Towards interoperability of neurophysiology data repositories

Jeffrey L Teeters[1], Petr Ježek[2] Christian Garbers[3] Christian J Kellner[3] Michael Sonntag[3]
Achilleas Koutsou[3] Jan Grewe[4] Friedrich T Sommer[1] Thomas Wachtler[3]

1. Redwood Center for Theoretical Neuroscience, UC Berkeley, USA
2. University of West Bohemia, Czechia
3. German Neuroinformatics Node, Ludwig-Maximilians-Universität München, Martinsried, Germany
4. Institute for Neurobiology, Eberhard-Karls-Universität Tübingen, Tübingen, Germany

Summary

Metadata stored in different forms (DataCite XML, NWB:Neurophysiology, NIX/odML) in the CRCNS.org and G-Node GIN data repositories can be converted to a standard format (RDF) which allows searching for and linking data across repositories.

1) DataCite Metadata

Both CRCNS.org and the G-Node GIN data repositories use DataCite to create DOIs for published datasets. The metadata used to create the DOIs can be retrieved in XML format, converted to JSON, then converted to RDF using the semantic framework software [1].

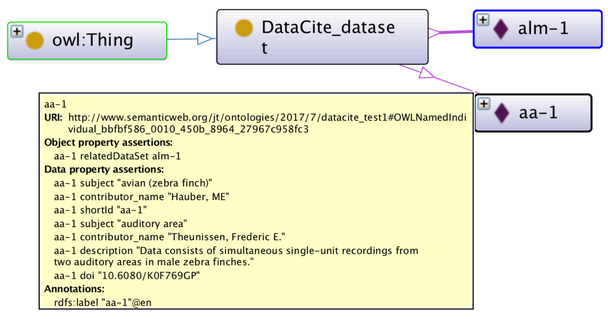

Fig. 1. Protégé OntoGraf representation of DataCite dataset object properties that were converted to RDF.

```
http://search.datacite.org/api?
q=datacentre_symbol:cdl.ucbcrcns&fl=doi,minted,updated,xml&fq=has_metadata:true&fq=is_active:true&rows=1000&start=0&sort=updated+asc&wt=json
```

Fig 2. Query to download Metadata from DataCite

```
{
    "aa-2": {
        "alt_title": "aa-2",
        "creators": [
            "Theunissen, Frederic E.",
            "Gill, Patrick",
            "Fremouw, Thane"
        ],
        "description": "This data set contains about 500 single...",
        "doi": "10.6080/K0JW8BSC",
        "minted": "2013-06-25T23:37:09Z",
        "publicationYear": "2011",
        "subjects": [
            "Neuroscience",
            "Electrophysiology",
            "auditory area",
            "avian (zebra finch)"
        ],
        "title": "Single-unit recordings from zebra finches",
        "updated": "2016-06-17T20:26:23Z"
    },
    "hc-2": {
    }
}
```

Fig 3. Sample downloaded metadata with XML converted to JSON

```
  <rdf:Description rdf:about="http://
cz.zcu.kiv.data.crcns.json#Aa2___26215382">
    <j.0:altTitle rdf:datatype="&xsd;string">aa-2</j.0:altTitle>
    <j.0:updated rdf:datatype="&xsd;string">2016-06-17T20:26:23Z</j.
0:updated>
    <j.0:description rdf:datatype="&xsd;string">This data set
contains about 500 single.</j.0:description>
    <j.0:minted rdf:datatype="&xsd;string">2013-06-25T23:37:09Z</j.
0:minted>
    <j.0:subjects rdf:nodeID="A2"/>
    <j.0:title rdf:datatype="&xsd;string">Single-unit recordings
from zebra finches</j.0:title>
    <rdf:type rdf:resource="http://
cz.zcu.kiv.data.crcns.json#Aa2__"/>
    <j.0:doi rdf:datatype="&xsd;string">10.6080/K0JW8BSC</j.0:doi>
    <j.0:publicationYear rdf:datatype="&xsd;string">2011</j.
0:publicationYear>
    <j.0:creators rdf:nodeID="A3"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A3">
    <rdf:_3 rdf:datatype="&xsd;string">Fremouw, Thane</rdf:_6>
    <rdf:_2 rdf:datatype="&xsd;string">Gill, Patrick</rdf:_2>
    <rdf:_1 rdf:datatype="&xsd;string">Theunissen, Frederic E.</
rdf:_1>
    <rdf:type rdf:resource="&rdf;Seq"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A2">
    <rdf:_4 rdf:datatype="&xsd;string">avian (zebra finch)</rdf:_4>
    <rdf:_3 rdf:datatype="&xsd;string">auditory area</rdf:_3>
    <rdf:_2 rdf:datatype="&xsd;string">Electrophysiology</rdf:_2>
    <rdf:_1 rdf:datatype="&xsd;string">Neuroscience</rdf:_1>
    <rdf:type rdf:resource="&rdf;Seq"/>
  </rdf:Description>
```

Fig 4. Metadata converted to RDF using Semantic Framework

a)

```
$ more query_small2.rq
# get datasets with subject avian*

PREFIX j.0: <http://cz.zcu.kiv.data.crcns.json#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#>

select  ?alt_title ?subject {
    ?object j.0:altTitle ?alt_title .
    ?object j.0:subjects ?subNode .
    ?subNode rdfs:member ?subject . FILTER regex (?subject,
"avian*")
    }
```

b)

```
sparql --data=ontologyDocument.rdf.xml --query query_small2.rq
-------------------------------------
| alt_title | subject               |
=====================================
| "aa-3"    | "avian (zebra finch)" |
| "am-3"    | "avian (zebra finch)" |
| "am-2"    | "avian (zebra finch)" |
| "am-1"    | "avian (zebra finch)" |
| "aa-2"    | "avian (zebra finch)" |
| "aa-1"    | "avian (zebra finch)" |
-------------------------------------
```

Fig 5. Query to search converted DataCite metadata. a) SPARQL query. b) command to search using Apache Jena.

NWB:Neurophysiology format metadata can also be converted to RDF and searched.

2) Metadata storage and organization using odM

Metadata in electrophysiology are highly diverse and complex. What metadata experimenter may need to store cannot be predicted in advance without knowledge of the specific experiment. Therefore, odML [2] provides a format for metadata but does not constrain the content, thus achieving both human-readability and machine-readability while enabling to store any information needed to annotate a given dataset [3]. Since the organization of odML metadata reflects the structure of the experiment, such annotation is straightforward by linking between metadata entries and data records, as in the NIX format for neuroscience data [4].

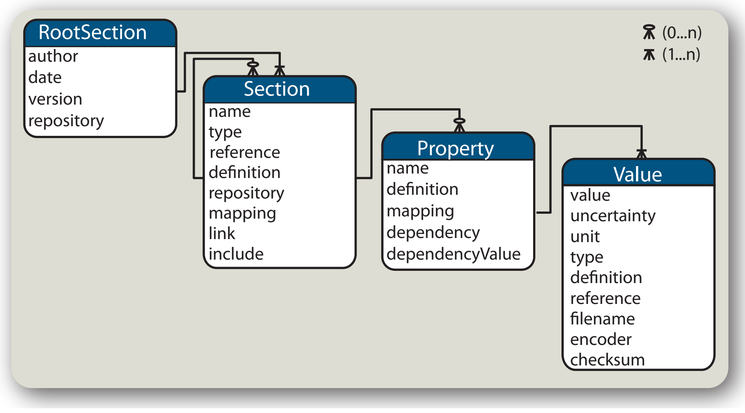

Fig 6. odML data model

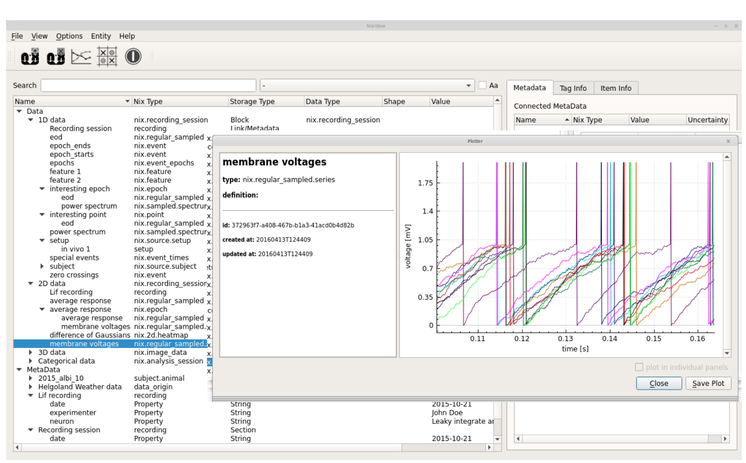

Fig. 7. Example NIX file with odML metadata and data linked together. Figure shows a screenshot of NixView, an advanced HDF5 viewer that makes use of NIX features for data exploration [5].

3) Metadata storage and indexing using document oriented databases

Managing collections of diverse experimental metadata likewise requires approaches that do not impose constraints on the content. We therefore evaluated the performance of odML for metadata search and integration with document-oriented databases and modern indexing technologies. We built different storage backends for odML that, due to schemaless designs, can easily be adapted to various formats.

Fig. 8. Schemas of the odML model in mongoDB and CouchDB (left) and Lucene (right).

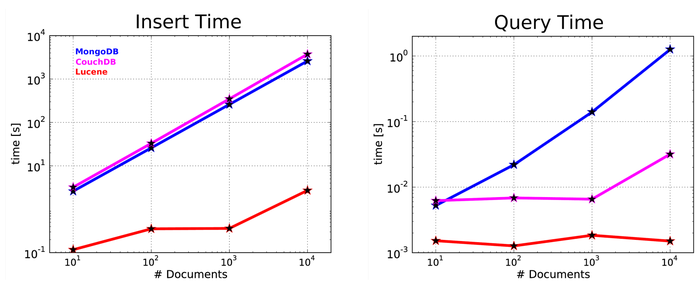

Fig. 9. Benchmarking results for odML metadata. Plots show insert and query times, respectively, as functions of number of documents for different databases. Insert times include rebuilding the index. Benchmarking was done on a DualQuad Core AMD Opteron™ Processor 2220 2.8Ghz 16GB used exclusively during testing. Inserting is costly if indexes are optimized. Lucene's advanced search features provide high performance, outperforming mongoDB and couchDB.

4) Conversion of odML to RDF

Semantic web technologies enable linking and searching datasets across the web. Conversion of odML to RDF offers a straightforward approach to interoperable databases.

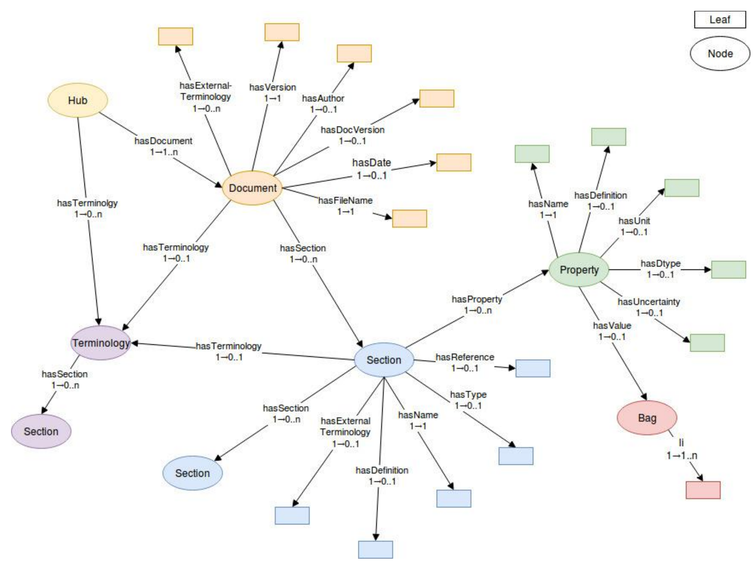

Fig. 10. Schema for mapping odML to RDF. odML sections and properties and their relations are expressed using RDF.

```
<odml:1532eef2-5a93-431a-859e-a40b4713e8ae> a odml:Section ;
    odml:hasName "Stimulus" ;
    odml:hasProperty <odml:04ba2361-64e3-44b5-9d93-aac911036b15>,
        <odml:2085f3e1-d177-4d71-854b-420950d6de86>,
        <odml:35661f50-e360-4bae-93ec-1b20fc28dd38>,
        <odml:571631fc-a20e-4509-8c86-8f6c595627c7>,
        <odml:86b47ea9-6bbf-4c0c-befd-19f1e80ff8ad>,
        odml:d9363a44-0560-414c-9a80-2546b87f533b ;
    odml:hasType "stimulus/white_noise" .

<odml:2085f3e1-d177-4d71-854b-420950d6de86> a odml:Property ;
    odml:hasDtype "string" ;
    odml:hasName "UpperCutoffFrequency" ;
    odml:hasUnit "Hz" ;
    odml:hasValue <odml:6558abef-8cba-4361-aa90-bf5494aee2f2> .

<odml:6558abef-8cba-4361-aa90-bf5494aee2f2> a rdf:Bag ;
    rdf:li "300.0" .

<odml:571631fc-a20e-4509-8c86-8f6c595627c7> a odml:Property ;
    odml:hasDtype "string" ;
    odml:hasName "Contrast" ;
    odml:hasUnit "%" ;
    odml:hasValue odml:b74a164e-a081-4b1b-b1e5-06ae8c500bcc .

odml:b74a164e-a081-4b1b-b1e5-06ae8c500bcc a rdf:Bag ;
    rdf:li "20.0" .

<odml:86b47ea9-6bbf-4c0c-befd-19f1e80ff8ad> a odml:Property ;
    odml:hasDtype "string" ;
    odml:hasName "Duration" ;
    odml:hasUnit "s" ;
    odml:hasValue <odml:5009b3e0-f4ef-4895-aaa4-858d647db00c> .

<odml:5009b3e0-f4ef-4895-aaa4-858d647db00c> a rdf:Bag ;
    rdf:li "10.0" .

<odml:35661f50-e360-4bae-93ec-1b20fc28dd38> a odml:Property ;
    odml:hasDtype "string" ;
    odml:hasName "SampleRate" ;
    odml:hasUnit "ms" ;
```

Fig 11. Example metadata in Turtle syntax. Representation of metadata in RDF enables linking data across repositories via the semantic web and thus offers a way to achieve interoperability between data repositories.

Conclusions

- DataCite metadata may be converted to a common form (RDF) that can be searched across repositories, thus achieving a first level of interoperability.
- odML is a flexible and extensible format for neuroscience metadata
- Searching and versioning of extensive metadata collections is efficient using odML with document-oriented databases or indexing.
- Exposing odML to RDF offers linking and searching of data across distributed repositories.

References

[1] Ježek P, Mouček R. (2015). Semantic framework for mapping object-oriented model to semantic web languages. Front. Neuroinform. doi: 10.3389/fninf.2015.00003
[2] Grewe J, Wachtler T, Benda J (2011). A bottom-up approach to data annotation in neurophysiology. Front. Neuroinform. 5:16. doi: 10.3389/fninf.2011.00016
[3] Zehl L, Jaillet F, Stoewer A, Grewe J, Sobolev A, Wachtler T, Brochier G, Riehle A, Denker M, Grün S (2016) Handling Metadata in a Neurophysiology Laboratory. Frontiers in Neuroinformatics 10:26 doi: 10.3389/fninf.2016.00026
[4] http://www.g-node.org/nix
[5] https://github.com/bendalab/NixView
[6] https://github.com/G-Node/python-odml/tree/dev-odml-rdf

Acknowledgements

NSF grant 1516527, BMBF grants 01GQ1302 and 01GQ1509.