CONNJUR_ML: An XML Schema for NMR Reconstruction Workflows

Douglas Heintz[1], Michael R. Gryk[1,2]

[1]Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL; [2]Molecular Biology and Biophysics, UCONN Health, Farmington, CT

Overview

The workflow for modern biomolecular NMR spectroscopy consists of three phases: spectral reconstruction, the process of converting time domain data into the frequency domain; spectral analysis, which includes peak identification and resonance assignment; and biophysical characterization, which includes all subsequent data analysis in which the spectroscopic data is used to draw biophysical inferences (such as structure determination). In this poster we describe an XML schema for describing structural, descriptive and administrative metadata for representing the intermediate datasets generated during spectral reconstruction. As such, this XML schema provides a provenance record of the spectral reconstruction, an essential step in supporting reproducible computation.

PREMIS 3.0 Model

Maintained by the Library of Congress, ***Pre****servation* ***M****etadata:* ***I****mplementation* ***S****trategies* has been developed in the archives and library communities since 2003 to provide digital preservation systems with a framework to build reliable systems for sustainable information stewardship.

Finalized in 2015, version 3.0 of the PREMIS model extended PREMIS' reach beyond providing administrative and preservation metadata for digital objects to providing for robust, object-oriented, description of physical objects, intellectual entities, and most importantly for this project, computational environmental dependencies.

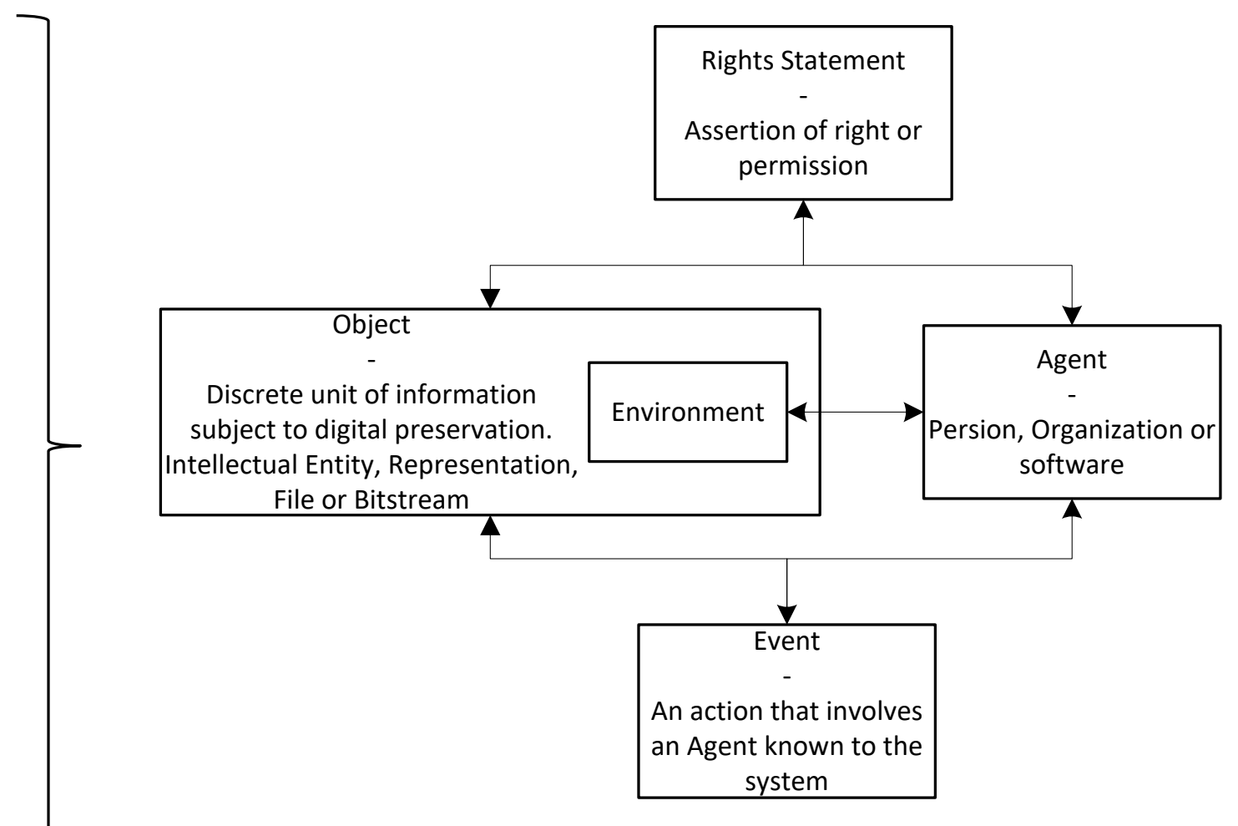
The PREMIS model is encoded as an XML namespace with some RDF semantics.

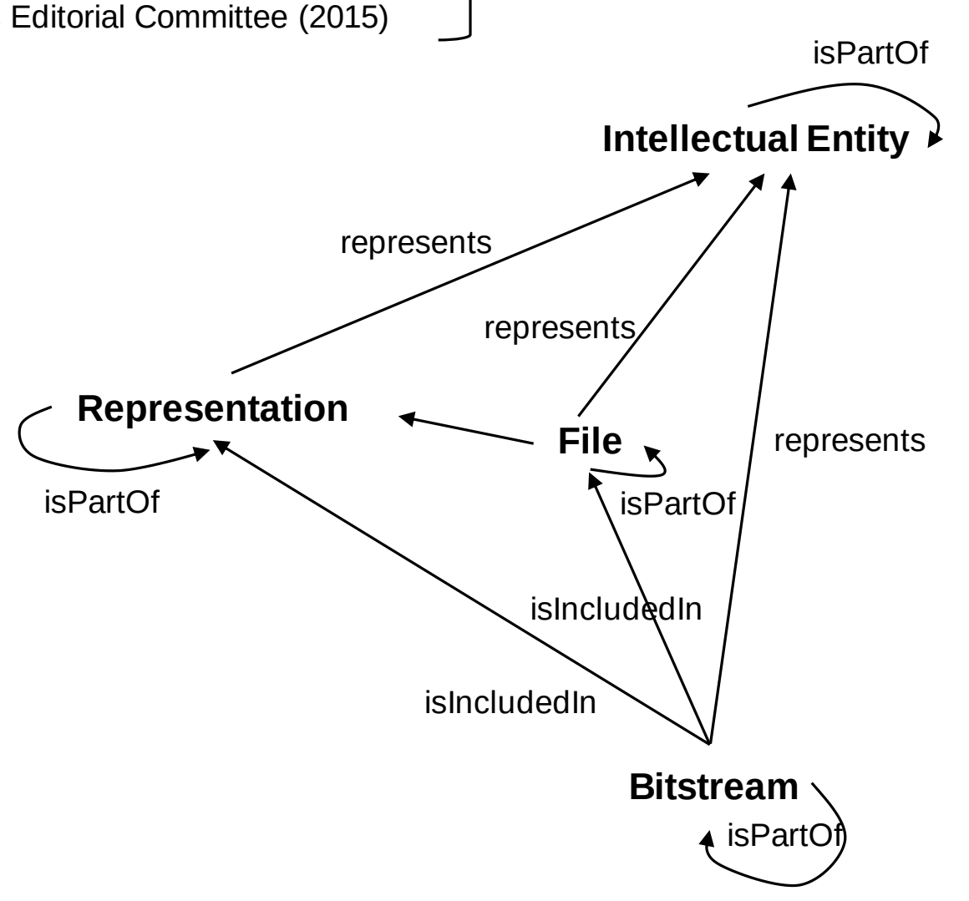
PREMIS Framework

Semantic Map of PREMIS Objects

Each data-entity is made of several discrete, interrelated objects. CONNJUR_ML focuses on Objects, Events, and Agents. Rights Statements could be used to trigger data access restrictions or data deletion.

Diagram Source:
Premis Editorial Committee (2015)


Rights Statement
-
Assertion of right or permission
Object
-
Discrete unit of information subject to digital preservation. Intellectual Entity, Representation, File or Bitstream
Environment
Agent
-
Persion, Organization or software
Event
-
An action that involves an Agent known to the system



Conceptual Map of Object Categories

Objects are modeled in interrelated categories based on the bibliographic FRBR conceptual model (Work, Expression, Manifestation, Item).

Diagram Source:
Premis Editorial Committee (2015)

CONNJUR_ML PREMIS Extension

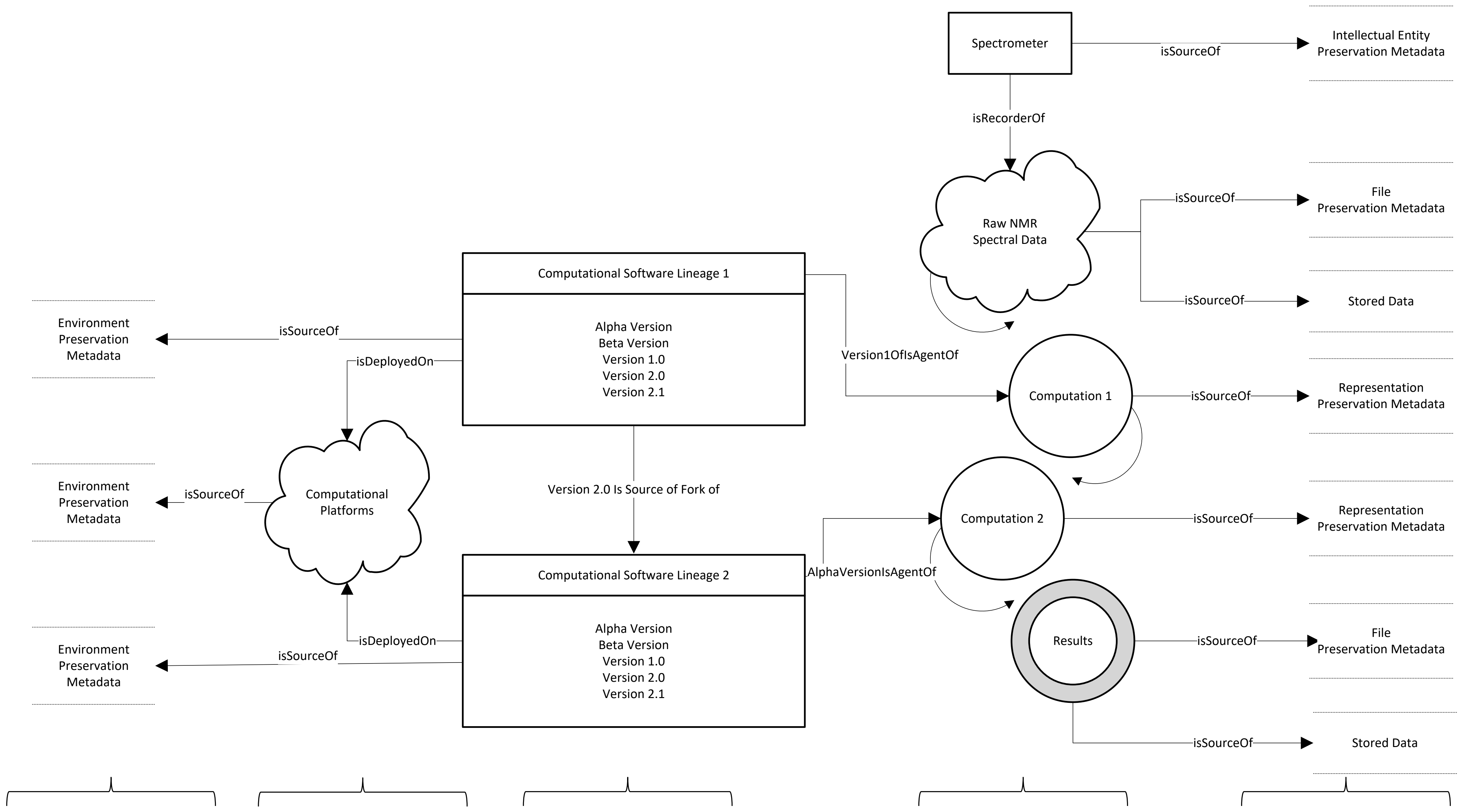
Computational Software Lineages

On ingest into the NMRbox virtualization tool, metadata on software tools are automatically logged as PREMIS objects connect to the NMRbox.org software registry. Related software is encapsulated in classes with group provenance and dependencies that can be accessed from the workflow system.

Self-Referencing Representations as Metadata Surrogates for Ephemeral Pipeline Data

The CONNJUR_ML namespace is embedded in PREMIS *significantPropertiesExtension* when describing workflow metadata for which no intermediate data are stored. Here, metadata transform into the primary data objects that can enhance reproducibility.

Model of a CONNJUR_ML Workflow with Integrated Metadata Capture



Environment Objects
- PREMIS XML in CONNJUR_ML Classes
- Linked to software registry on NMRbox.org

Environments
- NMR Box Backend
- Can include hardware
- Deep Dependencies of specific NMRbox Tools

Software
- NMR Box hosted applications and scripts
- Class representation is specific to preservation domain

Workflow
- PREMIS events
- Computational steps leading to final results

Workflow (Meta)data
- PREMIS XML with CONNJUR_ML Extensions
- References Environment Objects

Spectral Metadata Embedded in a PREMIS Object with CONNJUR_ML Extensions

```
<?xml version="1.0" encoding="UTF-8"?>
<premis:premis xmlns:premis="http://www.loc.gov/premis/v3"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:connjur="https://raw.githubusercontent.com/CONNJUR/CONNJUR_ML/master/connjur_ml.xsd"
    xsi:schemaLocation="http://www.loc.gov/premis/v3
                        http://www.loc.gov/standards/premis/premis.xsd" version="3.0">
    <premis:object xsi:type="premis:representation">
        <premis:objectIdentifier>
            <premis:objectIdentifierType>local</premis:objectIdentifierType>
            <premis:objectIdentifierValue>NHSQC_001</premis:objectIdentifierValue>
        </premis:objectIdentifier>
        <premis:significantProperties>
            <premis:significantPropertiesType>format</premis:significantPropertiesType>
            <premis:significantPropertiesValue>nmrPipe</premis:significantPropertiesValue>
            <premis:significantPropertiesExtension>
                <connjur:spectrum>  <!-- migration -->
                    <connjur:spectrumIdentifier>NHSQC_001</connjur:spectrumIdentifier>
                    <connjur:spectralFormat>
                        <connjur:formatName>nmrPipe</connjur:formatName>
                          <connjur:formatSpecification>
                          https://www.ibbr.umd.edu/nmrpipe/install.html
                          </connjur:formatSpecification>
                        <connjur:formatByteOrder>BigEndian</connjur:formatByteOrder>
                        <connjur:formatNumericType>IEEE 754-2008 binary32
                          </connjur:formatNumericType>
                    </connjur:spectralFormat>
                    <connjur:spectralDimensions>2</connjur:spectralDimensions>
                    <connjur:spectralLayout/>
                    <connjur:spectralAxis>
                        <connjur:sweepwidth units="Hz">6012</connjur:sweepwidth>
                        <connjur:pointType>complex</connjur:pointType>
                        <connjur:spacingUnits units="us">166.7</connjur:spacingUnits>
                        <connjur:totalPoints>1024</connjur:totalPoints>
                        <connjur:sign>=</connjur:sign>
                        <connjur:phaseShift>none</connjur:phaseShift>
                        <connjur:encoding>echo-anti-echo</connjur:encoding>
                        <connjur:decoupledNucleus>15N</connjur:decoupledNucleus>
                        <connjur:constantPhase>0.0</connjur:constantPhase>
                        <connjur:linearPhase>0.0</connjur:linearPhase>
                        <connjur:domain>time</connjur:domain>
                        <connjur:sampling>uniform</connjur:sampling>
                    </connjur:spectralAxis>
                    <connjur:spectralAxis>
                        <connjur:sweepwidth units="Hz">2200</connjur:sweepwidth>
                        <connjur:pointType>complex</connjur:pointType>
                        <connjur:spacingUnits units="us">454.5</connjur:spacingUnits>
                        <connjur:totalPoints>256</connjur:totalPoints>
                        <connjur:sign>+</connjur:sign>
                        <connjur:phaseShift>none</connjur:phaseShift>
                        <connjur:decoupledNucleus>1H</connjur:decoupledNucleus>
                        <connjur:constantPhase>0.0</connjur:constantPhase>
                        <connjur:linearPhase>0.0</connjur:linearPhase>
                        <connjur:domain>time</connjur:domain>
                        <connjur:sampling>uniform</connjur:sampling>
                    </connjur:spectralAxis>
                </connjur:spectrum>
            </premis:significantPropertiesExtension>
        </premis:significantProperties>
    </premis:object>
    <!-- … Additional objects, events, agents and rights …  -->
```

Conclusions and Future Work

Implementation Strengths

- Extends 2015 NMR Star Metadata work by Fenwick *et al.*
- Integrates best-practices from information management into scientific workflow
- Creates human readable metadata in a standard archival format with little customization
- XML/RDF allows for mapping and workflow reconstruction with network diagrams

Solicit Feedback

- Understand user needs and potential uses through open dialog with Bio-NMR Community

Future Work

- Middleware automatically generates environment metadata as part of NMRbox registry
- Integration of CONNJUR_ML into existing workflow processes
- Automatic visualization of RDF relationships and network analysis
- Integrate with NMR Star

School of
Information Sciences
The iSchool at Illinois

UCONN
HEALTH

Acknowledgments

This work is partially supported by National Institutes of Health (NIH) / National Institute of General Medical Sciences (NIGMS), grant 1P41GM111135 through the NMRbox project.

Dr. Rhiannon Bettivia at the Illinois iSchool provided expert guidance as we developed this schema.

Repository Access

https://github.com/CONNJUR/CONNJUR_ML

Contact

Douglas Heintz - dheintz@Illinois.edu
Michael R. Gryk – gryk@neuron.uchc.edu

References

Denenberg R (ed) (2014) PREMIS: Preservation Metadata XML Schema version 3.0. Library of Congress, Washington, DC

Fenwick M, Hoch JC, Ulrich E, Gryk MR (2015) CONNJUR R: an annotation strategy for fostering reproducibility in bio-NMR—protein spectral assignment. *Journal of Biomolecular NMR* 63:141–150 . doi: 10.1007/s10858-015-9964-1

PREMIS Editorial Committee (2015) PREMIS Data Dictionary for Preservation Metadata version 3.0. Library of Congress, Washington, DC