NMR Task Force Meeting

Date:

19.07.2023 @ 09:00 - 18:00 CET

Type:

Hybrid

Location:

In person:

Address: Lessingstraße 8, 07743 Jena
Conference room: Seminar room 127
(Building door has a password, please contact Noura (call, whatsapp, telegram) to open the door when you arrive. Number: 017658071461)

Online: Zoom

https://uni-jena-de.zoom.us/j/64825838609
Meeting ID: 648 2583 8609
Passcode: 750774

One tap mobile
+496938980596,,64825838609#,,,,*750774# Germany

Dial by your location
+49 69 389 805 96 Germany
Meeting ID: 648 2583 8609
Passcode: 750774

Find your local number: https://uni-jena-de.zoom.us/u/cbBD8ZpGLa

Workshop folder

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

NFDI4Chem
ENHANCE YOUR DATA.

Time-table

Time (CET)	Session no.	Title	Speaker/Moderator
09:00		Welcome	Christoph Steinbeck
09:10	1	MIChI for NMR Analysis of Synthetic Compounds	Christoph Steinbeck
10:00	2	Scientific applications development - NMRium	Luc Patiny
11:30		*Lunch - Self-paid*	*Mensa Philosophenweg 20, 07743*
12:30	3	Towards open NMR data formats - what we want and what we can have	Johannes Liermann
~~14:00~~		~~Discussions~~	
14:30		*Group picture* + *Coffee break*	*Seminar room 127*
15:00	4	nmrXiv + NMRium as the open source choice for NMR platforms	Nils Schlörer
16:00	5	NMR Data Packaging with Metadata in NFDI4Chem	Noura Rayya
17:00		Wrap-up and next steps	
18:00		End of the Workshop	
18:30		*Dinner - Self-paid*	

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

Sessions description:

1. Recommendations for Reporting Liquid-State NMR Analyses of Synthetic Compounds Metadata: A proposal from NFDI4Chem of what metadata to report for liquid-state NMR Analyses of Synthetic Compounds.
2. Scientific applications development - NMRium: Modular approach for the development of complex scientific applications like NMRium.
3. Towards open NMR data formats - what we want and what we can have: Looking into JCAMP alternatives (Bruker, nmrML. NMReData..). A well-specified format vs. an open one. What do we need for a standard?
4. nmrXiv + NMRium as the open source choice for NMR platforms: A proposal of an (ideal) workflow for NMR labs and some key features to be included.
5. NMR Data Packaging with Metadata in NFDI4Chem: Looking into available approaches of data packaging with metadata to use for data download and transfer - packaging our NMR MIChI in JSON format.

Technical Details

- Please **avoid using Zoom chat** to keep all the discussion in one place here.
- For presenters:
  - We will use one laptop for presenting, please make sure to make your presentation available in this folder.
  - Please mute zoom while people joining online are talking.
- For people joining in person:
  - Please keep your **laptop** volume muted as we will use the conference room speaker to hear people joining online.
  - Please keep your **zoom** muted as we will keep only the presenter laptop unmuted.
- For people joining online:
  - Please raise a hand before talking, or add notes to the document and we will discuss them.

Please feel free to add your notes, questions, remarks and discussion to this document.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

NFDI4Chem is supported by DFG under project number 441958208

Notes and Discussions:

Session 1:

- Chris kicking off with the NMR MiChI
- Noura: How we started: Looking into metadata extractable with NMRium and crucial ones, along with sources of impurities.
- Steffen: We would like to broaden the scope of our michie to be about small compounds, including np, synthetic compounds and mixtures where we have known molecules.
- Luc: We need to know if 1D or 2D, and some details are not that crucial like the instrument.
- Johannes: MI about instrument is relevant, especially the probe with the temperature because it affects the correlation.
- Stefan: Instrument is easily available for new instruments.
- Johannes: We have many required metadata because it is easily there.
- Luc: the only mandatory thing should be the sample details, not the instrument.
- Nicole: Q: What data we want in publications and repositories are two different perspectives and we need to know what we are discussing here. Because the mandatory here is easy only for repositories. Metadata for repositories should be more htan the ones for publications.
- Luc: Sample has metadata that we want to link to the spectrum, and for each spectrum we need nucleus, dimensionality.., the pulse, the instrument the probe, We have one sample to many spectra relation. For the sample we need to know the supplier the code.
- Stefan: I think this is only related to packaging and databases, but the relation between the sample and the spectrum here are not the main focus.
- Chris: I agree with both but maybe we can focus later on the raising problem with their relation.
- Johannes: We can keep the sample details out here from the michi.
- Pavel: Why number of points is scans? Chris: It is what you get for free.
- Luc: You don’t always get it for free so shouldn’t be mandatory.
- Chris: For historic data, then authors will fail to submit because of such issues.
- Johannes: There are two aspects: What I need to make use of the data? Nucleus.., and for data publication it is another aspect where we need to focus on traceability and reproducibility and such details like the number of scans and the instruments are very useful for reproducibility.
- Chris: Then we move them to optional instead of level1 as we get them anyway so we will have them anyway.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

- Johannes: making them mandatory encourages people to use automatic workflows.
- Chris: Can we see a situation where author fails to submit important dataset for failing to get unnecessary mi?
- Stefan: This scenario can happen with all sort of data even the important ones.
- Pavel: New parameter about the type of experiment was introduced recently and it should be added with the pulse sequence too.
- Steffen: https://docs.google.com/presentation/d/1d6PEaLVS3PLAonWkKXqO95Qq-FvrbkIMPyOLyVP2qv4/edit#slide=id.gf20c620c77_0_0
- Steffen: M4.1 MiChI Process
- What others have done mostly is the first column guidelines and it is kind of educations. What we are discussing in our table is kind of the check list. What luc discusses about the relations is mostly about the model, and what Noura mentioned is the ontology, and lastly we have the implementation.So now we are at the checklist to tell people what to check when publishing NMr data.
- Chris: What do we do now? What are the next steps? Is it publishing an article after we agree on the table or is there something in between? I suggest that the document should be circulated like with the IUPAC group.
- Steffen: FAIRspec people might have goldbook things related to us and we need to discuss with the rest to get things in shape.
- Chris: So should we aim to agree with FAIRspec about a joint MI? And then publish it with them?
- Steffen: FAIRspec published something related and we need to put here. IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles
  And the open Preprint IUPAC Specification for the FAIR Management of Spectroscopic Data in Chemistry (IUPAC FAIRSpec) - Guiding Principles
- Chris we still need to think about next steps.
- Stefan: I am not sure if FAIRspec is in the same direction. I suggest we first agree on the table and then we reach out. We need to go through it section by section and find a consensus. Otherwise it is tricky.
- Nicole suggested in the Chat to number the information in the MI table for referencing. Done by Noura now.
- We need to make our recommendations modular, we can focus on the experiment for now and not focus on the sample.
- Table discussion:
  1. NMR Sample:
     1.1 Solvent: Agreed on, if you mix two then it is a list, so the input format should be an array with at least one entry.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

1.2 chemical shift standar: Agreed on, it can be the same as the solvent and this makes the solvent more crucial.

2. NMR Instrument
   - 2.1. NMR instrument manufacturer: move to recommended
   - 2.2. NMR probe: We need to know the kind of the probe not the name or serial number. We need to add it to ontologies. move to recommended
3. NMR Acquisition Parameter
   - 3.1. NMR pulse sequence: agreed on, We need the type not the name. and we need terminology for that
   - 3.2. Magnetic field strength: agreed
   - 3.3. Number of scans: move to recommended for historic reasons and it will be there anyway if it is there.
   - 3.4. Acquisition nucleus: Agreed on.
   - 3.5. Relaxation delay: move to recommended
   - 3.6. Number of acquisition data points: move to recommended
   - 3.7. Sample temperature information: move to recommended
   - 3.8. Flip Angle: definition not clear.
   - 3.9. Pulse duration: move to recommended
   - 3.10. shaped pulse file: data or metadata? It makes things complex. Maybe better to use true false of shaped pulse was used. Move to recommended
   - 3.11. Pulse power: Move to recommended
   - 3.12. acquisition time: Move to recommended
4. NMR Data Processing
   - 4.1. Number of zero filling points: Move to recommended
   - 4.2. window function for apodization: Move to recommended
   - 4.3. window function parameters: Move to recommended
   - 4.4. baseline correction: Move to recommended
   - 4.5. baseline correction parameters: Move to recommended
   - 4.6. phase correction: Move to recommended
   - 4.7. zero order phase correction (ph0): Move to recommended
   - 4.8. first order phase correction (ph1): Move to recommended
   - 4.9. Absolute correction: Move to recommended

Remove optional section and move to recommended as more data the better.
We need to make clear that what we want is for people to publish their data where we can extract the metadata and not providing long lists of metadata in the experiment section.
When it comes to publishers, we should think of our impact on them to ask for more restricted data, but also not to miss crucial MI.

Chris: We stick to the time table and we close the session and we can continue

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

discussion later.

Session 2:

- Luc: We are trying to make programming fair, We have created some libraries that are very popular, react components and others. We have parsers, jcamp, or Bruker converters or to create jcam from other data..
- We are strict in our programming, we use conventional committing and based on that we have automatic generation of the change log based on bug fixes and new features and this can close issues automatically.
- We have a lot of continuous integration and automattic github testing.
- Coverage tells how much of the code was tested usually it is over 80% and we know which lines have'nt been tested.
- We publish our projects on zenodo and we get dois for them.
- CITATION.cff helps in generating citation for the project as txt or bibtex
- Why open source? From slide.
- Building react application: We use many little libraries for GUI, react components.. And we use libraries we have control on.
- In NMRium we use more than 60 libraries of our own.
- We use GitHub project to have an overview of the works
- We can archive what was closed
- Pareto principle: 80% of wealth belongs to 20% of people. We have something similar in our work where we use most of our time to fix bugs and such.
- We added automatic bug report on NMRium. To tell us what went wrong.
- We have important compomemnt in NMRium like converters, structure viewing, load and save data..
- nmr-load-save : from slide. And it can extract most of the metadata in NMRium from the raw files
- Some NMRium features: multiplet analysis NMRium demo.It doesn't need download as it works in the browser. You can display multiplicity trees. You can show the structure and do the assignment there.
- It is interactive between the tables and the molecule and the spectrum.
- It can generate reports
- NMRium allows to have a database like solvent or reference database.which you can search by structure or substructure or shifts..Also we have advance search to look for specific atoms.
- Prediction: We can 2D too, we provide structure and we predict the spectrum. It is also interactive, We predict Carbon, COSY, HSQC, HMBC
- There is a new feature from yesterday about simulation. You enter value sin the table to get a spectrum which is helpful for teaching.
- Metabolomics: Even in the browser we are still able to handle huge gigabytesand it allows filtering those hughes files for instance based on the type.
- Discussion:
    - License: MIT
- We do e-learning where you process the spectrum and enter your suggested structure and you find if correct or not.
- Future developments: from slides
- Pixelium: Works with images processing, not related to NMR, but uses a lot of libraries we created for NMRium.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

- Discussion:
    - We need to distinguish between formats for archiving and the ones for processing. mzML has been stable but people didn't like it for applications.
    - Massbank has many types of formats.
    - Bruker NMR is well documented while MS Bruker is not.
    - NMRium uses one json containing all the data related to the original and processed data.
    - Pavel: audit trails are needed and we need to be able to read the data in 30 years from now,
    - Luc: The data is in json which will be readable in 30 years.

We go back to the michi table, go up for the continuation.

Session 3:

- Johannes presentation: moving forward to an open NMR format
    - The situation is that we prepare a sample with a molecule and in the spectrometer we get information that we want to assign to the molecule. NMR data has several layers: raw, processed, assignment.
    - Another layer is the metadata. (details from slide)
    - In higher levels in NMr it is difficult to tell what is data and what is metadata.
    - Available NMR formats:
        - JCAMP: text based. With tags to define the metadata and also the actual data is in the jcamp along with the binary data which will be very length if not compressed.
            - Pros and cons: from slides.
        - Bruker: has folder structure with different files. FID has the binary data from the acquisition. Acqu is related to jcamp and provides metadata too.
            - Bruker dataset is all you need to reproduce an NMR experiment.
        - nmrML: has embedded ontology, mostly the only format in the NMR world with ontology use. It is not famous out of metabolomics and biology world as it is biology oriented, but this orientation can be fixed.
        - NMReData doesn't deal with spectral data but it goes to the sdf structure and extends the sdf with 1D, 2D, assignments, coupling constants, correlations…
        - The landscape: JCAMP and Bruker cover raw and processed data and if at all assignment data.
        - nmrML covers a little assignment but not designed for that.
        - NMReDara covers only assignment data.
    - Challenges:
        - Example: flip angle in Bruker: you need the pulse sequence for that. Clear for chemists but difficult for developers.
        - How can the repository know this is HMBC? There are many different pulse sequences names and difficult to know from it. Still they all follow a machine readable syntax that can identify the experiment.
        - For TOCSY: The names are more complicated without TOCSY in them
        - Pulse program: d8 is the mixing time and in the dataset file you have a D field and you have to find the 8th one.
        - Coupling constant: you can find it in the pulse sequence

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

        - Hypotheses: Using ontologies can also help in the unit, and it helps in making the solvent always the same value or the experiment type.
            - We need to contribute to available NMR ontologies.
- Discussion:
    - Pavel: We need a mapping between the ontologies descriptions and the vendors datasets.
    - Software can take ontologies and take a part of it for example to ask the user to pick a solvent.
    - What do we need to start using ontologies in a format?
        - We need identifiers, and the program doesn't need to understand the meaning of it, and we can add unit ontology .
        - Pavel: I don't recommend that because this modifies JCAMP in a way the chemists might not be interested in or understand.
    - What do we want from an open NMR format?
        - Do we need one to start with? If Bruker becomes very well documented and open their specifications could it be enough? Bruker documentation is already open.
        - If everyone in NMR can work with Bruker then why to have an open one? It will take too much human power.
        - It is better to modernize Bruker than maybe working with nmrMl and other formats.
        - JCAMP is popular for ELNs and difficult to replace by Bruker.
        - It is easier and cheaper to document an existing formt than developing a new one.
        - Can we work with IUPAC by taking available format and work on it and add from Bruker and ask for IUPAC recommendation.
        - Several analytical techniques all using JCAMP are suffering from the overlapping documentation.
        - IF we want to enhance jcamp, we need to decide the minimum to be included in that format. This makes it capable to have a lot of metadata and still not have important ones
    - JCAMP being textual is very old for parsers. Making it not encouraging to modernize it.
    - We can use the dot in JCAMP for optional fields to add what interests us un NMR.
    - There is no way to validate jcamp other than writing parsers which needs implementing all "standard" versions of it.

We are still stuck with the question of what format to go with:

- JCAMP misses important things can be an argument to enhance existing format to get better data.
- Johannes: we all agreed we do not want to develop new format, but to enhance existing formats. We need to modify it as little as possible to get what we need.
- What about taking jcamp model and provide another serialization of it as a json.
- Pavel: we need to have validation and testing for industry .
- Take home message:
    - Luc: If you want more metadata about the experiment you can add it to JCAMP.
    - An option: To get the files ae have and to provide a json on the side with metadata.
    - We need use cases when the available formats don't work for the end users.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

- Available NMR formats are not bad enough to push people to offer something else.

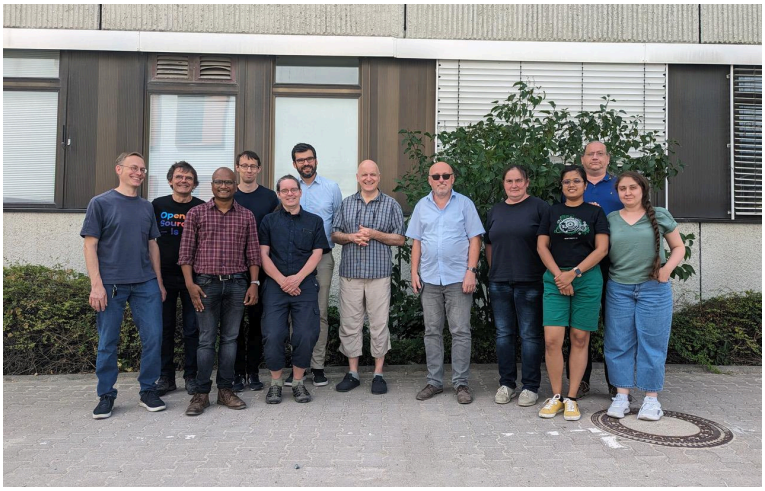
Session 4: Nils presentation

- We are focussing on academia not the industry. What tasks in the lab and the ideal workflow and what bottlenecks we have. What features we could have by merging nmrium with nmrXiv in such environment.
- In NMR lab: things are not standardized, some uses lims, other servers, what about open access, automation?? If you use lims things become easy for data submission without submitting a paper. Other option is providing the data to the user with accounts.
- The repository is important cause if it includes apps like NMRium it helps not only in visualization but it helps in processing, and also data assignment.
- There are 2 aspects: The side of the administrators where quality check and ranking can happen. And the aspect of the user expectations, if you give checkboxes for chemists with 50 ones, we should expect low submission.
- It is common to not provide 2D assigned spectra which is frequently needed in the common routines. A repository or a tool like NMRium can provide an added value by suggesting assignment.
- What is ideal for laboratory NMR workflow to generate data ready for submission:
    - The data should be hosted instead of local storage, which unfortunately people prefer.
    - You also need at least one structure linked to the spectrum. It is still common that people don't submit structures.
    - The local and public format for data storage should be the same to avoid switching between systems.
    - What to do with the data once created? People usually process on specific processing software, and it would be great for repositories to provide tools for directly visualize and process data on the same platform.
    - NMRium in nmrXiv gives living data.
    - Currently this is not the practice, people distribute some digital data and do paper assignment and the 2D information is usually mainly ignored. And this information can become digitally available instead .
    - Good to provide one interface for all tasks.
    - If the platform can give suggested assignment this is very attractive to users.
    - There should be some assignment control to check the plausibility by chemical shift comparisons but also can be used with case environment including logic, and if people do some assignment the system should be capable of continuing the job.
- Success factors:
    - Chemists don't like to handle several interfaces and instead they might go to paper.
    - An ideal repo should be able to do slide 15
    - Good to have a streamline for all the steps.
    - Very important to provide the structure always in the repo.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

    - The privacy issue is important and giving the people the option to publish data with a restricted group instead of just missing the data.
    - Slide 18 is the compilation of all the steps.
- Discussion:
    - Chris: We need to include ELNs in the vision
    - Steffen: Core facilities should have the role of raising the bar in making data digital. But not all facilities are allowed to do that. But they should push to get better.
    - Steffen: DFG hands out money only if you adhere to certain level. So the control over the data quality can happen from the funders side.
    - Chris: There are commercial management system in Jena platform and it is used instead of paper.
    - Nils, An issue with our platform software is that you can search by people and group, not by molecules, this is because structure submission is not mandatory.
    - Chandu: In nmrXiv, the repo layer is ready, but we are enhancing the submission layer and the molecule layer. We also want to build the prediction and search layer to be separate from nmrXiv to allow others to use the same projects. Collaboration is still needed with Nils to push things faster.
    - Pavel: Are you planning to work with ELNS? Chris: We already work with Chemotion, an open source ELN which we found to be the best in the chemistry area.

Group pic



Session 5: Noura's presentation

- nmrXiv data model and (Bio)schemas export
- Metadata is multiple pieces all over the data download ZIP
- Packaging options
    - BagIt has a poor schema support

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208

    - Ro-crate has some restrictions on datasets and media files
    - ... and the combination thereof :-)
- Pavel: is there a toolset to verify BagIt archives ?
    - https://github.com/joehand/bagit-tools
    - https://en.wikipedia.org/wiki/BagIt
- Pavel: Is an internet connection required for “identifier”: “https://doi ...”
- Bioschemas Reaction Prototype:
    - https://drive.google.com/file/d/1LfchAtmQ8tRGJEGV9_z9Pdp61IS22gIm/view
    - Some `additionalProperty` were needed to encode more of the internal chemotion/nmrXiv datamodel
- We agreed to not drop any ro-create.json into any Bruker directories for data hygiene reasons
- Suggestion
- Session 6: nmrCV ontology:
    - Steffen: nmrCV was a part on nmrML development, and we still can use it as controlled vocabulary.
    - nmrCV def=velopment was done before 2019, then there was a group who worked on pulse sequences, and there is TIB team who works on ontologies in NFDI4Chem.
    - nmrCV needs:
        - Structure cleaning in terms of hierarchy.
        - There are missing terms here and there.
        - There are white spots there that we can maintain and import,
        - Link fom Johannes
    - What do we have in terms of NMR ontologies?
        - We have NMR section in CHMO, also nmrCV.
    - Preliminary draft: 09-11-2022 - nmrCV & CHMO shortcommings call for NMR ontology from Philip / Johannes
    - nmrCV would need an update w.r.t the RODK (?) in its github actions
    - Ontologies4Chem Workshop 2023 - Brainstorming Content
    - Discussion on NFDI4Chem maintenance of nmrCV, with nmrXiv contribution.
    - The pulse sequenece ontology
      https://openaccess.uoc.edu/handle/10609/126306
      Is also linked from an earlier document NMR spectroscopy

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
NFDI4Chem is supported by DFG under project number 441958208