IJDC | *General Article*

Curating Scientific Workflows for Biomolecular Nuclear Magnetic Resonance Spectroscopy

Douglas Heintz
University of Illinois, Urbana-Champaign

Michael R. Gryk
University of Illinois, Urbana-Champaign

Abstract

This paper describes our recent and ongoing efforts to enhance the curation of scientific workflows to improve reproducibility and reusability of biomolecular nuclear magnetic resonance (bioNMR) data. Our efforts have focused on both developing a workflow management system, called CONNJUR Workflow Builder (CWB), as well as refactoring our workflow data model to make use of the PREMIS model for digital preservation. This revised workflow management system will be available through the NMRbox cloud-computing platform for bioNMR. In addition, we are implementing a new file structure which bundles the original binary data files along with PREMIS XML records describing the provenance of the data. These are packaged together using a standardized file archive utility. In this manner, the provenance and data curation information is maintained together along with the scientific data. The benefits and limitations of these approaches, as well as future directions, are discussed in this paper.

Received 22 January 2018 ~ Accepted 22 January 2018

Correspondence should be addressed to Michael R. Gryk, School of Information Sciences, University of Illinois, Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820 USA. Email: gryk2@illinois.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v13i1.657
DOI: 10.2218/ijdc.v13i1.657

Introduction

An acknowledged goal in the field of data curation is to move curation tasks upstream closer to the creation and origination of the data. When considering a scientific study, it is often the case that the digital dataset is not the product of a single, experimental observation. Rather, multiple observations are collected, digitized, normalized, cleaned, and otherwise transformed. Several different and potentially disparate datasets are then analysed in concert along an involved computational pipeline or workflow (Bowers and Ludäscher, 2005).

As the production of computational data through the aforementioned workflows has become increasingly complicated, there has been a growing concern for the lack of a detailed reporting of the workflows along with the intermediate datasets, particularly as these are necessary for the reproducibility and/or repeatability of the computation (Stodden et al., 2016). In the case of such scientific workflows, data curation must be an ongoing process along the entire computational pipeline.

In this paper we discuss efforts to improve the reproducibility of scientific computation by adding curation tasks directly within the construction and execution of the computational workflows. This effort is being conducted within the context of the field of biomolecular nuclear magnetic resonance spectroscopy (bioNMR) using the NMRbox[1] platform for bioNMR computation. The workflow management system used within NMRbox is CONNJUR Workflow Builder (Fenwick et al., 2015a). CONNJUR Workflow Builder (CWB) is currently being refactored such that curation metadata will be stored as XML using a hybrid of the PREMIS metadata schema for digital preservation (Denenberg, 2014) and a bioNMR specific metadata schema currently referred to as CONNJUR_ML[2]. The short-term goal is to package this PREMIS XML file together with the scientific dataset (typically a binary file) using standard file archive utilities.

Biomolecular NMR Spectroscopy

Biomolecular NMR spectroscopy is a biophysical technique which exploits the magnetic moments of the nuclei comprising the matter all around us. A close sibling of Magnetic Resonance Imaging (MRI), which uses this intrinsic magnetism to image human tissue, nuclear magnetism is used in bioNMR studies to explore the structure and dynamics of biological molecules at atomic detail. These studies include determining the three-dimensional structure of proteins and nucleic acids, drug discovery, kinetics and mapping the interfaces of protein-protein and protein-ligand interactions.

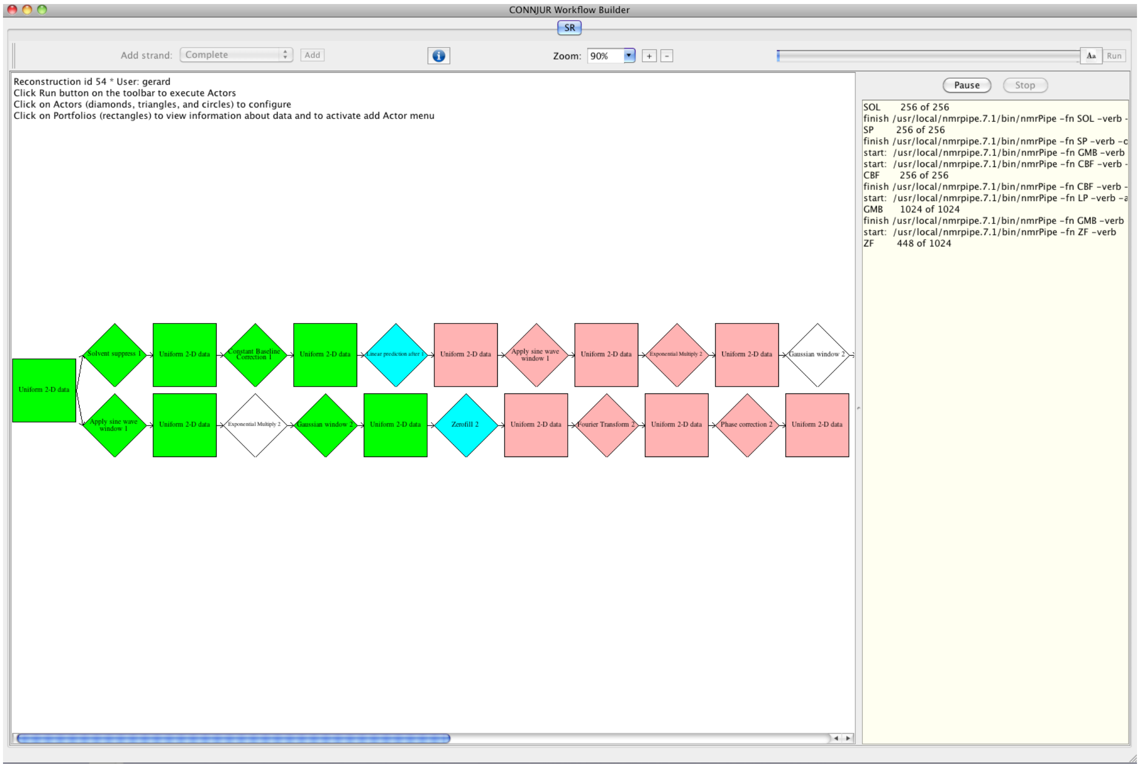
The computational workflow for modern bioNMR spectroscopy consists of three phases: spectral reconstruction, the process of converting time domain data into the frequency domain; spectral analysis, including peak identification and resonance assignment; and biophysical characterization, including all subsequent data analysis in which the spectroscopic data is used to draw biophysical inferences (e.g. structure determination) (Ellis et al., 2011; Verdi et al., 2007). The data semantics vary throughout these phases, from primary 'raw' data of the nuclear precessions to various

[1] NMRbox: http://www.nmrbox.org

[2] CONNJUR_ML: http://www.github/CONNJUR/CONNJUR_ML

levels of derived or interpreted data, including resonance frequencies and interatomic distances. This computational workflow uses more than a dozen, academically-developed software tools with many file translation and data cleaning steps along the workflow. Proper curation of the bioNMR workflow is an ongoing challenge affecting data sharing, the archival of research results, and the reproducibility of prior studies (Maciejewski et al., 2017).

NMRbox (Maciejewski et al., 2017) is a recent initiative to foster computational reproducibility for the bioNMR community by (a) establishing an archive of the various software tools for bioNMR and (b) provisioning a virtual machine for bioNMR computation. NMRbox uses CONNJUR for semantic data management within these virtual machines. CONNJUR[3] is a long-standing project for developing a software integration environment for bioNMR, and currently supports CONNJUR Workflow Builder (Figure 1), a scientific workflow management system for bioNMR spectral reconstruction (Fenwick et al., 2015a). The benefits of using CWB are that the metadata are stored in a relational database making them available throughout the computational workflow, multiple software tools can easily be interleaved within the workflow, and the workflows can be exported as XML to facilitate reuse and sharing between researchers (Fenwick et al., 2015a). An example of CWB used for spectral reconstruction is found on our YouTube channel under the title 13C HSQC[4].



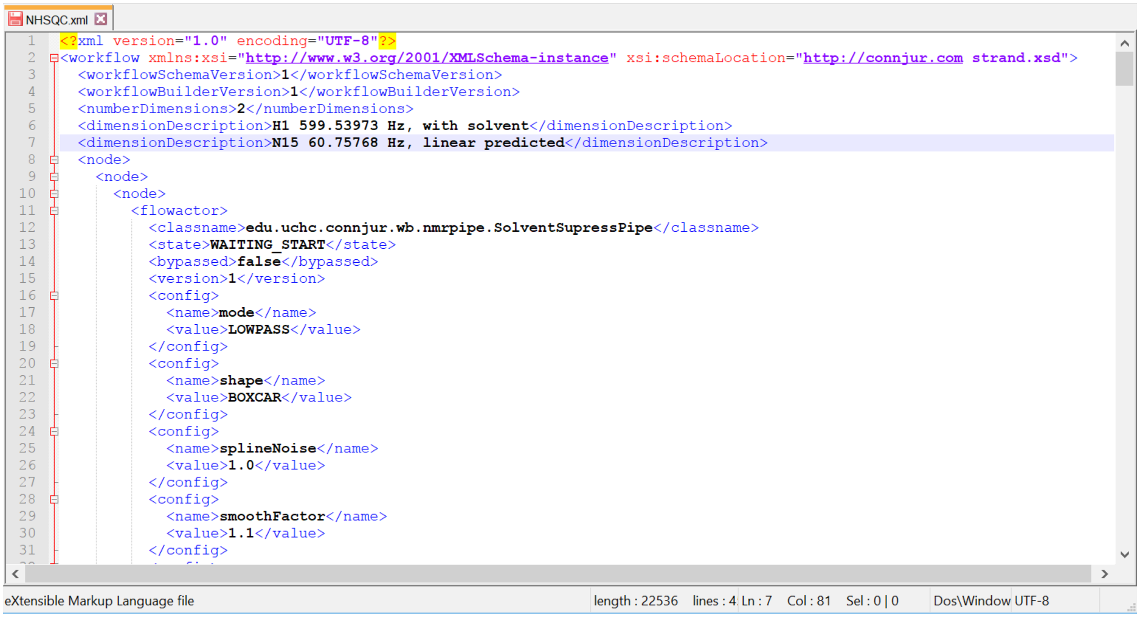
Reconstruction id 54 * User: gerard
Click Run button on the toolbar to execute Actors
Click on Actors (diamonds, triangles, and circles) to configure
Click on Portfolios (rectangles) to view information about data and to activate add Actor menu

```
SOL     256 of 256
finish /usr/local/nmrpipe.7.1/bin/nmrPipe -fn SOL -verb -
SP      256 of 256
finish /usr/local/nmrpipe.7.1/bin/nmrPipe -fn SP -verb -c
start:  /usr/local/nmrpipe.7.1/bin/nmrPipe -fn GMB -verb
start:  /usr/local/nmrpipe.7.1/bin/nmrPipe -fn CBF -verb -
CBF     256 of 256
finish /usr/local/nmrpipe.7.1/bin/nmrPipe -fn CBF -verb -
start:  /usr/local/nmrpipe.7.1/bin/nmrPipe -fn LP -verb -a
GMB     1024 of 1024
finish /usr/local/nmrpipe.7.1/bin/nmrPipe -fn GMB -verb
start:  /usr/local/nmrpipe.7.1/bin/nmrPipe -fn ZF -verb
ZF      448 of 1024
```

Figure 1. Screenshot of the graphical canvas for CONNJUR Workflow Builder. Squares represent datasets while diamonds represent actors. The above workflow is in the process of being executed. Green objects are those who have completed successively, blue are in progress, and pink actors have yet to be invoked. While actors are being bypassed and not executed in the above workflow.

[3] CONNJUR: http://www.connjur.org
[4] 13C HSQC: https://www.youtube.com/watch?v=7sB8oQFHE1E

While useful in this context, the data model for CWB is restricted to the context of bioNMR and the implementation details for constructing and executing the workflow itself. As pointed out by Willoughby and Frey (2017), such customization has the drawback that the reuse and sharing enabled by workflow management systems such as CWB is limited in scope to the users of CWB. This can be more readily appreciated by examining the XML output from version 1 of CWB. (Figure 2). The XML for the exported workflow contains information about the Java classes required to configure and execute the workflow within the CWB program. With appropriate knowledge of the operation of CWB and the XML itself, it should be possible to translate such a workflow into a more generic description amenable to other workflow management systems. However, this knowledge requirement is quite extensive and has prompted us to examine a more generic workflow representation which is both useful for the bioNMR community it serves, as well as the broader audiences of workflows, provenance and digital curation. The broad XML architecture chosen for this purpose is that of PREMIS, a general framework used by the digital preservation community.



```
<?xml version="1.0" encoding="UTF-8"?>
<workflow xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://connjur.com strand.xsd">
  <workflowSchemaVersion>1</workflowSchemaVersion>
  <workflowBuilderVersion>1</workflowBuilderVersion>
  <numberDimensions>2</numberDimensions>
  <dimensionDescription>H1 599.53973 Hz, with solvent</dimensionDescription>
  <dimensionDescription>N15 60.75768 Hz, linear predicted</dimensionDescription>
  <node>
    <node>
      <node>
        <flowactor>
          <classname>edu.uchc.connjur.wb.nmrpipe.SolventSupressPipe</classname>
          <state>WAITING_START</state>
          <bypassed>false</bypassed>
          <version>1</version>
          <config>
            <name>mode</name>
            <value>LOWPASS</value>
          </config>
          <config>
            <name>shape</name>
            <value>BOXCAR</value>
          </config>
          <config>
            <name>splineNoise</name>
            <value>1.0</value>
          </config>
          <config>
            <name>smoothFactor</name>
            <value>1.1</value>
          </config>
```

eXtensible Markup Language file   length : 22536   lines : 4   Ln : 7   Col : 81   Sel : 0 | 0   Dos\Window UTF-8

Figure 2. Notepad++ screenshot of original CONNJUR Workflow Builder (CWB) XML schema. While this XML document could be shared between CWB users, the metadata schema was specific for CWB and not intended for broader distribution. It contains information about the software state as well as specific Java classes of software code.

PREMIS

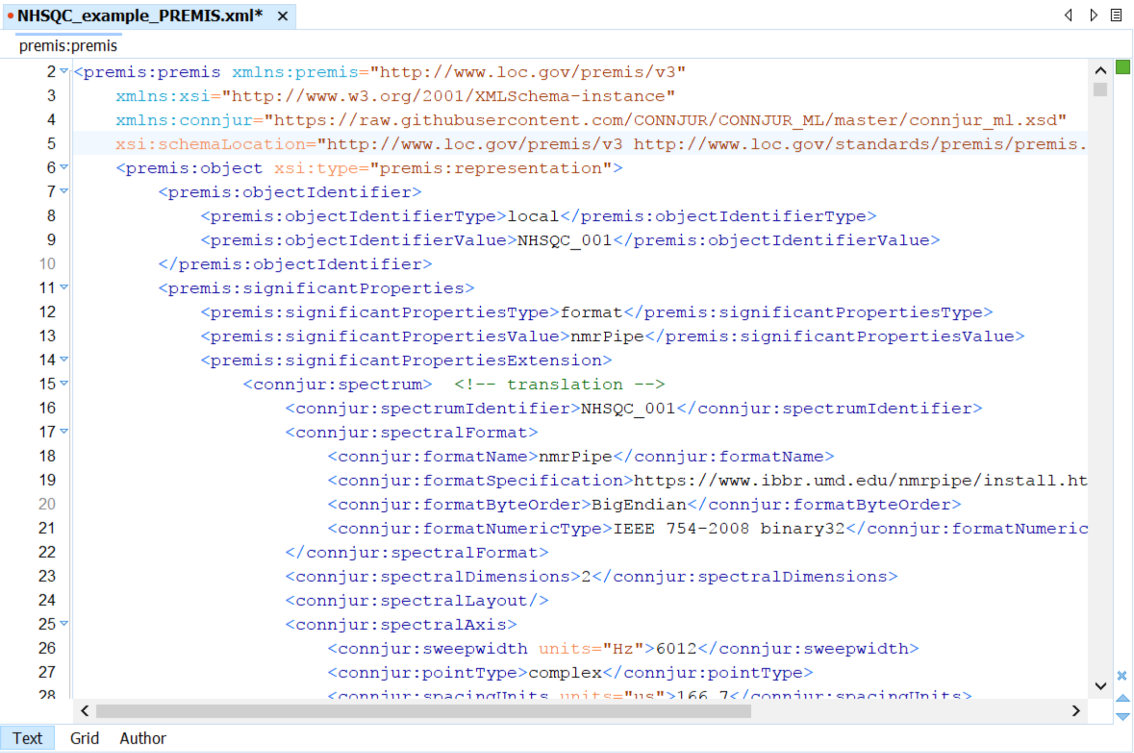
Maintained by the Library of Congress, Preservation Metadata: Implementation Strategies (PREMIS) has been developed by the archival and library communities to provide digital preservation systems with a framework to build reliable systems for sustainable information stewardship (Deneberg, 2014). With the release of version 3 of PREMIS, it is possible to embed custom structural, descriptive and administrative metadata within a PREMIS XML record. We have developed a bioNMR XML for spectral reconstruction which is currently made available on GitHub. Called CONNJUR_ML, this bioNMR XML is intended to provide metadata suitable for the significant_properties field in PREMIS objects, as well as the extension field for PREMIS agents to provide metadata about the software and hardware environments.

Digital preservation is itself a type of workflow, albeit a very general workflow supporting a wide assortment of digital objects. CWB was re-assessed in the context of PREMIS as it became clear that the panoply of workflow components for spectral reconstruction are all variations on the data cleaning, data translation, data normalization and data transformation operations typical for digital preservation workflows. The important metadata being captured by CWB throughout the workflow are structural and administrative metadata important for both describing how the component processing steps were configured, and also for identifying classification properties of the intermediate datasets. Willoughby and Frey (2017) call attention to the importance of the intermediate data generated during a scientific workflow. In the case of bioNMR, the intermediate datasets can be quite large, often gigabytes in size. We have found that a detailed metadata record can be a suitable surrogate to storing the actual intermediate files.

CONNJUR Workflow Builder

The data model for CONNJUR Workflow Builder is being refactored to support the more general concepts provided by PREMIS, as well as the bioNMR-specific classifications of the bioNMR data. The next version of CWB will import and export spectral reconstruction workflows as PREMIS digital preservation records, widening the audience for data sharing and reuse. As part of its integration within the NMRbox platform, CONNJUR/PREMIS workflows will also be translated into NMR-STAR, the file format supported by the primary biomolecular NMR data repository, the BioMagResBank[5] (Ulrich et al., 2008). By supporting linked digital preservation events, the PREMIS metadata standard provides provenance of what operations were performed on any digital object (in our case, bioNMR datasets), when they were performed, by what human agents and using what software tools – up to and including the entire software environment. As part of the NMRbox platform, this entire software environment is archived and will be accessible as a virtual machine for several years to come.

[5] BioMagResBank: http://www.bmrb.wisc.edu/



```
<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.
    <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>  <!-- translation -->
                    <connjur:spectrumIdentifier>NHSQC_001</connjur:spectrumIdentifier>
                    <connjur:spectralFormat>
                        <connjur:formatName>nmrPipe</connjur:formatName>
                        <connjur:formatSpecification>https://www.ibbr.umd.edu/nmrpipe/install.ht
                        <connjur:formatByteOrder>BigEndian</connjur:formatByteOrder>
                        <connjur:formatNumericType>IEEE 754-2008 binary32</connjur:formatNumeric
                    </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>
```

Figure 3. Oxygen screenshot of PREMIS XML record with CONNJUR_ML metadata embedded within the 'significantProperties' PREMIS tag. CONNJUR_ML is an ongoing modelling task and the XML can be found on GitHub.

Figure 3 shows a snippet of a CWB metadata record using PREMIS in conjunction with CONNJUR_ML (the actual metadata record for this workflow is over 1,000 lines long). PREMIS is used to define the intellectual objects represented by each intermediate dataset along the computational pipeline. As these intermediates are only intended to be stored local to the execution environment, they are given local identifier types and values.

Each intermediate dataset is manifested (perhaps only transiently) in a file format which can be interpreted by the underlying software tool which will process the data. The format of the data is recorded using the PREMIS significant properties tag. All additional structural metadata about the intermediate dataset is contained in the significant properties extension section. This makes use of the bioNMR specific CONNJR_ML to record important features of the intermediates – such as the levels of signal coverage, resolution, as well as various state information regarding the types of transformation which have been applied. This structural metadata can be used as a sort of fingerprint for (a) classifying which types of transformations can be made on which datasets as well as (b) for validating the workflow.

Each subsequent processing step, in which an actor ingests a dataset and produces another, is recorded as a PREMIS event with associated metadata describing and linking the events to the intellectual objects with which they are involved. Additionally, each PREMIS event is associated with a PREMIS agent which may be a human or a computer. For the computational workflows executed by CWB, additional metadata on the software and software versioning is added to the PREMIS agent extension.

Limitations and Future Directions

Limitations of this approach are related to that of virtual machines themselves. While the digital image of a virtual machine can be stored indefinitely, its operation is dependent on suitable computing hardware and hypervisor software capable of emulating the software environment. Virtual machines are currently ubiquitous and are expected to remain so for years to come; however, future migration to new hardware/software would seem unavoidable.

It is for this reason that many scholars concerned with reproducibility are advocating the use of container technology such as Docker. Docker containers provide a lighter footprint of virtualization and might be expected to remain stable farther into the future than full-fledged VMs. However, NMRbox currently supports over 100 individual software tools. To package each of these tools within an individual container and have the individual containers all interoperate is a significant task. Nevertheless, this usefulness of containers is being explored by the developers at NMRbox.

In this context of long-term sustainability, one of the added benefits of describing workflows within a digital preservation standard such as PREMIS is that the transformations along the workflow are described using a broader language, making the context and purpose of the workflows accessible to a wider audience. In this sense, our workflow management system is being repurposed as a data curation tool.

Our current metadata model and workflow implementation is for well-structured workflows using primarily mathematical transformations. Additional future directions will be to provide for more general data curation activities and more free-form annotation schemes such as done for spectral analysis using the reproducibility extensions of NMRFAM-Sparky (Fenwick et al., 2015b).

Acknowledgements

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.

References

Bowers, S. & Ludäscher, B. (2005). Actor-oriented design of scientific workflows. In: L. Delcambre, C. Kop, H.C. Mayr, J. Mylopoulos, & O. Pastor (Eds) Conceptual Modeling – ER 2005. ER 2005. *Lecture Notes in Computer Science, 3716* (pp. 369-384). doi:10.1007/11568322_24

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

Ellis, H.J.C., Nowling, R.J., Vyas, J., Martyn, T.O., & Gryk, M.R. (2011). Iterative development of an application to support nuclear magnetic resonance data analysis of proteins. *Proceedings of the International Conference on Information Technology: New Generations 2011*, 1014–1020. doi:10.1109/ITNG.2011.215

Fenwick, M., Weatherby, G., Vyas, J., Sesanker, C., Martyn, T.O., Ellis, H.J.C, & Gryk, M.R. (2015a). CONNJUR workflow builder: A software integration environment for spectral reconstruction. *Journal of Biomolecular NMR, 62*, 313-326. doi:10.1007/s10858-015-9946-3

Fenwick, M., Hoch, J.C., Ulrich, E.L., & Gryk, M.R. (2015b). CONNJUR R: An annotation strategy for fostering reproducibility in bio-NMR-protein spectral assignment. *Journal of Biomolecular NMR. 63*, 141-50. doi:10.1007/s10858-015-9964-1

Maciejewski, M.W., Schuyler, A.D., Gryk, M.R., Moraru, I.I, Romero, P.R., Ulrich, E.L., Eghbalnia, H.R., Livny, M., Delaglio, F., & Hoch, J.C. (2017). NMRbox: A resource for biomolecular NMR computation. *Biophysical Journal, 112*(8), 1529-1534. doi:10.1016/j.bpj.2017.03.011

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*, 1240-1241. Retrieved from http://doi.org/10.1126/science.aah6168

Ulrich, E.L, Akutsu, H., Doreleijers, J.F. Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Wenger, R.K., Yao, H. & Markley, J.L. (2008). BioMagResBank. *Nucleic Acids Research*, *36*, D402–D408. doi:10.1093/nar/gkm957

Verdi, K.K., Ellis, H.J., Gryk, M.R. (2007). Conceptual-level workflow modeling of scientific experiments using NMR as a case study. *BMC Bioinformatics, 8*, 31. doi:10.1186/1471-2105-8-31

Willoughby, C. & Frey, J.G. (2017). Documentation and visualisation of workflows for effective communication, collaboration and publication @ source. *International Journal of Digital Curation, 12*, 72-87. doi:10.2218/ijdc.v12i1.532