [cit. 4.3.2015]. Available from: http://www.incf.org/.

[16] *INCF* [online]. 2015. [cit. 4.3.2015]. Available from: http://www.incf.org/activities/our-programs/datasharing/electrophysiology-task-force/.

[17] KEMP, B. – OLIVAN, J. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. *Clinical Neurophysiology*. sep 2003, 114, 9, s. 1755–1761. 10.1016/s1388-2457(03)00123-8. Available from: http://dx.doi.org/10.1016/S1388-2457(03)00123-8.

[18] *Neo* [online]. 2014. [cit. 03.07.2014]. Available from: http://pythonhosted.org/neo/.

[19] *NeuroHDF* [online]. 2014. [cit. 03.07.2014]. NeuroHDF Interest Group. Available from: https://neurohdf.readthedocs.org/.

[20] *NeXus Format* [online]. 2014. [cit. 03.07.2014]. NeXus International Advisory Committee. Available from: http://www.nexusformat.org/.

[21] *NIX* [online]. 2014. [cit. 03.02.2015]. G-Node. Available from: https://github.com/G-Node/nix/wiki.

[22] *Ovation* [online]. 2014. [cit. 3.7.2014]. Physion LLC. Available from: http://physion.us/.

[23] *The Open Biological and Biomedical Ontologies* [online]. 2015. [cit. 10.5.2015]. Available from: http://www.obofoundry.org/.