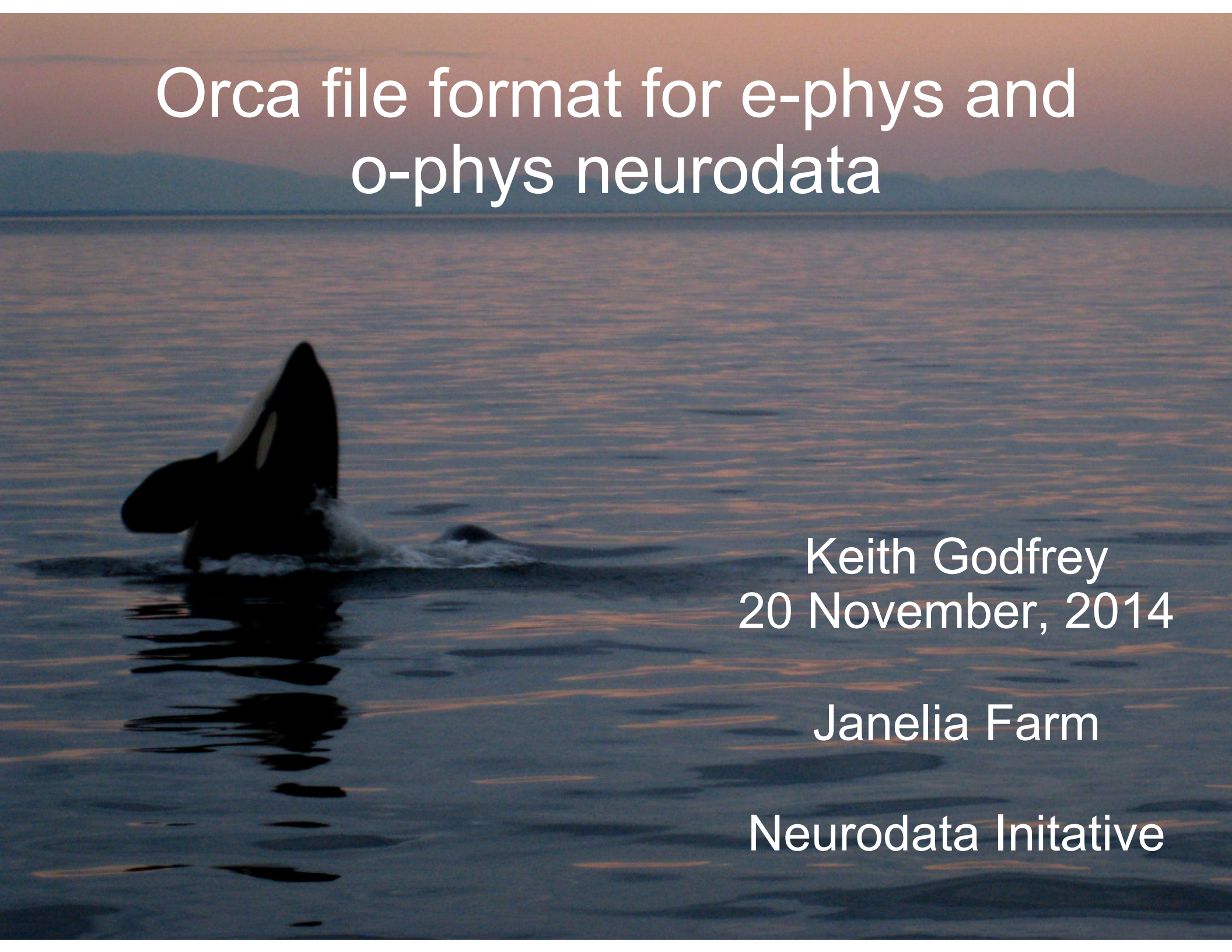

Orca file format for e-phys and o-phys neurodata

Keith Godfrey
20 November, 2014

Janelia Farm

Neurodata Initative

Orca design goals

Represent Neurodata and Allen Institute data sets, including intra- and extra-cellular e-phys; 2-photon o-phys; electrical and visual stimulus; and video tracking

Use a defined schema to facilitate software tool support

Represent data in a way that can be quickly understood and/or explained

Provide easy access to data through Matlab, python and compiled languages

Use object-oriented design, for extensibility plus backward compatibility

Orca file structure – top level

Group/dataset	Description
orca_version	File version string. E.g., “Orca-0.2.10”
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Data sequences

The orca format is designed around data structure called a 'Sequence'. A sequence is a superset of many INCF types.

Dataset / attribute	Description
description (attribute)	Verbal description of what sequence represents
ancestry (attribute)	Text array with subclass hierarchy, including self
data	Polymorphic field for storing data
si_unit (attribute)	Base SI unit (e.g., volt, amps, etc)
conversion (attribute)	Scaling factor to convert values in data to specified SI unit
filter (attribute)	Description of filtering that's been applied to data
resolution (attribute)	Minimum data resolution (e.g., bits per volt)
sampling_rate	Samples per second, approximately
num_samples	Number of samples
timestamps	Array of timestamps (seconds since experiment began)
t_interval	'N', or number of samples between each timestamp

Example of sequence subclassing

Group/dataset	Description
orca_version	File version string. E.g., "Orca-0.2.10"
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Acquisition

Acquired data, including ephys, ophys, and tracking, plus graphical documentation of experiment (e.g., photos of surgery, histology, etc).

Datasets and subgroups	Type	Description	Comments
sequences	folder	Acquired sequences	When importing acquisition data to neurodata file, all acquisition/tracking/stimulus data must already be aligned to a common time frame. It is assumed that this task has already been performed.
acquisition_sequence_x	generic sequence		Name is arbitrary
images	folder		
image_x	binary	Photograph of experiment or experimental setup (video also OK)	Name is arbitrary
format (attr)	text	format of image	eg, jpg, png, mpeg
description (attr)	text	human description of image	

Group/dataset	Description
orca_version	File version string. E.g., "Orca-0.2.10"
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Stimulus

Presented stimuli and stimulus templates. Templates can be re-used and/or stored in a remote file.

Datasets and subgroups	Type	Description	Comments
templates	folder	Template stimuli	Time stamps in templates are based on stimulus design and are relative to the beginning of the stimulus. When templates are used, the stimulus instances must convert presentation times to the experiment's time reference frame.
template_stimulus_x	generic sequence		Name is arbitrary
epochs (attr)	text array	Epochs where this template was used	names in array correspond to those in /epochs
presentation	folder	Stimuli presented during the experiment	
stimulus_sequence_x	generic sequence		Name is arbitrary

Group/dataset	Description
orca_version	File version string. E.g., "Orca-0.2.10"
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Epochs

Experimental intervals, with links to and windows into all acquisition and stimulus sequences that are a part of the interval.

Datasets and subgroups	Type	Description	Comments
epoch_x		Different experimental epoch	Name is arbitrary but must be unique within the experiment.
description	text		
start_time	double		
stop_time	double		
ignore_intervals	2D double array		
acquisition_series_x		input stream recorded during epoch	Name is arbitrary and does not have to match the sequence that it refers to
t_start	double	Starting time of sequence in epoch	Start and stop time are technically redundant, as this information is available within the series, but computing it from an arbitrary point in the time[] and discontinuity[] arrays is not necessarily fast
t_stop	double	Ending time of sequence in epoch	
idx_start	int	Epoch's start index in sequences data[] field	
idx_stop	int	Epoch's stop index in sequences data[] field	
sequence		link to sequence under /acquisition	
stimulus_series_x		Output stream presented during epoch	Name is arbitrary and does not have to match the sequence that it refers to
t_start	double	Starting time of sequence in epoch	Start and stop time are technically redundant, as this

Group/dataset	Description
orca_version	File version string. E.g., "Orca-0.2.10"
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

General

Experiment metadata (example format)

Datasets and subgroups	Type	Description	Comments
session_id	text		Only 1 session_id per file, with all time aligned to experiment start time.
animal	folder		
id		ID of animal used in experiment	
species		Species of recording	
genotype		Genetic strain	If absent, assume WT. Otherwise indicate genes knocked-in/-out.
age		Age of animal	
area		Recording site(s)	
experiment			
lab			
notes			
protocol			
devices	folder		Eg, monitors, ADC boards, microscopes, etc
NEC234WM	folder	Example device	
description	text	eg, "24-inch NEC LCD monitor"	
serial_number	text		
resolution	text	eg, dpi for x and y	
distance	text	distance of monitor from animal	
ephys		Meta-data for ephys modality, if any ephys modules are present	See metadata description in the Ephys module description (above).
2photon-ophys		meta-information for 2-photon ophys	

Group/dataset	Description
orca_version	File version string. E.g., "Orca-0.2.10"
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Processing

Processing modules (e.g., like Kwik). Modules are objected-oriented, like Sequences. Each stores trail back to pre-processed data.

Datasets and subgroups	Type	Description	Comments
EventDetection	folder	Detects events from raw ephys data	Name is arbitrary. Eg, there may be parallel ephys modules named "events_shank_1", "events_shank_2", etc. Provenance information is stored as part of module.
FeatureExtractor	folder	Extracts features from identified events	Name is arbitrary (eg, "features_shank_1")
SpikeClustering	folder	Automatic clustering module (e.g., Kwik)	Name is arbitrary (eg, "clusters_shank_1")
ClusterCutting	folder	Manual clustering or manual "cleaning" of automatic clustering	Name is arbitrary
2photon-ophys	folder	Module for ophys data providing estimated spike/event times and/or processed image stack	Name is arbitrary
position-tracking	folder	Position tracking module	
eye-tracking	folder	Eye-tracking module	

Group/dataset	Description
orca_version	File version string. E.g., “Orca-0.2.10”
identifier	Unique string to identify file
file_create_date	Date + time (ISO format)
session_start_time	Date + time (ISO format) of experiment
acquisition	Data streams recorded from the system
stimulus	Data streams pushed into the system
epochs	Experimental intervals and/or sub-experiments
general	Metadata, including protocols, notes and devices
processing	Modules for intermediate processing
analysis	Lab-specific and scientific analyses

Analysis

Lab-specific and scientific analysis. The content of the folder is free-form, but use INCF-based data structures is encouraged.

Evaluation criteria

Uses specific schema to facilitate software support

Data is organized so it's descriptive and easily accessible

Supports python, MATLAB and compiled languages using standard open-source HDF5 libraries

Designed to store intra- and extra-cellular ephys plus 2-photon and intrinsic imaging ophys.

Issues: Functionally meets Initiative requirements, but can use refinement, particularly regarding metadata

Thank you

Anton Arkhipov
Jim Berg
Tim Blanche
Saskia de Vries
Severine Durand

Aleena Garner
Nathan Gouwens
Ken Harris
Chris Lau
Kenji Mizuseki

Josh Siegle
Karel Svoboda
Jeff Teeters
Barry Wark
Rob Young