Deliverable 4.2

Development of data handling and archiving protocols

of Remote NMR (R-NMR):

Moving NMR infrastructures to remote access capabilities

Authors: Óscar Millet (CICBIO), Tammo Diercks (CICBIO), Rubén Gil (CICBIO), Miquel Pons (UB)

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement N. 101058595

TECHNICAL REFERENCES

Project acronym:	R-NMR
Project Title:	Remote NMR: Moving NMR infrastructures to remote access capabilities
Grant Agreement number:	10105859
Project coordinator:	Prof. Dr. Harald Schwalbe
Organization:	J.W. Goethe Universität, Frankfurt
E-mail:	Schwalbe@nmr.uni-frankfurt.de
Project website address:	http://www.r-nmr.eu/

Deliverable No.:	D4.2
Lead Beneficiary:	CICBIO
Type and dissemination level:	Report - Public
Due Date:	M30 (31 December, 2024)
Delivery Date:	27 December, 2024

CONTENTS

1. Introduction

The R-NMR project is building upon a foundation that has already established critical protocols for managing the initial stages of NMR experiments, from the submission of project requests through to the admission of samples at NMR facilities. These earlier stages, particularly the admission phase, have been well-documented in previous tasks and deliverables, such as **D3.1**, which standardizes operating procedures for remote NMR measurements and sample shipment. This initial work includes ensuring that samples are received, inspected, and stored properly, with all necessary metadata captured regarding their condition, type, and handling requirements. The metadata gathered in this phase ensures that the sample is ready for subsequent NMR measurement phases, such as pre-acquisition, acquisition, and post-acquisition.

However, to fully support the following three phases (**Pre-Acquisition Settings**, **Data Collection**, and **Post-Acquisition Operations**), there is a pressing need to develop a comprehensive data model that handles the growing complexity of NMR experiments. This data model will serve as the backbone for organizing, storing, and transferring the large volumes of raw data generated during the experiments, along with the rich metadata required to contextualize and reproduce each experiment. Without such a model, there is a risk of inconsistency in how data is captured, stored, and shared across the various participating NMR facilities.

Another key consideration for ensuring the long-term findability and accessibility of NMR data is the assignment of **Digital Object Identifiers (DOIs)** to relevant datasets. DOIs provide a persistent, unique identifier that allows datasets to be located and cited independently of their physical storage location, whether on institutional servers, repositories like Zenodo, or other platforms. Incorporating DOIs into the data management framework aligns with FAIR principles, enhancing the discoverability, traceability, and reusability of research outputs, while also facilitating proper attribution in academic and collaborative environments.

A critical aspect of ensuring the integrity and reusability of datasets is the classification of individual experiments as 'valid' or 'non-valid'. This classification is based on predefined criteria such as proper calibration, correct experimental parameters, and the overall quality of the collected data. Not all experiments will be valid—for instance, some may be preliminary tests or suffer from misconfiguration—but tagging them appropriately ensures that only high-quality data is prioritized for further analysis and DOI assignment. Non-valid data can still be retained

for troubleshooting or internal review, providing a valuable feedback loop for process improvement.

This validation process is key to ensuring that only high-quality data is linked to the overarching experimental framework. At the heart of this model is the **unique Sample ID**, which connects the admission phase with the subsequent phases. This identifier will ensure that all the metadata collected during the admission of the sample (such as sample type, storage conditions, safety requirements, and experiment goals) can be linked directly to the data generated in the later stages of the workflow. The Sample ID will allow users and facilities to seamlessly trace the journey of each sample, ensuring that all relevant information is available at every stage of the experiment.

The data model developed in **Task T4.2** must meet several critical requirements to align with the goals of the R-NMR project. First, it must be designed in accordance with **FAIR principles** (Findable, Accessible, Interoperable, Reusable) to ensure that data can be easily discovered, accessed, and reused by both local and remote users. Additionally, the model must comply with **GDPR regulations** to ensure that all personal and sensitive data associated with experiments—such as information about the users and sample origin—are handled securely and ethically. This becomes especially important in cases involving human-derived samples or sensitive proprietary research.

Beyond compliance, the data model must be **flexible and scalable**, able to support not only the **standardized protocols** defined in **D4.1**, but also future special applications. These protocols provide a comprehensive framework for conducting NMR experiments across different facilities and ensure that the experimental setup and data acquisition are consistent and reproducible, while also allowing for **site-specific modifications**. This flexibility is essential, as different facilities may require adaptations based on their specific hardware or software configurations. The data model must be designed to accommodate these modifications while maintaining a consistent core structure. For instance, in the area of metabolomics, which will serve as a case study for special applications, the model must accommodate data from complex, high-throughput experiments that generate large datasets of metabolic profiles. Furthermore, it should be adaptable to other future experimental needs that arise as the project evolves, ensuring long-term utility.

To achieve these goals, this deliverable will build on the foundational work of other deliverables and milestones. Key components, such as the standard operating procedures defined in **D3.1**, the standardized protocols from **D4.1**, and the technical requirements for remote spectrometer

access outlined in **D3.3**, will inform how data and metadata should be structured and managed. Furthermore, milestones related to data protection and sample handling from **WP2** and **WP3** will be referenced to ensure that the data model aligns with ongoing developments in those areas. This deliverable will ensure a seamless integration of data handling protocols, creating a robust and interoperable framework for the management of NMR data across the consortium.

2. Overview of data and metadata generated in each phase

Pre-Acquisition Settings Phase

Once the sample has passed the admission phase and its **Sample ID** is assigned, the process moves into the **Pre-Acquisition Settings** phase. During this step, several key pieces of data and metadata are generated to ensure that the sample is correctly prepared and that the instrument is configured appropriately for the upcoming NMR measurement.

Project and Funding Metadata

For effective management and transparency, it is important to associate each experiment with the relevant **research project**, whether it is externally funded or an internal initiative. This metadata should include:

- **Project Association**: The project(s) associated with the experiment. This could include internal research projects, collaborative initiatives, or externally funded projects (e.g., Horizon Europe, national research programs). This ensures the experiment can be traced back to its respective research goals, whether formalized with external funding or internal resources.
- **Funding Source**: The source of funding for the project, where applicable. This could include specific funding agencies (e.g., European Union, national research funds) or internal funding sources. This metadata is key for auditing and ensuring that infrastructure usage aligns with the project's financial guidelines.

Sample Preparation Data

- **Sample ID**: The unique identifier assigned to the sample, ensuring it is tracked throughout the experiment.
- **Sample Preparation Details**:
  - Type of sample preparation: Whether the sample is placed in an NMR tube (for solution NMR) or packed into a rotor (for solid-state NMR).
  - Sample labeling scheme.
  - Sample volume, concentration, and any buffer or solvent used (for solution NMR).
  - Information on rotor type and packing method (for solid-state NMR).
- **Storage Metadata**: If there is a delay between preparation and measurement, storage conditions (e.g., temperature, light sensitivity) are recorded.

Instrument Setup Metadata

- **Probe Type**: The type of probe used for the experiment (e.g., solution-state NMR probe or MAS rotor for solid-state NMR).
- **Temperature Settings**: The temperature to be maintained during the experiment, ensuring the correct conditions for the sample.
- **Shimming Information**: The configuration and optimization of the magnetic field homogeneity.
- **Calibration Data**: Calibration details, including tuning and matching of the probe, and RF pulse calibration settings for the nucleus being observed (e.g., ^{1}H, 13C, 15N).

Experiment Configuration Metadata

- **Experiment Type**: The type of NMR experiment to be performed (e.g., 1D ^{1}H NMR, 2D NOESY, solid-state NMR).
- **Planned Acquisition Settings**: Number of scans, pulse sequences, and expected acquisition time for the measurement.
- **Experiment Schedule**: The time and date when the experiment is expected to take place, as well as the facility or spectrometer being used.

- **Standard Protocol/Application Type**:
  - **Standard Protocol ID**: The specific protocol or application type being used, as defined in **D4.1** (e.g., Metabolomics Standard, Structural NMR for Proteins, Solid-State NMR Protocol). This identifier ensures that the experimental setup follows the standardized procedures developed in **D4.1**, which includes the protocols for both liquid- and solid-state NMR platforms.

Data Collection Phase

After the sample is prepared and the instrument is set up, the process moves into the **Data Collection** phase, where the actual NMR measurement takes place. This phase generates both raw data and metadata, which are essential for ensuring the accuracy, reproducibility, and traceability of the experiment.

Raw Data

- **Free Induction Decay (FID) or SER Files**: These are the primary outputs of the NMR spectrometer and represent the unprocessed signals captured during the experiment. These files are typically stored in proprietary formats, such as Bruker's FID or multi-dimensional SER files, depending on the type of experiment being performed (e.g., 1D, 2D, or 3D NMR).
- **File Size and Location**: The size of the raw data file and its location (whether stored locally or transferred to a centralized repository for further processing and analysis).

Acquisition Metadata

- **Sample ID**: The unique identifier that links the collected raw data back to the sample, ensuring that the data can be accurately associated with the correct sample and experiment.
- **Acquisition Parameters**:
  - Pulse sequence used (e.g., NOESY, COSY, ^{1}H NMR).
  - Number of scans and spectral width.
  - Acquisition time and delay times between scans.

- Receiver gain and digital filtering applied during the experiment.

- **Instrument Settings**: Detailed information about the spectrometer, including:
  - Spectrometer model and exact probe, including serial numbers.
  - Field strength (e.g., 400 MHz, 600 MHz).
  - Temperature during the experiment.
  - RF power levels and pulse widths applied for the nucleus being observed (e.g., ^{1}H, ^{13}C, etc.).

- **Shimming and Tuning Information**: Data on the quality of the magnetic field homogeneity and the tuning of the probe to the appropriate frequency.

Timing and Environmental Metadata

- **Date and Time**: The exact date and time when the experiment was conducted, ensuring that the experiment can be traced back in time.

- **Room Conditions**: Any environmental factors that could affect the experiment, such as room temperature and humidity (if applicable), especially relevant for sensitive samples.

Spectrometer Calibration Metadata

- **Probe Calibration**: Details of the calibration performed before data acquisition, including the tuning and matching of the probe.

- **Calibration Standards**: If any reference or calibration standards were used during the experiment (e.g., for chemical shift referencing or quantification purposes), these must be recorded.

Post-Acquisition Operations Phase

In the **Post-Acquisition Operations** phase of Task T4.2, the focus shifts from raw data to **processed data**, which can undergo multiple levels of additional processing depending on the application. The raw data collected during the experiment is first processed into usable spectral data, but this processed data can itself serve as input for further analysis, potentially leading to

new processed data and results. This chain of processing can continue through several stages, each generating its own metadata to ensure traceability and reproducibility.

Processed Data and Reprocessing

- **Initial Processed Data**: The raw data collected (e.g., FID or SER files) undergoes initial processing to produce spectral data. This includes steps such as Fourier transformation, phase correction, and baseline correction. The resulting processed files (e.g., 1r, 2rr) allow for visual and quantitative analysis of the NMR signals.
- **Reprocessing of Processed Data**: Processed data can serve as input for further processing using specialized software. For example:
    - In **metabolomics**, processed spectra can be used in software tools (e.g., Chenomx, Bayesil) to quantify metabolites by identifying peaks corresponding to known compounds.
    - In **structural biology**, processed 2D or 3D NMR data can be used in chemical assignment and structure determination software (e.g., CYANA, CARA) to generate protein or nucleic acid structures from NMR-derived distance restraints and chemical shift assignments.
- **Chain of Processing**: Each level of processing generates new **processed data** that may feed into further analysis. This can become an iterative process, where data from one processing step serves as the basis for new insights or as input for additional software tools.

Combining Multiple Experiments and Processed Data

- **Multiple Experiments per Sample**: A single sample may be subjected to multiple types of NMR experiments (e.g., 1D, 2D NOESY, HSQC). Each experiment generates its own raw and processed data, but these datasets can also be combined in subsequent processing steps. For example, combining data from different experiments can provide a more comprehensive picture, such as correlating chemical shifts from multiple dimensions in structural biology or combining different spectral data in metabolomics.
- **Combining Processed Data**: In some cases, processed data from different experiments or analysis steps can be combined to generate a final result. This could involve using data from multiple processed spectra in a single quantification step (e.g., combining multiple peaks for accurate metabolite quantification) or merging different datasets to refine protein structure determination.

Processing Metadata

- **Processing Steps**: For each processing step, detailed metadata must be captured, including:
  - The specific processing techniques applied (e.g., Fourier transformation, phase correction).
  - Parameters used during processing (e.g., window functions, pulse sequences).
  - Any manual adjustments or corrections made during the processing.
- **Software Information**: Metadata must include details on the software and version used at each stage of processing, whether it is an initial processing tool (e.g., Bruker TopSpin) or a secondary tool for further analysis (e.g., CYANA for structure determination or Chenomx for metabolite quantification).
- **Link to Previous Data**: Each processed data file must be linked back to the raw data or previously processed data from which it was derived. This ensures a clear chain of provenance, allowing users to trace the lifecycle of the data through multiple stages of processing.
- **Combining Data Metadata**: When processed data from different experiments or analysis steps are combined, metadata documenting the combination process must be captured. This includes which datasets were merged, how they were aligned or normalized, and the rationale for combining them.

3. Analysis of Implementation Alternatives

As we move forward with Task T4.2, we are confronted with two major challenges that are central to the successful implementation of the R-NMR data management model:

1. **How to organize the data and metadata into a flexible and scalable data model** that can accommodate raw data, processed data, and the potentially complex chain of data reprocessing across different NMR experiments.
2. **How to efficiently capture data and metadata in an automated or semi-automated way** and ensure that it is correctly entered into the data repository without burdening users with excessive manual input.

In this section, we will explore several alternatives for addressing these two challenges, considering their strengths, weaknesses, and suitability within the context of the R-NMR project.

Organizing Data and Metadata in a Flexible Data Model

The data and metadata generated throughout the NMR experiment lifecycle—including raw data, processed data, and further reprocessed datasets—requires a robust data model that is both structured and adaptable. Below, we examine three potential database models for organizing this information.

Relational Database (SQL) with Flexible Schema

Overview: A traditional relational database, such as **PostgreSQL** or **MySQL**, organizes data into structured tables with well-defined relationships between them (e.g., linking samples to experiments, raw data to processed data).

- **Advantages**:
  - Strong data integrity, with well-structured relationships between datasets.
  - Highly efficient for querying structured data (e.g., retrieving all data linked to a Sample ID or a specific experiment).
  - Mature ecosystem with plenty of tools for managing, querying, and visualizing data.
- **Challenges**:
  - Rigid schema design that may not easily accommodate new types of data or metadata, especially as new experimental protocols evolve.
  - Limited flexibility for handling unstructured or hierarchical metadata (e.g., nested processing steps).

NoSQL Document Database (e.g., MongoDB)

Overview: A NoSQL document-based database like **MongoDB** stores data in flexible, schema-less formats (e.g., JSON), making it ideal for handling varying metadata across different experiments or facilities.

- **Advantages**:
  - Highly flexible and scalable for storing unstructured or semi-structured data.
  - Can easily handle the diverse and hierarchical nature of metadata (e.g., raw data processed multiple times or across different tools).
  - Great for capturing new types of experimental data without needing to redesign the schema.
- **Challenges**:
  - Less efficient for handling highly structured data, which might lead to more complex queries for relational data like linking samples to experiments.
  - Querying and indexing can become complicated with large volumes of unstructured data.

Hybrid Model (SQL + NoSQL)

Overview: A hybrid approach combines the strengths of both SQL and NoSQL databases, where structured data (e.g., Sample ID, experiment metadata) is stored in an SQL database, while flexible metadata (e.g., experiment-specific processing parameters) is stored in a NoSQL document-based system.

- **Advantages**:
  - Balances structured and unstructured data handling: SQL for relational data and NoSQL for hierarchical metadata.
  - Scalability and flexibility for future experiments, while maintaining data integrity for core structured data (e.g., samples, experiments).
- **Challenges**:
  - More complex implementation, requiring synchronization between SQL and NoSQL databases.
  - Integration between the two systems needs to be seamless to avoid data fragmentation.

Automating the Capture of Data and Metadata

Once we define the data model, the next challenge is **how to capture data and metadata efficiently and automatically** throughout the NMR experiment lifecycle, minimizing manual

input while ensuring that all essential information is stored in the repository. Below, we consider three potential approaches for data capture.

Integration with NMR Instrument Software

Overview: Directly integrating with NMR software platforms (e.g., **Bruker TopSpin**, **Mnova**) to automatically capture raw data and metadata from experiments, such as acquisition parameters and sample information.

- **Advantages**:
  - Captures core metadata (e.g., pulse sequences, scan settings) automatically, reducing user input.
  - Ensures that all critical metadata is consistent with the actual experiment setup.
- **Challenges**:
  - Requires developing or leveraging APIs for each NMR software platform, which may vary by manufacturer or software version.
  - Limited flexibility if certain metadata fields are not natively captured by the instrument software.

Automated Metadata Extraction Scripts

Overview: Using scripts (e.g., **Python**) to extract key metadata directly from raw data files (e.g., FID, SER) or processed data files, generating the necessary metadata fields dynamically.

- **Advantages**:
  - Automates the extraction of essential metadata, such as acquisition settings or processing parameters, from existing data files.
  - Flexible solution that can be adapted to multiple file formats (e.g., Bruker, JEOL).
- **Challenges**:
  - Requires maintenance of scripts for each file format and updating them as formats or metadata requirements evolve.
  - Some data (e.g., user input, special experiment conditions) may still require manual entry.

User Input Forms with Metadata Automation

Overview: Creating user-friendly web-based forms where users can input key experimental metadata, while automation tools populate other fields (e.g., sample ID, storage conditions) based on templates or previous entries.

- **Advantages**:
  - Allows users to quickly provide critical metadata that cannot be automatically extracted (e.g., specific sample conditions, custom experimental objectives).
  - Reduces manual entry errors by using templates to populate standard metadata fields.
- **Challenges**:
  - Some user input is still required, which could lead to inconsistencies if the input is not standardized.
  - Less automated than direct integration with NMR software.

Assigning DOIs to Enhance Data Findability and Traceability

The integration of Digital Object Identifiers (DOIs) into the R-NMR data management framework represents a powerful solution to ensure the findability, traceability, and citation of datasets. DOIs serve as permanent, unique identifiers that provide a reliable way to locate and reference data, regardless of where it is physically stored. This approach is especially beneficial in multi-institutional and remote-access contexts, where datasets may be distributed across various repositories or servers.

Key Benefits of DOIs

- **Persistent Identifiers:** DOIs offer a permanent link to datasets, ensuring accessibility even if the storage location changes.
- **Improved Citability:** Researchers and collaborators can properly cite datasets in publications, enhancing the scientific impact of the project.

- **Alignment with FAIR Principles:** DOIs directly support the "Findable" and "Reusable" aspects of the FAIR principles, facilitating data sharing and reuse.
- **Interoperability:** DOIs are widely supported by repositories like Zenodo, Dryad, and institutional servers, enabling seamless integration with existing infrastructures.

Implementation Strategy

To implement DOIs in the R-NMR framework, the following steps are recommended:

- **Integration with Existing Repositories:** Partner with repositories such as Zenodo or institutional platforms that support DOI minting. These platforms can automatically assign DOIs when data is uploaded.
- **Metadata Synchronization:** Ensure that metadata linked to each dataset complies with DOI standards, including information about the data origin, authorship, and versioning.
- **Automated DOI Generation:** Develop scripts or workflows within the data management system to automatically assign DOIs to datasets deemed relevant by predefined criteria (e.g., datasets linked to published results or large-scale experiments).
- **User Awareness and Training:** Educate users about the purpose and use of DOIs, including how to cite datasets and retrieve data using DOI links.

Challenges and Considerations

While implementing DOIs offers significant benefits, there are some challenges to address:

- **Cost of DOI Minting:** Some repositories or platforms charge for DOI assignment, which may require budgeting and funding allocation.
- **Data Selection Criteria:** Not all datasets need DOIs; defining criteria for relevancy is essential to avoid unnecessary overhead.
- **Metadata Quality:** High-quality metadata is critical for ensuring that DOIs remain useful for long-term discoverability and access.

Classifying Experiment Validity within Sample Workflows

In the context of NMR experiments, each SampleID can be associated with multiple experiments, ranging from preliminary tests to final measurements. Not all experiments yield valid results due to factors such as incorrect calibration, parameter misconfigurations, or unexpected experimental conditions. To ensure that data workflows remain accurate and transparent, it is crucial to establish a system for classifying the validity of each experiment independently.

Proposed Solution: Experiment-Level Validity Classification

1. **Defining Validity at the Experiment Level:**
    - Valid experiments meet predefined quality criteria, such as:
        - Correct calibration and shimming during the run.
        - Accurate acquisition parameters.
        - Absence of significant noise or artifacts in the data.
    - Non-valid experiments may include:
        - Preliminary test runs.
        - Experiments with known errors (e.g., incorrect temperature or misaligned pulse sequences).

2. **Tagging Experiments within Sample Workflows:**
    - Each experiment conducted under a SampleID should be tagged as:
        - **Valid:** Meets all quality criteria and is ready for further analysis or DOI assignment.
        - **Non-Valid:** Does not meet criteria but is retained for troubleshooting or protocol refinement.
    - These tags should be stored in the metadata of each experiment to facilitate tracking and filtering.

3. **Filtering and Querying Data by Experiment Validity:**

- Implement filtering mechanisms to allow users to:
  - Retrieve only valid experiments for analysis or sharing.
  - Review non-valid experiments to identify and address recurring issues.

4. **Conditional DOI Assignment at the Experiment Level:**
   - Assign DOIs only to datasets generated by valid experiments, ensuring the integrity of publicly shared data.
   - Maintain traceability by linking non-valid experiments to their associated SampleID for internal reference.

Benefits of Experiment-Level Classification

- **Granular Control:** Allows users to manage data at the level of individual experiments rather than entire SampleIDs.
- **Improved Data Quality:** Ensures that only valid experiments contribute to downstream processes or publications.
- **Enhanced Troubleshooting:** Retains non-valid experiments for internal analysis and iterative improvement of workflows.

Challenges and Mitigation

- **Automating Experiment Classification:** Develop tools to automate preliminary checks, such as verifying calibration and acquisition parameters.
- **User Oversight:** Allow users to manually review and finalize the validity tags, supported by clear guidelines and intuitive interfaces.
- **Integration with Existing Systems:** Ensure that experiment-level tags are seamlessly integrated into the metadata structure of the repository.

Implementation Pathway

To implement this system, the following steps are recommended:

1. Integrate automated quality checks into the data acquisition workflow to suggest preliminary validity tags for each experiment.
2. Provide users with an interface to review and confirm or adjust the validity tags.
3. Link experiment-level tags to DOI assignment workflows and data filtering mechanisms within the repository.

Managing Project and Funding Metadata

In the context of centralized infrastructures and research facilities, it is essential to link each experiment to its corresponding project, whether internally funded or supported by external grants. The goal is to provide transparency and ensure that the use of resources can be traced back to the projects and funding that support them. The following steps should be considered:

- **Project Linkage**: While not all experiments will be linked to externally funded projects, every experiment should be associated with a project (internal or external) to provide context for its purpose and scope.
- **Source of Funding**: For projects that are externally funded, the source of funding (e.g., Horizon Europe, national grants) should be recorded. This helps ensure compliance with funding conditions and facilitates auditing of project resources.
- **Flexibility for Internal Projects**: Internal projects, which may not have a formal funding source, should also be registered with a project identifier, ensuring that the experiment is tied to a specific initiative, even without external funding.

4. Use cases

General case: spectral data handling system at UB

Figure 1 shows a schematic view of the spectral data handling system in the University of Barcelona. Its main characteristics are described below.

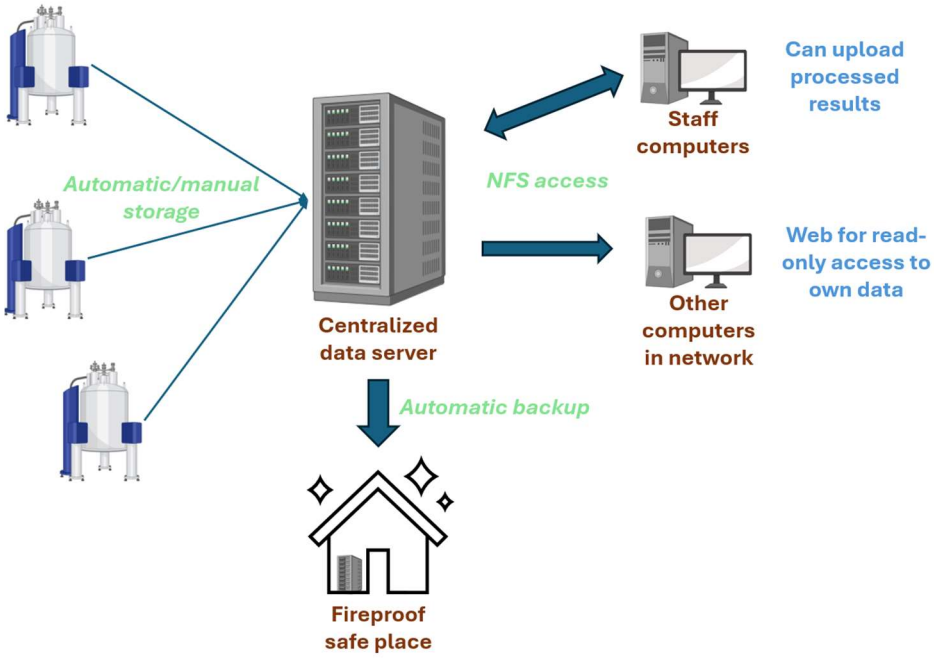

Figure 1. Schematic view of spectral data handling system in the University of Barcelona.

Centralized Data Storage

- All final NMR data collected on instruments within the Scientific and Technical Services of the University of Barcelona are stored on a centralized data server.
- This server is highly protected against data loss or manipulation and ensures the retention of data for a period of 10 years.

Automated Data Storage

- Data storage is automatic for spectra collected in automatic mode or by facility staff.
- Autonomous users are encouraged to store their data on the server, though this is optional.

Data Protection

- To protect against accidental data loss, the system utilizes a double backup mechanism with checksum verification.
- One of the backups is physically stored in a fireproof safe located away from the spectrometers.

Protection Against Data Manipulation

- Access to the data server is restricted to prevent unauthorized data manipulation. Write access is highly limited.
- Data is accessible only via read-only NFS connections from two separate servers:
  - One server is accessible exclusively by facility staff. Through this server, staff can upload transformed spectra (e.g., for users lacking specific processing software, such as for NUS acquired spectra), but cannot alter raw data.
  - The second server is used to generate specific web pages for read-only access to individual datasets. Users receive a link to access and transfer only the data pertaining to their request.

Data Confidentiality

- Each dataset is linked to registered users or user groups, as defined in the service request forms, ensuring user-specific confidentiality.
- Confidentiality during data transfer is maintained by avoiding the use of cloud-based intermediate storage.

Current Limitations

At present, the Barcelona system is not configured for public access or reuse of stored data. For example, it does not allow users to search for examples of specific pulse sequences used on a particular instrument or track repeated experiments on user samples across different time points to monitor instrument performance changes (beyond the standard tests routinely conducted on standard samples).

NMR Metabolomics Workflow at CICBIO

Figure 2 shows a summary of this workflow. Typically, CICBIO receives urine or serum/plasma samples, which are stored at -80°C. Each tube comes with its origin code, which is used during analysis. Although not having a unique code might result in duplicate codes between samples, this does not pose a problem as such cases belong to different projects and

will not be analyzed together. Ideally, however, we should assign a unique code upon receipt of each sample.

On the day the metabolomics measurements are to be performed, the first step is to calibrate the spectrometer, which is assumed to have the appropriate probe already installed. Calibration involves verifying the correct temperature, ensuring accurate quantifications, and confirming that the magnetic field is homogeneous (shimming). All of this is managed using the **TopShim** software, once the calibration samples are loaded: methanol for temperature calibration, QuantRef (a prepared artificial sample) for quantification, and sucrose for shimming. These tests are executed, and the results are saved in the calibration folder, which in our case is "IVDr/RefData/nmr". This folder contains subfolders for each of the three parameters evaluated. Folder names include the date, and within each, a small report summarizes the calibration results—for example, whether the temperature is correct, if the quantifications are accurate, and whether the magnetic field is properly shimmed. The parameters and artificial samples differ depending on whether the calibration is for urine or serum/plasma.

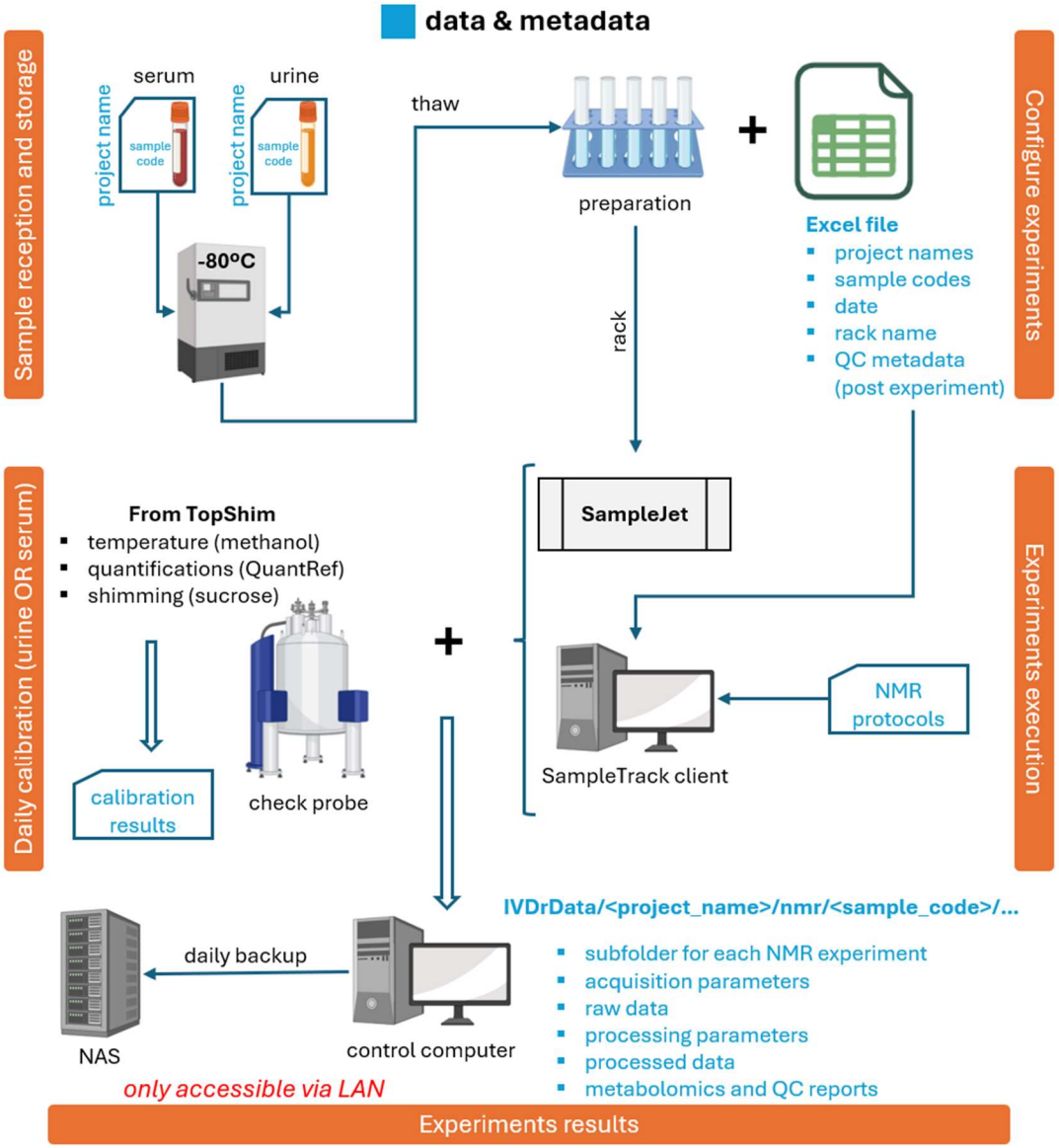

Figure 2. Overview of NMR metabolomics workflow at CICBIO

If the spectrometer is determined to be properly calibrated for metabolomics measurements on that day, the samples to be measured are selected and thawed. An Excel template is then filled out, listing the code of each sample and specifying the exact position where the NMR tube will be placed in the rack after sample preparation. The Excel file also includes a project name, which will become the name of the folder on the magnet containing the results of all analyzed samples. This project name is entered manually and usually consists of a short identifier for the project, along with the measurement date. The Excel file also includes the rack name that holds the samples and has several columns that will be filled in during preparation and after the analysis:

1. **Correct Sample**: Filled in at the end to indicate whether the measurement was successful.
2. **Alanine**: Indicates if the alanine peak is correctly identified in the right position.
3. **H2O**: Indicates if water suppression was correctly performed.
4. **Comments**: Used for additional notes, such as the sample's appearance before measurement or any other relevant information.

Some of these fields are filled in using information from quality control reports generated by Bruker during the measurement process.

Once the Excel file is complete and the samples are in their rack on the spectrometer, the file is uploaded to the **SampleTrack Client**. The user selects the protocol to use from several predefined options and initiates the process. For example, there is one protocol for urine and another for serum/plasma. The protocol includes a series of NMR experiments (e.g., 1D NOESY, JRES) along with subsequent processing, which in the case of metabolomics includes contacting the Bruker server for metabolite quantification. For urine samples, software that quantifies 150 metabolites is used, whereas for serum/plasma samples, three software tools are used: quantification of 41 metabolites, quantification of 112 lipoproteins, and quantification of parameters related to **PACS** (Post-Acute Covid Syndrome). The quantification reports are generated in PDF and XML formats, as are the quality control reports.

Thus, each day of metabolomics analysis on a spectrometer generates a folder named with the project and date. Within this folder is the "nmr" subfolder, which contains a subfolder for each analyzed sample. The sample folder names are derived from the names provided in the Excel file. Each sample folder contains a subfolder with a numeric code for each NMR experiment conducted. For serum/plasma, "10" typically represents NOESY, "11" is for CPMG, and so on. Each experiment folder contains the raw data and acquisition parameters, often organized into "pdata" subfolders for processed data. The main folder for NOESY ("10") usually contains the quantification and quality control reports in PDF format, while the "pdata/1" subfolder holds the processed data, processing parameters, and XML reports.

Metabolomics data are stored on the computers associated with the magnets where they were generated, under the "IVDrData" folder. An automatic incremental backup of all spectrometer

data is performed nightly on a NAS. Users who need access to the data must be connected to the internal network to retrieve it directly from the magnet computers or the backup.

Custom metadata database at IBS Grenoble (NMRlib)

NMRlib is a set of tools developed by the Institut de Biologie Structurale (IBS) in Grenoble, France, specifically designed for Bruker spectrometers running TopSpin versions 3.5 to 4.0. It provides a framework for setting up and managing NMR experiments, aiming to streamline the process of configuring pulse sequences and ensuring consistency across different spectrometer setups. The toolset is compatible with a variety of magnetic fields, making it adaptable for multi-instrument laboratories, and includes functionalities for both solution-state and solid-state NMR applications.

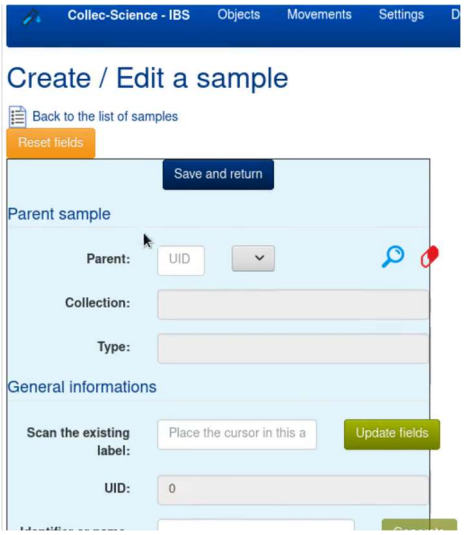

Figure 3. Screen to create a new sample in NMRlib

The NMRlib system is installed on spectrometer computers, with its database hosted on an online server accessible through a browser via HTTPS. Access is securely managed with a username and password. When creating a new sample (Figure 3), users are presented with a form featuring several fields to fill out, including dropdown menus offering various options such as the reference user (Figure 4). Free-text comments can also be added for additional context.

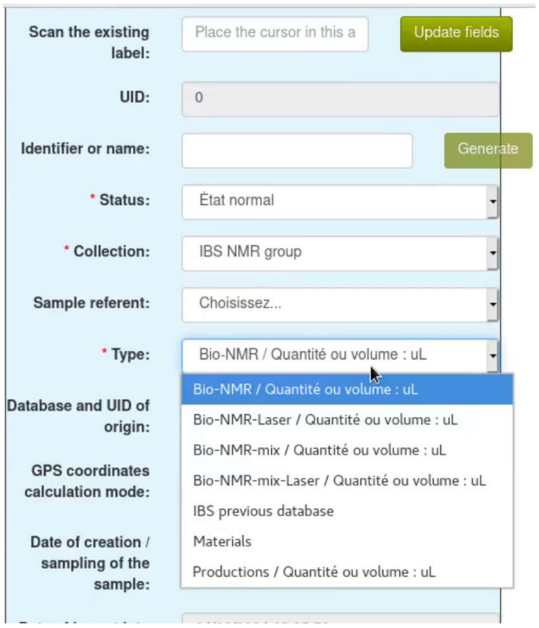

Figure 4. Example of available fields in NMRlib

The platform includes powerful search forms that allow users to filter and locate samples efficiently (Figure 5).

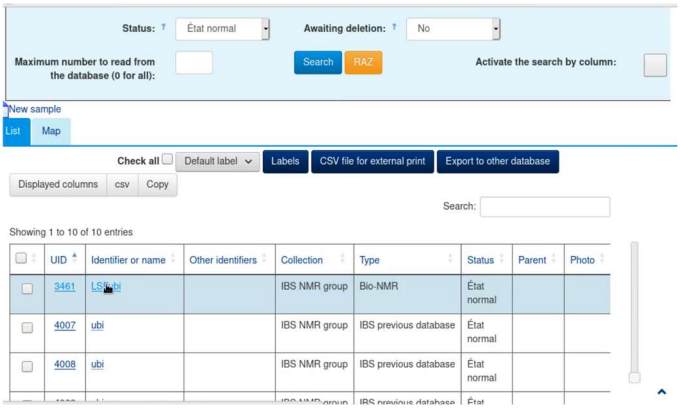

Figure 5. Search form and results in NMRlib

Moreover, the information stored for a sample in the database can be seamlessly transferred to the **SAMPLE** tab within Topspin (Figure 6), streamlining data handling and experiment setup.

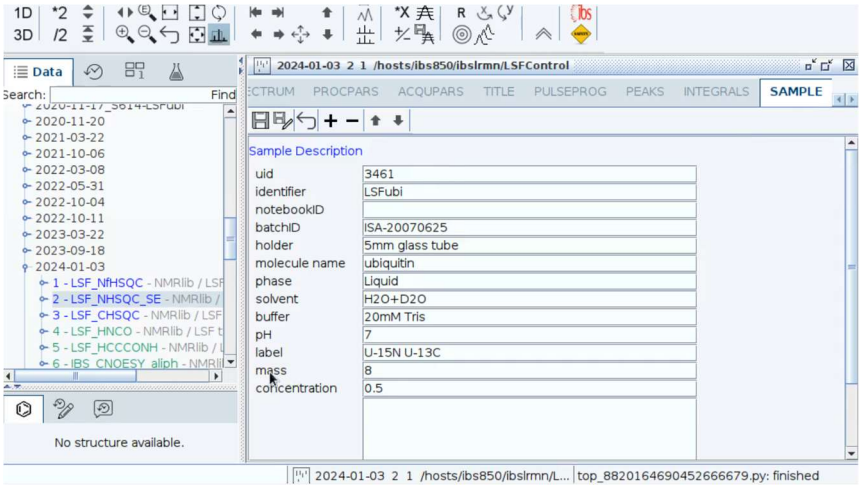

Figure 6. Example of information integrated into TopSpin

Additional information can be found here:

https://www.ibs.fr/en/communication-outreach/scientific-output/software/nmrlib-2-0-ibs-pulse-sequence-tools-for-bruker-spectrometers
(Accessed on November, 2024)

Commercial solution: LOGS – A Scientific Data Management Platform

LOGS is a versatile scientific data management platform that facilitates the automatic upload, management, and communication of research data generated by laboratory instruments (Figure 8). Designed as a web-based system, it provides login-based access to research data via a browser and enforces a fine-grained permissions system based on user roles and project memberships. The platform is highly adaptable, being installable on PCs with moderate specifications and hostable on virtual machines or cloud services.

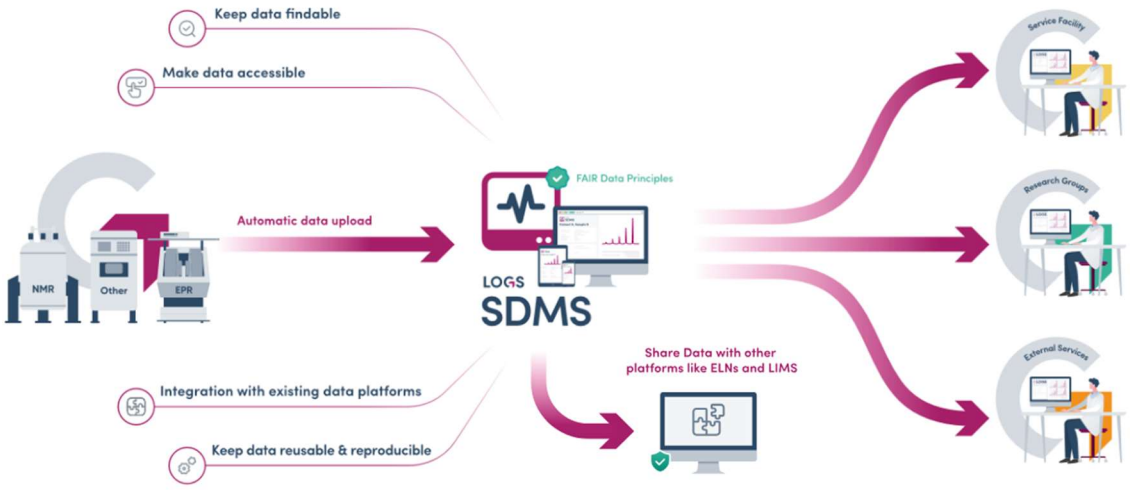

Figure 8. Schema of use of LOGS system

Key Features

- **Data Search and Metadata Integration**: LOGS supports connected metadata and text search through the browser or Python API. Uploaded data formats are automatically recognized and parsed, with metadata extracted directly from the files and additional sources like directory pathways or filenames. All metadata is stored in a PostgreSQL database.
- **Data Upload Options**: Datafiles can be uploaded using a "push" mechanism via the LOGS Data Bridge—a lightweight client program that transfers data from spectrometers to the LOGS server. Alternatively, a "pull" mechanism can be employed, allowing LOGS to fetch data using sFTP.
- **Integration Capabilities**: LOGS complements other platforms through REST and Python APIs, enabling data integration with systems like ELNs (Electronic Lab Notebooks) and LIMS (Laboratory Information Management Systems).

Advanced Metadata Handling

LOGS generates metadata automatically from datasets and configurable sources such as filenames, text files, and directory paths. This ensures comprehensive and accurate metadata capture with minimal manual effort.

Data Sharing

LOGS enables secure and efficient data sharing in two ways:

1. **Collaborator Accounts**: Collaborators can access shared datasets with role-based permissions.
2. **Link Sharing**: Scientists can generate a secure, password-protected link for specific datasets, which leads to a dynamically created webpage accessible without requiring a LOGS account.

Use Cases

LOGS serves as both an internal database for research groups and a collaborative tool for sharing data within consortia, offering seamless and secure data management for diverse research workflows.

Additional information can be found here:

https://logs.sciy.com/
(Accessed on November, 2024)