NAN
Network for Advanced NMR

Network for Advanced NMR Use Cases

Chris Bontempi, Director of IT

University of Connecticut Health Center

Farmington, CT

“Mid-scale RI-2 Consortium: Network for Advanced NMR” Award #1946970

What is the Network for Advanced NMR (NAN)?

NAN is a joint NSF-funded project (NSF grant “Mid-scale RI-2 Consortium: Network for Advanced NMR” Award #1946970) across the University of Connecticut Health Center (UCHC), the University of Georgia (UGA) and the University of Wisconsin (UW-Madison) that entails the following:

1. Purchase and installation of two 1.1Ghz Nuclear Magnetic Resonance (NMR) spectrometers, one at UGA and one at UW.
2. The automated transmission of spectrometer acquisition data (aka “experiments” or “spectra”).
3. A Data Browser – The ability for users to manage, share and use the data collected and stored in the NAN repository.
4. A Resource Connector – The ability for the public to see and utilize the resources available at the various sites including spectrometers, services, available options, etc.
5. A Knowledge Base – The ability for users of various experience levels with NMR to learn, understand and make use of previous work to advance the appropriate use of NMR.
6. A strong emphasis on democratization of NMR, the principles of FAIR and the concerns and considerations of NSF and associated federal agencies.

Overview of Some Data Elements

1. Spectrometer (aka instrument, magnet, workstation) – This is the instrument used to gather “spectra” from a “sample”.
    a. Challenges
        i. This is a multi-part device, made up of a control workstation, a computer that operates the mechanisms, a magnet, cooling units, and a variety of options that affect how data is collected. It is challenging to uniquely identify what a “spectrometer” is – there is no single, essential, core element of a spectrometer. Virtually every piece can be swapped out.
        ii. The computers that run these devices often run only the original software that was installed on them (Linux mostly) and are not kept up-to-date, so many of them are limited in what they can do. For example, many run CentOs 5, which does not support anything later than TLS-1.0. We also need to be wary of anomalies that uncorrected bugs might introduce into the data.
    b. Sample attributes

        i. Workstation operating system (Linux or Windows) and version
        ii. Workstation spectrometer software (Topspin or VNMRJ) and version
        iii. Technical attributes (e.g., field strength (typically between 500Mhz and 1.1Ghz, bore, channel count, receiver count, etc.).
        iv. Identifying attributes (e.g., model, serial number, year configured, year commissioned, etc.)
        v. Availability (e.g., online, free schedule, etc.).
        vi. Cost
        vii. Grant/Funding
        viii. Service records
    c. Associated data elements
        i. Magnet (manufacturer, age, etc.)
        ii. Probe (manufacturer, specifications, type, specifications, etc.)
        iii. Sample Changer
        iv. Etc.

2. Experiment (aka acquisition, spectra) – This refers to both the act of acquiring spectra (when, where, how) and the results of the acquisition (the spectra itself).
    a. Challenges
        i. There are differing applications for NMR, of which the main three we are focusing on are
            1. Solution State – typically a single, relatively long-running acquisition
            2. Solid State – often many very short-running acquisitions
            3. Metabolomics – acquisitions across scores or hundreds of samples at once
        ii. In addition, there are often test, calibration or “shim” acquisitions the results of which are unimportant. There is no distinctive way to tell these from important acquisitions at acquisition time. We have proposed allowing users to review experiments and mark them as important or unimportant, but in some cases these are 100s or 1000s of experiments. We can think of ways to make it practical, but it is still overhead for the user.
        iii. Spectrometer software does not uniquely identify an instance of an acquisition – there is no internal unique identifier. We use the spectrometer and acquisition start time to uniquely identify an instance of an acquisition, since it is impossible for more than one acquisition to happen at exactly the same second. We assign a unique UUID4 to each experiment.

   b. Sample attributes
      i. Start time – which is often questionable because these devices are highly isolated due to the likelihood of interference from virtually anything, and so the computers that run them are often entirely isolated also (because of physical limitations), so often no NTP, so the clocks continuously drift and are not often corrected.
      ii. End time
      iii. Original location on the spectrometer workstation of the data
      iv. Pulse sequence
      v. State of the spectrometer at the time of acquisition (probe, maintenance history, etc.)
      vi. Sample information
      vii. Solvent
      viii. Dimensions
      ix. Temperature (which is notoriously inaccurate)

3. Sample(s) – This refers to the item that is being evaluated with the spectrometer.
   a. Challenges
      i. Different users of the system have very different ideas of what a sample is and what it is composed of. Some view it as a known molecule in some sort of solution, with buffers, etc. Others view it as the process that was used to create the sample, because they are trying to understand what is actually in the sample.
      ii. Some users deal with individual samples, others with a small number of similar samples, and others, as noted above, with hundreds of samples on which they conduct the same experiment.
      iii. Samples can also be reused. This challenges a simple concept of an Experiment and a Sample, because an experiment can correspond to a single sample, multiple samples, or many samples over a significant period of time. Conversely, a sample can belong to a single experiment or multiple experiments. Samples can even belong to multiple people.
      iv. Furthermore, through titration and other techniques, samples can change over time, in their concentration and content.
   b. Sample attributes
      i. Focus
      ii. Buffer
      iii. Titration
      iv. Owner
      v. Original concentration

        vi. Creator (often a lab or sample vendor)

4. Other miscellaneous data elements
    a. Institution – school or institute that owns or hosts the spectrometer – has funding and ownership/control implications
    b. Facility – the unit within the institution that manages and operates the spectrometer – has full time staff that operates one or more spectrometers
    c. People (see Audiences below for details):
        i. User
            1. Anonymous
            2. Authenticated
        ii. Experimenter
        iii. Principal Investigator (PI)
        iv. Content Provider
        v. Facility Manager
        vi. Institution Representative
    d. Study – may have multiple specific meanings, but generally is a collection of related experiments with a common goal. May involve multiple Experimenters.
    e. Project – similar to a Study, but may involve other activities besides experiments and may, although not necessarily, correspond more closely with one or more grants.

Sample Audiences

1. User – users of the system
    a. We are obligated to support public, unauthenticated (anonymous) users of the system who are granted unimpeded read access to public information.
    b. We must also support the concept of authenticated users to facilitate private ownership (e.g., of experiment data that is not public) and areas of responsibility.
2. Experimenter – a sub-category of User, specifically a person who conducts an experiment. Specifically, someone who is interested in the results.
    a. This is distinct from a person who may physically conduct the experiment on behalf of an experimenter, who is a person who does the work but is not interested in the results. They are staff of the facility.
    b. This person is a part owner of the experiment information, and controls the public/private status of it to some extent.

3. Principal Investigator – the person ultimately responsible for the spending of funds related to experiments, spectrometers, spectrometer facilities, etc. Ordinarily associated with grants and other funding sources.
    a. This person typically has a group of experimenters, and is often somewhat removed from the day-to-day activities involved (i.e., they may never conduct experiments themselves).
    b. They conduct and control the resources that belong to them, and are part owners of the results obtained under their auspices.
    c. They are held responsible for the use of resources under their control.
4. Content Provider – this is a generic term for a subject matter expert who is permitted to contribute to the overall system, is normally identified, and is recognized as an expert in the field.
    a. May create new content.
    b. May verify the accuracy of content created by others.
5. Facility Manager - Overall manager of a given facility that contains NMR spectrometers.
    a. May have other responsibilities.
    b. Maintains spectrometers and the environment, schedules usage, supports spectrometer users, resolves issues.
    c. Is held responsible for the smooth operation of the facility, as well as spending related to the facility.
6. Institution Representative – An authoritative representative from a given institution that can resolve issues around funding, operations, ownership of resources.
    a. Is the top decision-maker with regard to compliance with the policies and procedures we are required to implement based on the rules of our grant.
    b. Represents the interests of the overall institution.

Sample Use Cases

Use We Expect

In general, we expect the data we've collected to be used in the following ways:

1. Ideally, every experiment we collect will have everything required for a given user to:
    a. Process the results of the experiment, which can be done in a number of ways using a wide variety of applications (200+ and counting).
    b. Reproduce an experiment.

2. We expect scientists to be able to use information across many experiments to find patterns, discover anomalies, and perform other analytic tasks, including the use of AI/ML.
3. We expect facility managers, PIs and others with administrative interests to want statistics regarding the use of spectrometers, project/grant-related statistics, and other information across a subset of metadata, rather than the actual raw data itself.

Use Case 1 – Experimenter Conducts an Experiment

Typical Process

The Experimenter books time on a spectrometer, goes to the spectrometer on the given date/time, places their sample in the device, calibrates and shims, and then runs the experiment (acquisition). The Experimenter will usually review the results as the acquisition is running, and will often stop it and make adjustments and restart it. Or they will conduct multiple acquisitions and vary some of the conditions. Sometimes they will be interested in the results of all the acquisitions, and other times they will only be interested in some of them or only the last one. Acquisitions can take anywhere from seconds to days.

Notable Variations

- Solid State Experimenters (material science and chemistry users) often run hundreds of discrete acquisitions over a several-day period. We do not fully understand yet how much of this data is useful to them.
- Metabolomics Experimenters will run hundreds of instances of the same experiment against hundreds of samples (think urine or blood tests). An example of this is evaluating the results of a drug trial, looking for elements that have been metabolized in the samples to predict adverse drug effects in a test population before those adverse effects manifest.
- An Experimenter may book time at a remote facility, send their sample to the facility, and have a facility staff member conduct the experiment(s) on their behalf. The staff will have to properly handle the sample, properly (meaning according to what the experimenter wants) calibrate and shim, and accurately represent the conditions at the time of the experiment.

Use Case 2 – An Experimenter Uses the Results of their Experiment

Typical Process

Once an experiment has been conducted and collected, the Experimenter will leave the facility and do things with the results of their experiment. These things may include:

1. Processing the results to visualize the spectra generated – this is generally done in a copy of the spectrometer software (Topspin or VNMRJ) on another computer.
2. Using sophisticated processing software to process the spectra – programs besides the original spectrometer software.
3. Sharing the results with a colleague.
4. Making the results public. This is generally done for sharing purposes, often for publication, for collaboration, or what have you. There is currently a bigger push to make results acquired at a government funded facility public sooner.
5. Using the results as the basis for a publication.
6. Using the results or the parameters used to obtain the results as the basis for a Knowledge Base article to help others conduct similar experiments.
7. Using the results, they may create a protocol that others can follow to produce the same or similar results.
8. They may combine these results with others to produce AI/ML models that surface trends or make predictions.
9. Collecting experiment results for usage in the following example variations:
   a. All of my experiments across all facilities and spectrometers.
   b. All of my experiments from a single spectrometer.
   c. All experiments using a given pulse program.
   d. All experiments from a single spectrometer.
   e. All experiments that ran for more than 3 days.
   f. Any experiments done on E. Coli samples.

Use Case 3 – A Content Provider Shares Their Expertise

Typical Process

Content Providers are people we expect will help to democratize the use of NMR spectroscopy. They may do so in a variety of ways, which include:

1. Producing introductory processes that novice users can follow.
2. Describing complex processes that intermediate and advanced users can follow.
3. Reviewing and enriching other people's content.
4. Producing specific parameters (pulse sequences, sample/solution handling, etc.)

Use Case 4 – A Facility Manager Describes Their Facility and Services

Typical Process

A facility manager at a given facility describes the resources they are making available to members of the general population who wish to avail themselves of NMR spectroscopy services. Including

1. The spectrometers at the facility, along with the various options like probes, sample changers, etc.
2. The services provided at the facility such as
    a. Sample handling, preservation, disposal, etc.
    b. Remote acquisition – users can conduct their own acquisition via remote desktop.
    c. Surrogate acquisition - a surrogate conducts experiments on behalf of another user, lending both their time and their expertise to the process.
    d. On-site support.
3. Availability and scheduling of specific spectrometers, probes and services.
4. Pricing for use of the facility spectrometers and services.
5. Areas of expertise (solution, solid, metabolomics, ultra-high-field, room temperature/cryogenic probes, etc.).

Use Case 5 – A Facility Manager Wishes to Optimize/Expand Their Facility

Typical Process

A facility manager will want to understand and optimize the use of their facility, to ensure smooth operating and with an eye toward expanding their capabilities based on evidence, in some of the following ways:

1. They will want to know what experiments and types of experiments are being conducted, and by whom, to spot trends that they can either better cater to or suggest alternatives around.
    a. For example, if a given user repeatedly conducts an experiment on a given spectrometer that could be conducted on a lower-cost spectrometer without impacting the results, the facility manager might want to make that recommendation to the user.
2. They may want to balance usage across their various spectrometers and optimize the use (minimize idle time).
3. They will want to understand usage patterns so they can schedule necessary maintenance without impacting demand.
4. They will want to spot trends in variation, to spot areas requiring maintenance before outright failure (which in some rare cases can be catastrophic to the spectrometers).
    a. For example, they may want to repair or replace a probe that is degrading or starting to fail before experimenters waste time and money using it.

Use Case 5 – A Principal Investigator (PI) Wishes to Understand and Manage Their Resources

Typical Process

A Principal Investigator is responsible for the resources related to the grants they have been given. As a result, they may be interested in the following:

1. Usage of equipment (e.g., spectrometers) purchased as part of a given grant.
2. Grant money spent on projects and studies related to a given grant (e.g., the cost of using time on a spectrometer).
3. Volume and type of data acquired or generated for a given grant.
4. Findings published based on work done under a given grant, or using resources of a given grant.
5. Based on evidence, areas in which they can pursue other grants to better serve the community and advance the science.

Use Case 6 – Changes in Personnel Occur

Typical Process

Specific individuals are identified as having specific roles. But individuals retire, resign, pass away, etc. Implications for personnel changes to NAN include:

1. A graduate student (Experimenter) leaves a PI's group for another opportunity. The PI has to be able to control the data they collected, and possibly assign it to someone else to manage.
2. A PI retires, and all of the resources collected as parts of their grant need to be managed and adjudicated by a representative of the institution.
3. A PI changes institutions, and may or may not take some of their grant-related resources with them.
4. A Facility Manager hires a colleague to share the workload, so now the facility has more than one Facility Manager.
5. A Facility Manager retires or moves to another institution. Since the Facility Manager is largely responsible for how the facilities operate, changes may occur when a new Facility Manager takes over.

Use Case 7 – A User Wishes to Add Experiment Data That Was Not Automatically Collected

As part of the knowledge base or other parts of the system, some participants may want to manually upload data they collected outside of the network of NAN spectrometers (in other words, the data was not automatically collected directly off a spectrometer). Considerations in this area include:

1. The data may not have all the same context (spectrometer details, software versions, conditions of the facility, etc.) as data that is automatically collected.
2. The lineage of the data may be lost or may be unreliable.
    a. This will be important to consumers of data for whom the lineage is vital.

Use Case 8 – A User Wishes to Blend NAN Data with Externally Sourced Data

An example of this might be Deepmind's Alpha Fold project and the protein structures they have managed to produce. Considerations include:

1. NMR data is particularly open to interpretation and variation. Alignment of data sources is a known challenge.
2. How do the units, periodicity and foci align with data that was produced from another source?
3. Can adjustments be made to the NAN data repository interface to allow the data to align with outside sources, or is it better to align the data from external sources with the NAN model as it is consumed?

A Summary of General Considerations

1. Since the database is intended to be multipurpose and for use by multiple audiences, it seems unlikely that a single, universal data model will serve any given audience well.
    a. We need to be able to present the data to consumers in a way that fits their paradigm, scholarly idiom, and concept of the data.
    b. These presentations have to be accurate, reliable, and easy to create.
2. Since one of the uses of this data will be for publishing, the following must hold:
    a. The lineage must be accurate and complete, and traceable back to the original conditions in which the data was collected.
    b. Considerations about the reliability of the data must be surfaced.
3. Access to the data has to be carefully considered:
    a. Private data must be protected in cases where the owner of the data has not yet made the data public.
    b. Data must be able to be made public easily, and we are obligated to allow public, unauthenticated, anonymous access to public data.
    c. A well-understood chain of ownership of the data must be maintained and adhered to.

4. Scale, units, precision, summarization and other individual data attributes must be carefully maintained and tracked in the system. Conversions with regard to any of these attributes need to be carefully tracked and auditable.
    a. For example, if a very small number is made effectively zero due to a conversion, it should be obvious to the user that this number is effectively zero, rather than actually zero.
5. If data is corrected or removed relative to the original source data, that should also be obvious to the user.
    a. We must always maintain the original source data for comparison.
6. We want the data to be as easy to characterize as possible. This may cut across a number of dimensions, but some examples are:
    a. What concentration of a given data attribute can be considered accurate or inaccurate, or missing or present?
    b. What concentration of a given data attribute is derived from lower versus higher precision?
    c. Are there patterns related to the collection environment, the person(s) conducting the experiment, the manufacturers of various components, etc.?
7. Overall, we want the entire system to be as self-service as possible for the users, since our resources are limited and users will have a strong need to move forward rapidly with the data they have collected and need access to.
8. Regarding alignment of multiple data sources, a couple of points are worth noting:
    a. UCHC houses two other important and complementary projects:
        i. Biological Magnetic Resonance Bank (BMRB) - a storehouse for NMR spectra and quantitative data derived from experiments.
        ii. NMRBox - a processing environment comprising hundreds of software packages, most of which work with NMR spectra.
    b. We are therefore highly motivated to support interchange capabilities with these and other (e.g., the PDB) NMR-oriented services.
    c. We are aware that even just within the biological NMR space there is some degree of conceptual variation. So while we can talk about ideas like “alignment”, for us to achieve true usefulness we will need to support some fluidity in the conceptual framework with the ability to continually modify and add to those frameworks. Simply put, one set of metadata and one conceptual schema is likely to limit the usefulness of NAN as a source of NMR experiments.

d. Add to the above statement the fact that we will have solid state NMR data and metabolomics data as well, and the variation in conceptual framework grows considerably.

Acknowledgements

PIs for Network for Advanced NMR

- Art Edison, University of Georgia, Athens, GA
- Kathernine Henzler-Wildman, University of Wisconsin, Madison, WI
- Jeffery Hoch, University of Connecticut Health Center, Farmington, CT