Curating Scientific Workflows for Biomolecular Nuclear Magnetic Resonance (NMR) Spectroscopy

February 20, 2018

13th International Digital Curation Conference

Douglas Heintz & Michael R. Gryk
Protein Structure Determination by NMR Spectroscopy
BioNMR Computation Pipeline

1. Data Collection
2. Transform to Frequency
3. Identify Signals
4. Biophysical Claim

Many software tools available for phases 2-4
BioNMR Computation Pipeline

1. Data Collection
2. Transform to Frequency
3. Identify Signals
4. Biophysical Claim

Many software tools available for phases 2-4
#! /bin/csh

```
nmrPipe -in 22k_3l_nhsqct_000320.pipe \
| nmrPipe -fn SOL -mode 1 -fl 16 -fs 1 -poly \n| nmrPipe -fn CBF -last 12 \n| nmrPipe -fn GMB -lb -7 -gb 0.1 -size 512 -c 0.5 \n| nmrPipe -fn ZF -size 2048 \n| nmrPipe -fn FT -verb \n| nmrPipe -fn PS -p0 68.6 -p1 -34.8 -di \n| nmrPipe -fn EXT -left -sw -verb \n| nmrPipe -fn TP \n| nmrPipe -fn LP -fb -ord 30 -x1 2 -xn 128 -pred 64 -fix -fixMode 1 -after \n| nmrPipe -fn SP -off 0.39 -end 0.98 -pow 2 -size 192 -c 0.5 \n| nmrPipe -fn ZF -size 256 \n| nmrPipe -fn FT -verb \n| nmrPipe -fn PS -p0 -9.0 -p1 20.0 -di \n-out 22k_3l_nhsqct_000320.ft2 -ov
```

Linux shell scripts automate command line operations. Similar to .bat files in Windows.
Many NMR spectroscopists claim the existence of this script guarantees the computation is **reproducible**.
Workflow Management Systems
CONNJUR Workflow Builder (CWB)

- Graphical design of workflow
- Manages execution
- Handles file translation
- Allows for forked workflows
- Interfaces with relational DB
CWB Exports Workflows as XML (sharing)
PROBLEM: Our XML is a custom representation of our workflow. We’ve solved one problem - tacit knowledge embedded inside shell scripts - by creating a similar problem. Tacit knowledge embedded in our XML workflow representations.
Preservation Metadata: Implementation Strategies: (PREMIS)

Figure 1: The PREMIS Data Model
Preservation Metadata: XML example

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xml version="1.0" encoding="UTF-8"
<premis:preservation xml:stylesheet href="http://www.loc.gov/premis/v3" type="premis:file">
  <premis:object identifierType="local">
    <premis:objectIdentifier>
      <premis:objectIdentifierType value="001"/>
    </premis:objectIdentifier>
    <premis:preservationLevelType value="logical preservation"/>
  </premis:preservationLevelType>
  <premis:preservationLevelRole value="intention"/>
  <premis:preservationLevelValue value="2015-02-23"/>
  <premis:significantProperties/>
    <premis:significantPropertiesType value="character count"/>
    <premis:significantPropertiesValue value="4063"/>
    <premis:significantPropertiesType value="word count"/>
    <premis:significantPropertiesValue value="711"/>
    <premis:significantPropertiesType value="line count"/>
    <premis:significantPropertiesValue value="33"/>
    <premis:significantPropertiesType value="paragraph count"/>
    <premis:significantPropertiesValue value="9"/>
  </premis:significantPropertiesType>
  <premis:objectCharacteristics/>
    <premis:resize/>
    <premis:format/>
      <premis:formatDesignation/>
      <premis:formatRegistry/>
        <premis:formatRegistrationName/>
        <premis:formatKey/>
      </premis:formatRegistrationRole>
    </premis:format>
</premis:preservationLevelType>
</premis:object>
</xml>
```
Preservation Metadata: XML example

```xml
<xml version="1.0" encoding="UTF-8">  
    <premis:object xml:type="premis:file">  
      <premis:objectIdentifier>  
        <premis:objectIdentifierType>local</premis:objectIdentifierType>  
        <premis:objectIdentifierValue>001</premis:objectIdentifierValue>  
      </premis:objectIdentifier>  
      <premis:preservation.evenType>logical preservation</premis:preservation.evenType>  
      <premis:preservation.evenValue> emulation </premis:preservation.evenValue>  
      <premis:preservation.evenRole valueURI="http://id.loc.gov/vocabulary/preservation/preservationLevelRole/int"  
        authorityURI="http://id.loc.gov/vocabulary/preservation/preservationLevelRole" authority="preservationLevelRole">intention</premis:preservation.evenRole>  
      <premis:preservation.evenRationale>institutional policy</premis:preservation.evenRationale>  
      <premis:preservation.evenDateAssigned>2015-02-23</premis:preservation.evenDateAssigned>  
    </premis:object>  
    <premis:significantPropertiesType>character count</premis:significantPropertiesType>  
    <premis:significantPropertiesValue>4063</premis:significantPropertiesValue>  
    <premis:significantPropertiesType>word count</premis:significantPropertiesType>  
    <premis:significantPropertiesValue>712</premis:significantPropertiesValue>  
    <premis:significantPropertiesType>line count</premis:significantPropertiesType>  
    <premis:significantPropertiesValue>33</premis:significantPropertiesValue>  
    <premis:significantPropertiesType>paragraph count</premis:significantPropertiesType>  
    <premis:significantPropertiesValue>9</premis:significantPropertiesValue>  
    <premis:objectCharacteristics>  
      <premis:objectFormatCharacteristics>  
        <premis:object:format:designtation name="WordPerfect for DOS">  
          <premis:object:format:version numeric="3.1"<premis:object:format:version>  
        <premis:object:format:registration name="PRONOM">  
          <premis:object:format:registration:key numeric="394"<premis:object:format:registration:key>  
        <premis:object:format:registration:role valueURI="http://id.loc.gov/vocabulary/preservation/formatRegistrationRole"  
          authorityURI="http://id.loc.gov/vocabulary/preservation/formatRegistrationRole" authority="formatRegistrationRole">specification</premis:object:format:registration:role>  
      </premis:objectCharacteristics>  
  </premis:premis>  
</xml>
```

XML Chunks:
- Intellectual Object
- Event
- Agent
- Rights Statement
Preservation Metadata: XML example

XML Chunks:
• Intellectual Object
  • In this metadata section, we can describe important features of our datasets. These features define what operations have been done to them, what can be done to them, and basic structural features of the dataset. Surrogate for intermediate dataset.
• Event
  • In this metadata section, we can record the workflow information. The stream of ‘actors’ or processing events to which the data was subjected
• Agent
  • This metadata section gives us the ability to record important information about the software tools used along with any library dependencies.
• Rights Statement
CONNJUR XML schema on GitHub

CONNJUR ML - NMR Spectral Workflow Markup Language

Summary: This project capitalizes on recent advances in the FREMS data model to create an object-oriented markup framework for recording complex computational processes regarding NMR spectra.

Why base our workflow system on FREMS? NMR spectroscopy experiments begin, run, and when data are written to disk by scientific instruments. In a typical computational experiment, the data are often then transformed and processed by tens of agents and software environments. FREMS object-oriented maps that the repository will be able to handle agents and environments systematically. For us, this is great for both our goal of recording a procedural workflow, useful software versioning, the provenance of the software and virtual machines the software. In a typical deployment, a client would
CONNJUR XML schema on GitHub
CONNJUR XML schema inside CWB
The hope is that now our workflows are understandable to a broader audience.
CWB inside NMRbox

Similar to the BitCurator Environment, NMRbox is a Virtual Machine (VM) provisioned with NMR software. Users can download the VM and run it on their local computer, or access cloud computing services at nmrbox.org
The goal of the VM is to provide stable instances of the bioNMR software - removing one obstacle to reproducibility. Hopefully, broader and richer description of processing workflows will address another obstacle.
PREMIS at the BMRB?

The major repository for bioNMR data is the BioMagResBank housed at the University of Wisconsin-Madison. The BMRB is partnered with NMRbox and is in the process of expanding their NMR-STAR data dictionary to include the workflow additions provided by CONNJUR.
NMRbox Container File (.nbx)

Similar to .docx or other xml based files, we have created a container file which stores the raw, binary data expected by the software tools, metadata parameters required by the software tools, as well as the PREMIS record and optionally an NMR-STAR representation for the BMRB.

```
<table>
<thead>
<tr>
<th>BINARY</th>
<th>Parameters</th>
<th>PREMIS</th>
<th>NMR-STAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Acknowledgments

- Special thanks to Dr. Rhiannon Bettivia, School of Information Sciences, University of Illinois at Urbana-Champaign
- NMRbox Staff - Farmington, CT, Madison, WI
- BioMagResBank Staff - Madison, WI
- This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number GM-111135.

- Thank you!
- Questions?
- Advice?