Data format for electrophysiological experiments

Jiří Vaněk
Faculty of Applied Sciences, Department of Computer Science and Engineering
University of West Bohemia
Pilsen, Czech republic
Email: vanek2@kiv.zcu.cz

***Abstract*—Currently there is no standardized data format for storing electrophysiological data. This standard is necessary for effective collaboration between scientists. This work deals with adjustments of existing data/metadata model for range of electrophysiological experiments and with proposal/implementation of data format for their storage.**

I. Introduction

Brain research has been very popular recently. Tens of measurements are conducted every year and it is very important to store measured data and metadata for later use. It was common that experiments were designed, measured and analyzed and after evaluation this recorded data was deleted. But during last years it has become more important to store experiments for later use and to share data with other researchers to provide further independent analyses.

It is important to develop an independent standard for storing and exporting experimental data and metadata. If this standard is accepted by a larger community, it will allow easy sharing and better understanding of conducted experiments.

At first we had to get familiar with formats for storing electroencephalography and biomedical data. I explored their model, terminology and ontology. I became acquainted with the International Neuroinformatics Coordinating Facility (INCF) [16] working group and their standard proposal. The storing options of measured experiments at the University of West Bohemia were checked. The formats containing data and stored metadata were analyzed. Then it was necessary to search for available formats for storing EEG data, compare them and choose the best one or develop a new one.

II. State of the Art

A. Data Sharing

A trend toward increased 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 [18] [22].
- **to improve data quality**
  The sharing data helps uncover mistakes as missing data, noise, errors, etc. and improves the quality of the data in the future experiments.
- **to reduce cost of research**
  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 Program on Standards for Data Sharing was established for the purpose of specification the standard for storing EEG data. INCF is an international non-profit organization devoted to advancing the field of neuroinformatics and was established in 2005 in Stockholm. INCF community consists of 17 member countries and associated research groups, consortia, funding agencies and publishers in the field. The National Nodes are institutions or networks that represent each member country. The nodes are established to coordinate neuroinformatics activity within a country [13].

Program Standards for Data Sharing 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.

The current focus of the Program on Standards for Data Sharing is in two areas: neuroimaging and electrophysiology. [13]. 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.

C. Present formats for Storing EEG Data

Most known formats for storing EEG data use the format HDF5. Also both INCF proposals use HDF5 and the Electrophysiology Task Force of the INCF Program on Standards for Data Sharing in Requirements for a standard recommends basing a standard on HDF5. [8] Some formats are proprietary and even though some of them are well documented, is due to licenses complicated to use them or edit them. So I focus on the open ones. For storing EEG neuroinformatics data many types of formats exist. The most known and used formats are Ovation [24], NeXus Format [21], NEO [19], NeuroHDF [20], EDF+ [17] and NIX (Pandora) [23].

1) NIX: This format also uses HDF5 as a data container. This format specification closely defines an inner structure of file, especially the data part. The meta data part is defined by

the odML. The NIX project (previously called Pandora) started in the context of the Electrophysiology Task Force which is part of the INCF Datasharing Program.

NIX is one approach to this problem: it uses highly generic models for data as well as for metadata and defines standard schemata for HDF5 files representing these models. Last but not least NIX aims to provide a convenient C++ library to simplify the access to the defined format. 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. [23]

This format's scheme (Figure 1) was taken as inspiration for EEGBase file format. The measurements are stored in blocks. Each block identifies measurement and related metadata section. Raw data (signals, stimuli) are saved in DataArrays and they are specified by Dimension, Sample, Set, Representation and Range (Figure 1) and could be specified by DataTag. The stimuli and artifacts are stored in SimpleTag (one stimulus) or MultiTag (more stimuli). The source of DataArrays or Tags could be specified by Source. Each section could contain link to the metadata part with measurement information.

2) Brain Vision Format: EEG data at University of West Bohemia are recorded by BrainVision Recorder [11]. This program records raw data and saves it to three files.

The BrainVision Recorder does not allow natively recorded data in any other format. Most recordings consist of three files. The format of these files is defined 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 is Brain Vision Data Exchange Header File Version 1.0 and includes basic information about measurements. The format of the header file is based on the Windows INI format. It consists of various named sections containing keywords/values. The file stores basic information about measuring: coding, name of data file, name of marker file, number of channels, sampling interval in microseconds, information about binary format (IEEE_FLOAT_32) and information about channels (Channel number, channel name, resolution of unit, unit).
- **vmrk file**
  This is Brain Vision Data Exchange Marker File, Version 1.0. The marker file is based upon the same principle of sections and keywords as the header file. This text file contains information about stimuli. The file stores stimuli number, type of the stimuli, description, position, size and channel number.

These files are stored in the EEGBase portal [5] with metadata about an experiment and a measurement.

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 is used by Cardiff University for resolving their problem with grid computing, Deutsche Bank for financial engineering, Diamond Light Source in synchrotron science, Laboratory for Neural Computation for bio-engineering and many others. 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." [15]

HDF is similar to XML documents, HDF files allow to specify complex data relationships and dependencies and are self-describing. Several APIs for programing 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 providing information about stimuli, data acquisition, and experimental conditions etc. are indispensable for the analysis and the management of experimental data within a lab. However, only rarely are metadata available in a structured, comprehensive, and machine-readable form. This poses a severe problem for finding and retrieving data, both in the laboratory and on the various emerging public data bases. [14] The odML defines the format, not the content, so 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 were assembled terminologies with definitions of commonly used terms. [9]

III. Design, Implementation, Testing

A. HDFExport Program

Program HDFExport was developed to conversion of Brain Vision files into EEGBase format.

1) Analysis and Design: The file format is divided into the two autonomous parts DATA and METADATA, which relate to each other, but could be read or written separately. The both parts are stored in one HDF5 container.

The data part stores all recorded binary data and the basic information about measurement for correct representation of measured data (Sampling interval, resolution and resolution unit). This section includes data from the eeg, vhdr and vmrk files. The file structure is defined by the NIX file model - (Figure 1). However our EEGBase model was simplified for our needs. More specific description follows in Section III-A2.

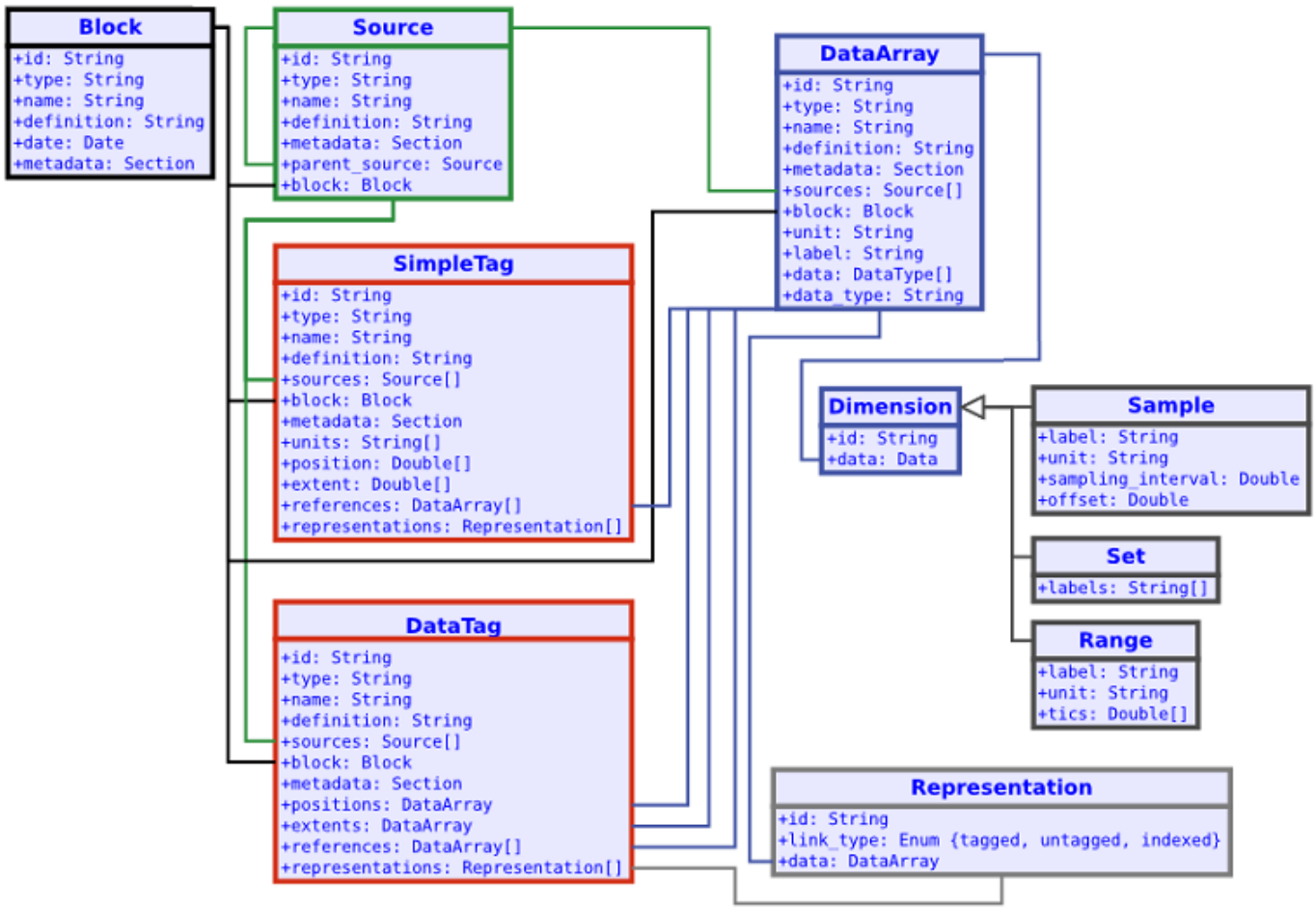

Figure 1. NIX data scheme. [23]

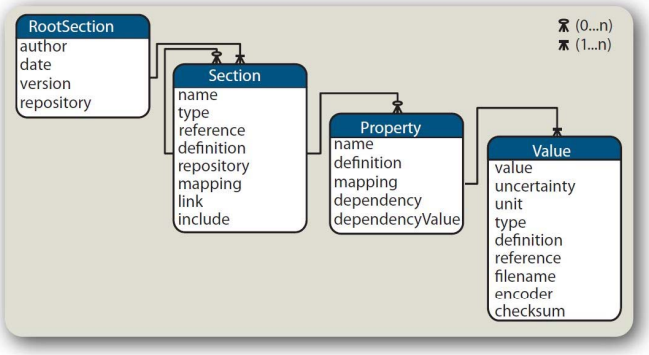

Figure 2. Open metadata Markup Language Entity-Relation diagram. [14]

Java was selected as a programming language for program developing, because there is already a parser of Brain Vision formats developed and also the EEGBase portal is written in Java. There is effort to include the export program into the EEGBase portal for easy export.

2) Data Model: The EEGBase data model is based on the NIX data model (Figure 1 and Section II-C1). The NIX model is able to save data from any electrophysiology experiment. But for EEG experiments the NIX model is too general. So we used only necessary parts of the model and other sections were omitted. The omitted parts are in the NIX model optional, so EEGBase format is compatible with the NIX definition. EEGBase data model is described in Figure 3. The data model uses the NIX scheme of Block, DataArray, MultiTag, DataTag and SimpleTag. The Block is used to divide measuring, DataArray stores raw data of signals and stimuli and MultiTag stores stimuli information and DataTag contains EEG channel information. DataArrays are divided for better distribution in my model to SIGNAL and MARKER parts. Also, the names of DataArrays correspond to names of channels. These adjustments allow better human readability and do not influence information or the model compatibility.

B. Model Ontology

1) Data Part: Ontology and terminology of the data part is based on the NIX model that is described above in Section III-A2 - Data model and in Figure 3.

2) Metadata Part - odML: Metadata are organized according to odML terminology. The G-Node odML scheme and

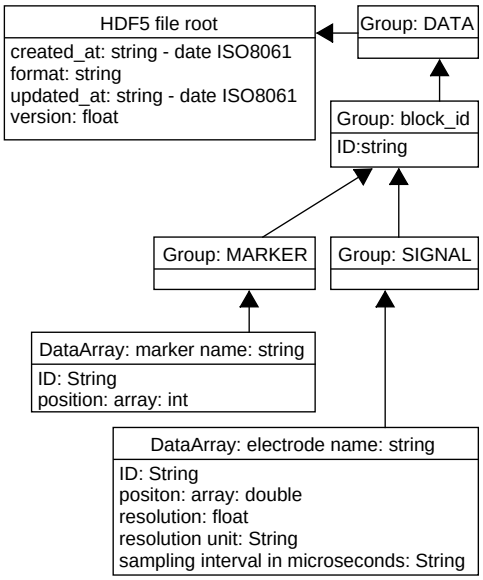

Figure 3. The final data model of proposed data format. The data are stored in tree structure with fixed terminology and structure. Each Group MARKER and SIGNAL could contain zero or more DataArrays with raw data.

terminology was used for the metadata part of file, but there was some more information in the metadata scheme used at UWB, which was difficult to save with existing odML terminology. Therefore the odML model and terminology were extended. The changes and adjustments are described in Section III-C (Metadata Terminology Extensions).

We also decided that it would be useful to store some more data, which are more specific and could help to describe experiments better. The existing ontologies were also searched, for example Ontology for biomedical investigations [25] [12]. However, I decided to use odML because I was able to extend odML to perfectly fit our needs and it is already used by the NIX model.

C. Metadata Terminology Extensions

In order to save all our metadata into the HDF5 container I extended the odML model for our metadata. These modifications were committed to G-Node respective 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 included into odML. All modification are listed in Table I.

D. HDFExport Program

The program is designed for several use cases divided by data and metadata location (Figure 4):

- data and metadata export from locally stored Brain Vision files only
- data and metadata export from locally stored Brain Vision files and metadata from EEGBase portal
- data and basic metadata export from in EEGBase postal stored Brain Vision files without experiments metadata
- data and metadata export from EEGBase portal and Brain Vision files stored in EEGBase portal

Other division is by type of exported data (Figure 5):

- Only raw EEG data and basic metadata are exported
  Only data and metadata from the Brain Vision files are exported. Stimuli are not exported (not all measurements uses stimuli).
- Raw EEG data, basic metadata and stimuli are exported
  All data, metadata and stimuli from Brain Vision files are exported.
- Raw EEG data, basic metadata, stimuli and experiments metadata are exported
  All data, metadata and stimuli from the Brain Vision files and experiments metadata from EEGBase portal are exported.
- Raw EEG data, basic metadata and experiments information are exported
  All data, metadata and experiments information are exported.

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.

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

Figure 4. Use cases of HDFExport Program.

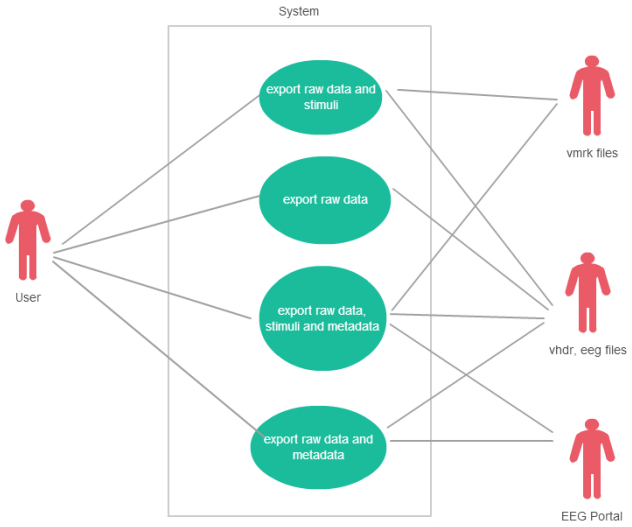

Figure 5. Use cases of HDFExport Program.

IV. Tests

The program was manually tested for several use cases and all created HDF5 files were verified manually (opened and the contents was checked) by official program of HDF Group HDFView [6] in version 2.11. The following test describing writing speed into HDF5 container were conducted. (Table II). The file sizes of created files were also tested. (Table III).

A. Performance Tests

The write performance tests were conducted to determine time consumption of export. The tests were conducted 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 amount of occupied memory by the EEGExport program for big (220 MB) measurements reached up to 4 GB of memory. Further testing showed that Java Virtual Machine, in attempt to speed up export, does not free allocated memory. However, if the program was paused the amount of allocated memory was lower; the program could run with less memory. In the end the disk writing speed was a limited factor. The conversion times with GZip compression are shown in Table II.

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

File size of Brain Vision 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

B. File Size Test

Several test were conducted to determine the resulting file size of the HDF5 container containing all data. Data from real experiments were used for the tests. The test shows that size of the HDF5 container is influenced by exported data and the original file size is not the only decisive factor. The size of the HDF5 container is always bigger then Brain Vision files, at worst case even four times bigger in the best scenario about two times bigger. The results are shown in Table III. The GZip compression was used to reduce file size. The GZip compression is integrated in HDF libraries and it is supported natively (reading of the data does not require any special actions).

Table III. File size tests.

File size of Brain Vision 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

V. Conclusion

We examined current file formats for storing electrophysiology data and data from experiments and measurements conducted at the University of West Bohemia. We became familiar with the data and metadata model of EEG measurements and its terminology. We examined the data and metadata model of EEGBase portal.

We analyzed two early file standard proposals from INCF and I tracked progress in both. We found several currently used

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

formats, which are using HDF as a container in neuroinformatics. We chose the most suitable format for our data and usage considering INCF recommendations.

We created implementation of the chosen format. We chose HDF5 container for the EEGBase format. We joined the INCF Electrophysiology Data Sharing Task Force and contributed to the odML terminology. We developed a program that transforms raw data and metadata from Brain Vision files to the EEGBase format, and We also included metadata which are stored in the EEGBase portal. We tested my format and program for several use cases and its performance.

The proposed EEGBase format is capable of storing all currently saved data and metadata and is able to store the future changes and modifications of metadata model. The developed program stores measured data into the EEGBase format. I also made a few suggestions for the UWB model. The program is currently using web services of the EEGBase portal for metadata loading. The developed libraries allow export of raw data or data with metadata. Exporting experimental data and metadata in the EEGBase format to the HDF5 container improves sharing capabilities of the EEGBase portal and overall attractiveness of stored experiments.

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] 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.

[14] 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.

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

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

[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] 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.

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

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

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

[22] 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.

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

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

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