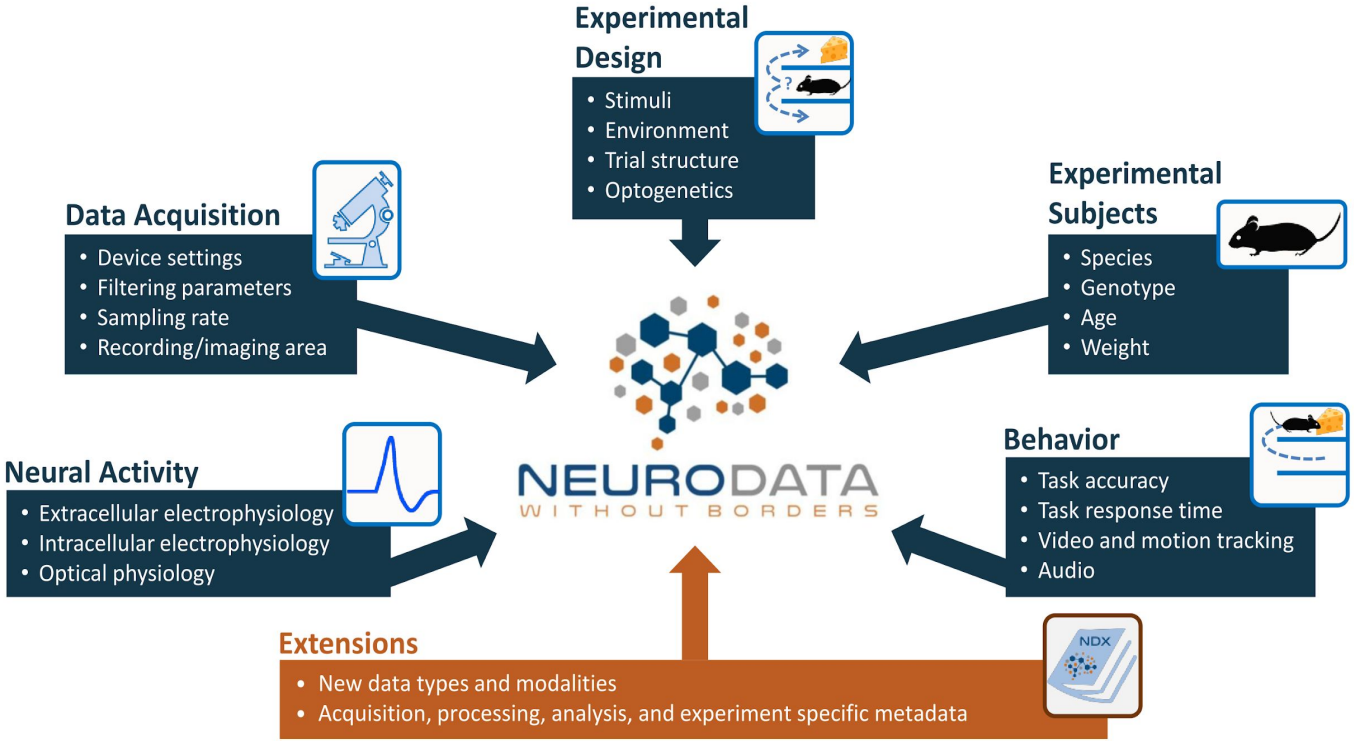
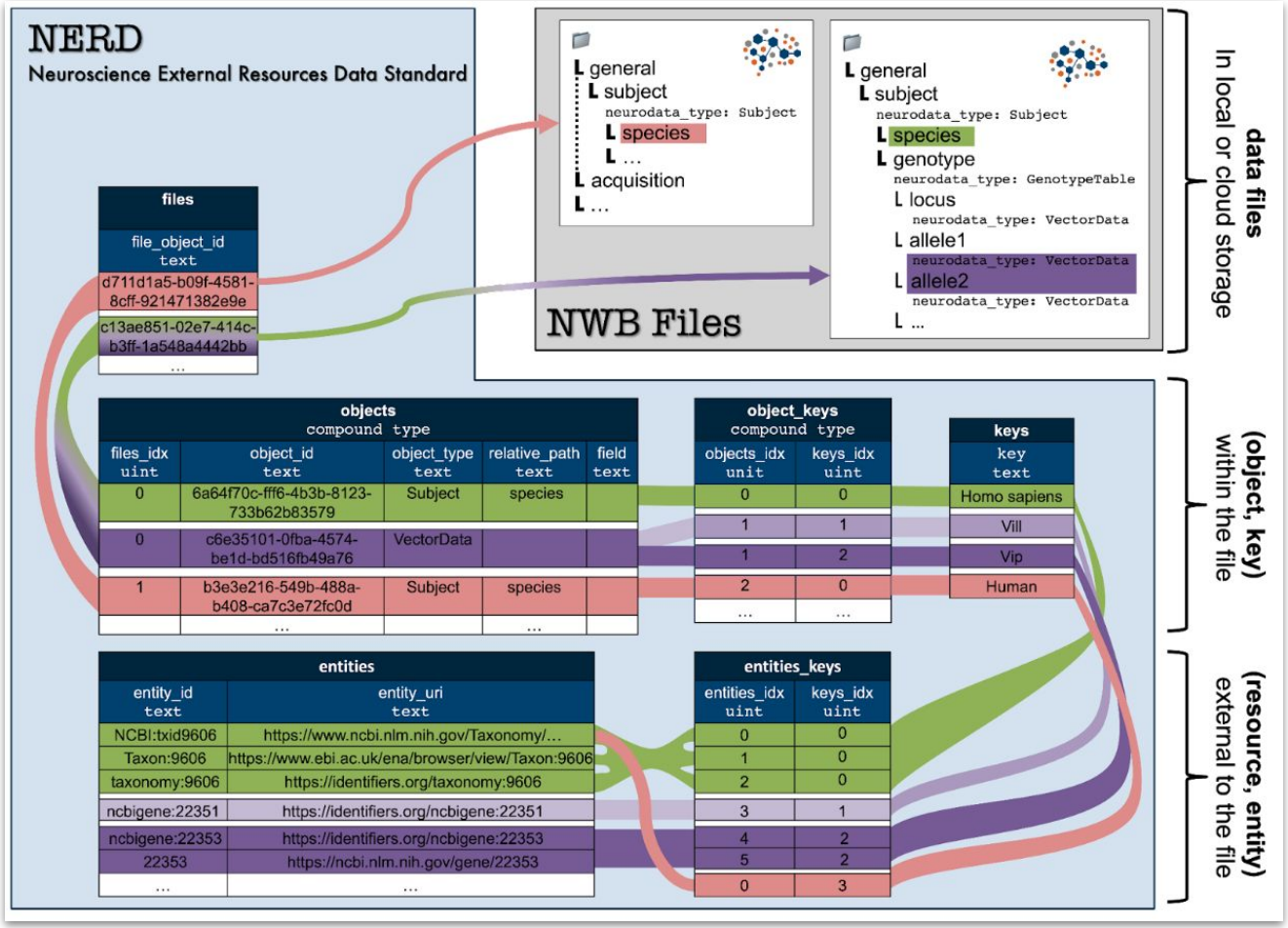
The Neuroscience External Resources Data Standard

Matthew Avaylon, Ryan Ly, Oliver Rübel

2023 NWB Developer Days
July 27, 2023

An Ecosystem for Neuroscience Data Standardization




NERD
Neuroscience External Resources Data Standard
NWB Files
general
subject
neurodata_type: Subject
species
…
acquisition
…
general
subject
neurodata_type: Subject
species
genotype
neurodata_type: GenotypeTable
locus
neurodata_type: VectorData
allele1
neurodata_type: VectorData
allele2
neurodata_type: VectorData
…
data files
In local or cloud storage
files
file_object_id
text
d711d1a5-b09f-4581-8cff-921471382e9e
c13ae851-02e7-414c-b3ff-1a548a4442bb
…
objects
compound type
files_idx uint
object_id text
object_type text
relative_path text
field text
0
6a64f70c-fff6-4b3b-8123-733b62b83579
Subject
species
0
c6e35101-0fba-4574-be1d-bd516fb49a76
VectorData
1
b3e3e216-549b-488a-b408-ca7c3e72fc0d
Subject
species
object_keys
compound type
objects_idx unit
keys_idx uint
0
0
1
1
1
2
2
0
keys
key text
Homo sapiens
Vill
Vip
Human
(object, key)
within the file
entities
entity_id text
entity_uri text
NCBI:txid9606
https://www.ncbi.nlm.nih.gov/Taxonomy/…
Taxon:9606
https://www.ebi.ac.uk/ena/browser/view/Taxon:9606
taxonomy:9606
https://identifiers.org/taxonomy:9606
ncbigene:22351
https://identifiers.org/ncbigene:22351
ncbigene:22353
https://identifiers.org/ncbigene:22353
22353
https://ncbi.nlm.nih.gov/gene/22353
entities_keys
entities_idx uint
keys_idx uint
0
0
1
0
2
0
3
1
4
2
5
2
0
3
(resource, entity)
external to the file

NERD Example

root (NWBFile)
session_description: Data from monkey Haydn performing ready-set-go time interval reproduction task. This file contains continuous segments of the full session on 2016-12-11 that can be used for training models for the Neural Latents Benchmark.
identifier: 8969f328-3929-11ec-8077-43176b153428
session_start_time: 2016-12-11 00:00:00-05:00
timestamps_reference_time: 2016-12-11 00:00:00-05:00
▶ **file_create_date**
experimenter: ('Hansem Sohn',)
related_publications: ('http://dx.doi.org/10.1016/j.neuron.2019.06.012',)
▶ **keywords**
epoch_tags: set()
▶ **electrodes**
▼ **electrode_groups (3)**
  ▼ **electrode_group_1**
    description: Electrodes on a neural probe
    location: Dorsomedial frontal cortex
    ▶ **device**
  ▶ **electrode_group_2**
  ▶ **electrode_group_3**
▶ **devices (3)**
▶ **intervals (1)**
▼ **subject**
  age: P4Y
  sex: M
  species: Macaca mulatta
  subject_id: Haydn

Using add_ref

Create and Link ER

```
er = ExternalResources()
read_nwbfile.link_resources(er)
```

NWBFile Experimenter

```
er.add_ref(
    container=read_nwbfile,
    attribute="experimenter",
    key="Hansem Sohn",
    entity_id='ORCID:0000-0001-8593-7473',
    entity_uri='https://orcid.org/0000-0001-8593-7473')
```

Electrode_Group Location

```
er.add_ref(
    container=read_nwbfile.electrode_groups['electrode_group_1'],
    attribute="location",
    key="Dorsomedial frontal cortex",
    entity_id="DB09",
    entity_uri="https://scalablebrainatlas.incf.org/macaque/DB09")
```

Subject Species

```
er.add_ref(
    container=read_nwbfile.subject,
    attribute='species',
    key='Macaca mulatta',
    entity_id='NCBI_TAXON:9544',
    entity_uri='https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/id=9544')
```

NERD Visualized

	file_object_id	objects_idx	object_id	files_idx	object_type	relative_path	field	keys_idx	key	entities_idx	entity_id	
0	9c3a5c45-316c-493d-a712-03a01b662ee9	0	9c3a5c45-316c-493d-a712-03a01b662ee9	0	NWBFile	general/experimenter		0	Hansem Sohn	0	ORCID:0000-0001-8593-7473	
1	9c3a5c45-316c-493d-a712-03a01b662ee9	1	f8641805-f93c-446f-8194-5fce08d22dbb	0	ElectrodeGroup	location		1	Dorsomedial frontal cortex	1	DB09	https://s
2	9c3a5c45-316c-493d-a712-03a01b662ee9	2	5ee39486-8625-4ac3-9691-ce9d724812a4	0	Subject	species		2	Macaca mulatta	2	NCBI_TAXON:9544	https://ww

TermSet

- **Validation of Data**
  - Currently supports only data sets.
- **Streamlines the NERD user experience**
  - Reduces the number of required fields in the add_ref method.

```
id: notebooks/species_example
name: Experimenter
prefixes:
  ORCID: https://orcid.org/
imports:
  - linkml:types
default_range: string

enums:
  Experimenters:
    permissible_values:
      Dichter, Benjamin K.:
        description: The ORCiD
        meaning: ORCID:0000-0001-5725-6910
      Rubel, Oliver:
        description: The ORCiD
        meaning: ORCID:0000-0001-9902-1984
```

NERD TermSet Example

```
terms = TermSet(term_schema_path='./experimenter_term_set.yaml')
er = ExternalResources()
```

```
session_start_time = datetime(2018, 4, 25, 2, 30, 3, tzinfo=tz.gettz("US/Pacific"))

nwbfile = NWBFile(
    session_description="Mouse exploring an open field",
    identifier="Mouse5_Day3",
    session_start_time=session_start_time,
    session_id="session_1234",
    experimenter=["Dichter, Benjamin K.", "Rubel, Oliver"],
    lab="My Lab Name",
    institution="University of My Institution",
    related_publications="DOI:10.1016/j.neuron.2016.12.011",
)
nwbfile.subject = Subject(
    subject_id="001",
    age="P90D",
    description="mouse 5",
    species="Mus musculus",
    sex="M",
)
```

Using add_ref_term_set

add_ref

```
er.add_ref(
    container=read_nwbfile,
    attribute="experimenter",
    key="Hansem Sohn",
    entity_id='ORCID:0000-0001-8593-7473',
    entity_uri='https://orcid.org/0000-0001-8593-7473')
```

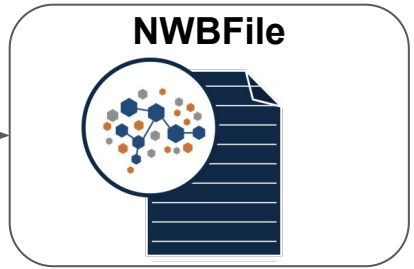
add_ref_term_set

```
er.add_ref_term_set(container=nwbfile,
                    attribute='experimenter',
                    term_set=terms
                   )
```

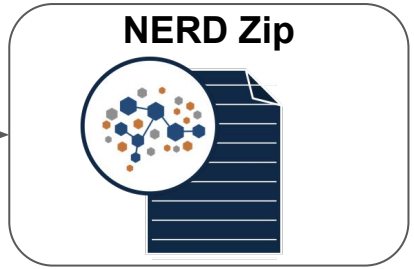
Write NWBFile and NERD

- The NWBFile and NERD are written separately.
  - NERD is written as a zip file containing the individual tables in the data structure.

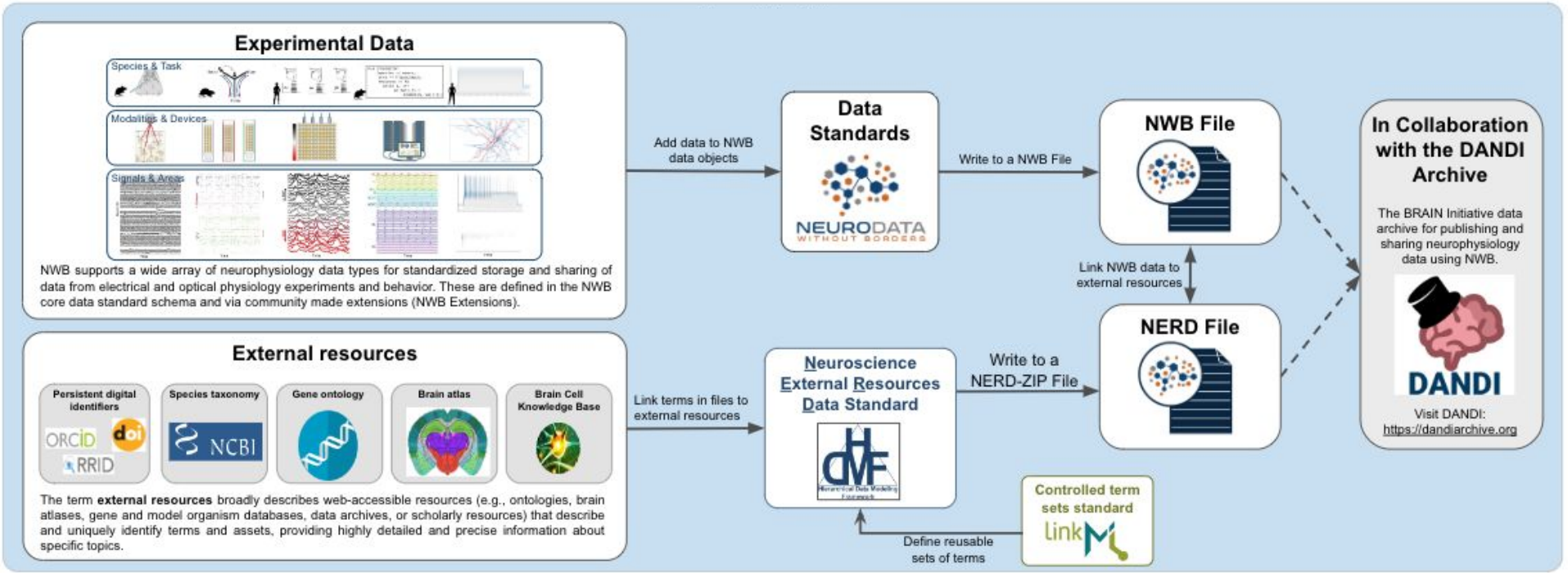
```
1 with NWBHDF5IO("NWBfile_ER_Example.nwb", "w") as io:
2     io.write(nwbfile)
```




```
1 er.to_norm_tsv(path='./')
```

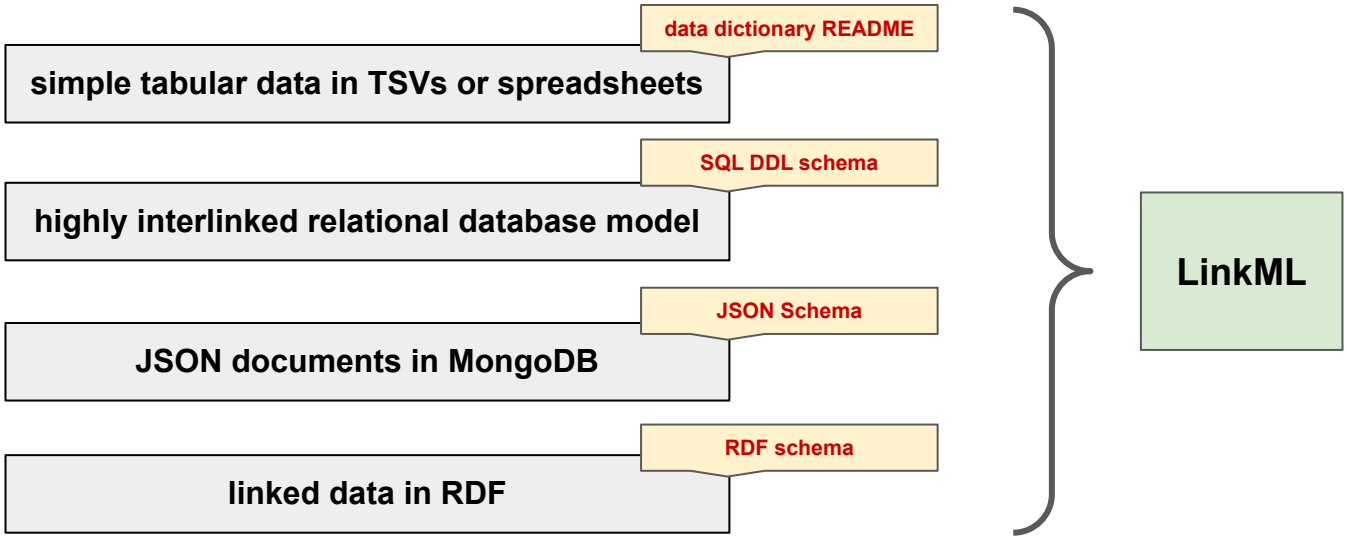


User Workflow



Introduction to LinkML

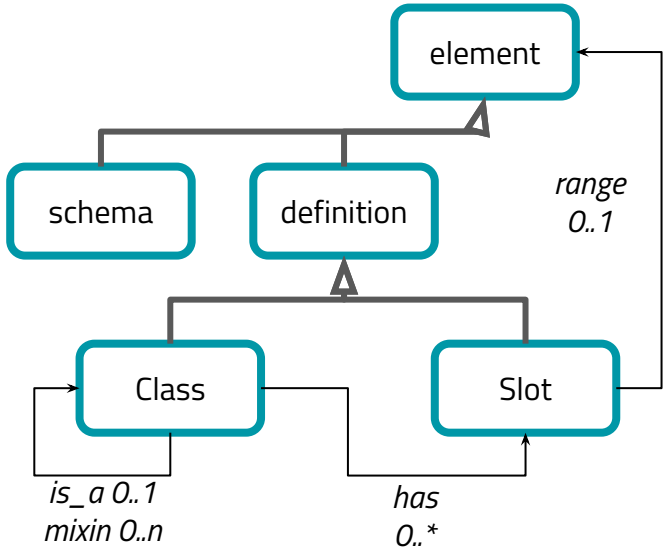
- All data follows some kind of schema / data model (“blueprint”)
- LinkML is a flexible modeling language that allows you to author schemas (“data models”) in YAML that describe the structure of your data



Introduction continued

THE STANDARD

*A **meta-standard** for structuring your data*



TOOLS

Pragmatic developer and curator friendly tools for working with data

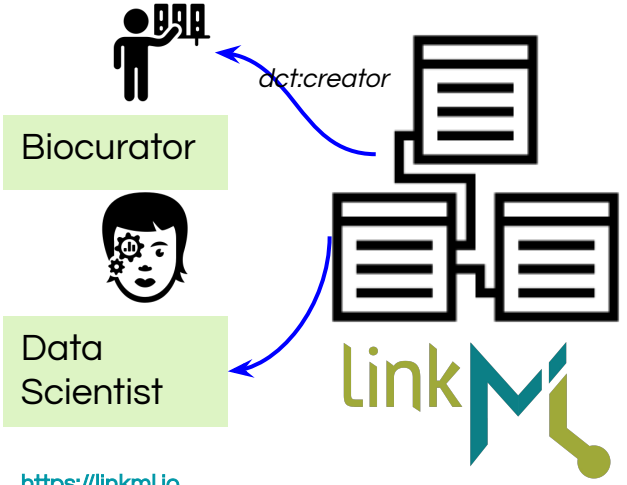
- **Validators**
- **Data Converters**
- **Compatibility tools**
- **Data entry**
- **Schema inference**

The LinkML landscape

Create data models / standards in simple YAML files, optionally annotated using ontologies

Compile to other frameworks

Choose the right tools for the job, no lock in



https://linkml.io
https://github.com/linkml/linkml

OWL
ShEx, SHACL
JSON-LD Contexts
JSON-Schema
Python Dataclasses
SQL DDL
TSVs

Semantic Web Applications And Infrastructure

RDF

JSON-LD

"Traditional" Applications and Infrastructure

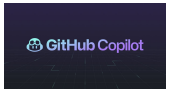
pandas

JSON

Authoring schemas in YAML

Option A: Author YAML *directly*

Optional productivity tools




```yaml
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
version: 0.0.1

prefixes:
 linkml: https://w3id.org/linkml/
 sdo: https://schema.org/
 ex: https://example.org/linkml/hello-world/

default_prefix: ex
default_curi_maps:
 - semweb_context

imports:
 - linkml:types

classes:
 Person:
   description: Minimal information about a person
   class_uri: sdo:Person
   attributes:
     id:
       identifier: true
       slot_uri: sdo:taxID
     first_name:
       required: true
       slot_uri: sdo:givenName
       multivalued: true
     last_name:
       required: true
       slot_uri: sdo:familyName
     knows:
       range: Person
       multivalued: true
       slot_uri: foaf:knows
```

Metadata

Namespaces

Dependencies

Actual data model

YAML conformant to LinkML standard

Schemasheets

Option B: Author using *schemasheets*

a person,living or dead

record	field	key	multiplicity	range	parents	desc	schema.org	wikidata	be
> class	slot	identifier	cardinality	range	is_a	description	exact_mappings	exact_mappings	in_
>								curie_prefix: wikida	
	id	yes	1	string		any identifier	identifier		
	description	no	0..1	string		a textual description	description		
Person		n/a	n/a	n/a		a person,living or dead	Person	Q215627	
Person	id	yes	1	string		identifier for a person	identifier		
Person|Organiza	name	no	1	string		full name	name		
Person	age	no	0..1	decimal		age in years			
Person	gender	no	0..1	decimal		age in years			
Person	has medical hist	no	0..*	MedicalEvent		medical history			
Event						grouping class for events		Q1656682	a
MedicalEvent		n/a	n/a	n/a	Event	a medical encounter			b
ForProfit					Organization				
NonProfit					Organization			Q163740	

SCHEMASHEETS
easy LinkML from spreadsheets

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
version: 0.0.1

prefixes:
 linkml: https://w3id.org/linkml/
 sdo: https://schema.org/
 ex: https://example.org/linkml/hello-world/

default_prefix: ex
default_curi_maps:
```

```
: a person
```

```
multivalued: true
slot_uri: foaf:knows
```

Metadata

Namespaces

Dependencies

Actual Datamodel

YAML conformant to LinkML standard

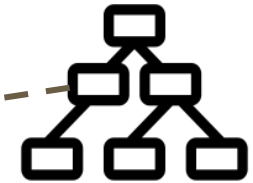
NEURODATA WITHOUT BORDERS

Enumerations in LinkML allow ontology mapping

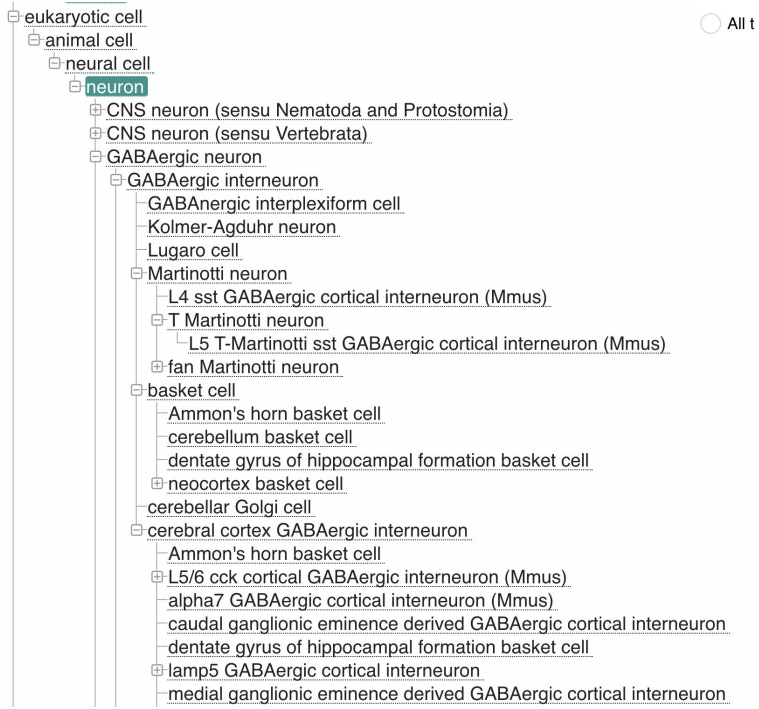
```yaml
prefixes:
  COB: http://purl.obolibrary.org/obo/COB_
  BFO: http://purl.obolibrary.org/obo/BFO_
  RO: http://purl.obolibrary.org/obo/RO_
  CHEBI: http://purl.obolibrary.org/obo/CHEBI_
  CHEMINF: http://semanticscience.org/resource/CHEMINF_
  SIO: http://semanticscience.org/resource/SIO_
  PUBCHEM.ELEMENT: https://pubchem.ncbi.nlm.nih.gov/element/
  LANL.ELEMENT: https://periodic.lanl.gov/
```

```yaml
enums:

  nanostructure_morphology_enum:
    permissible_values:
      nanotube:
        meaning: CHEBI:50796
      nanoparticle:
        meaning: CHEBI:50803
      nanorod:
        meaning: CHEBI:50805
      nanotubosome:
        meaning: CHEBI:50806
      quantum dot:
        meaning: CHEBI:50853
      nanofibre:
        meaning: CHEBI:52518
      nanocrystal:
        meaning: CHEBI:52529
      nanoribbon:
        meaning: CHEBI:52530
      nanosheet:
        meaning: CHEBI:52531
      nanowire:
        meaning: CHEBI:52593
```



Dynamic Enumerations in LinkML




```
enums:
  NeuronTypeEnum:
    reachable_from:
      source_ontology: obo:cl
      source_nodes:
        - CL:0000540 ## neuron
      include_self: false
      relationship_types:
        - rdfs:subClassOf
```

Topics to Start Discussion

1. Any questions in general?
2. Changes Just Around the Corner:
    a. HDMF_Zarr supporting the most up-to-data NERD tools (Next Release)
    b. Customize NERD Zip File name (Next Release)
3. Community Feedback Topics:
    a. Version Control on NERD
    b. NERD and NWBFile “write”