NFDI4Chem, Chemistry Consortium in the NFDI

Deliverable D3.3.1

Gap analysis report for selected repositories

(status 2021)

Authors of this deliverable: Felix Bach[1#], Kunigunde Binder[2], Christian Bonatto Minella[2], Benjamin Lutz[1], Matthias Razum[2]

[1] Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen

[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen

[#]now at FIZ Karlsruhe

Funded by the German Research Foundation (DFG)

Project number. 441958208

Research program: national research data infrastructure (NFDI)

Content

Executive Summary

The deliverable 3.3.1 “Gap analysis report for selected repositories” aims both to identify gaps in the coverage regarding data types or disciplines and to close them through adjustments or, if necessary, new developments. In order to accomplish that, the TA3-team performed a gap analysis of the existing relevant repositories by means of individual interviews with the repository leaders. The interview consisted of a series of questions ranging from general information up to metadata standards and ontology, data contents, technical information about Authorisation and Authentication Infrastructure (AAI), API, services and functionality, operating environment as well as software architecture and workflows. The interviews will serve to establish the current degree of maturity as well as the operational fitness of the selected repositories and to derive suitable recommendations aiming to fulfil the yet missing requirements.

Project objectives

With this deliverable, the project has contributed to the following objectives:
Key Objective 1 “Virtual environment of federated repositories” and 3 “Smart Laboratory Environments”

Background

Within the framework of the NFDI initiative and especially of the NFDI4Chem consortium, Task Area 3 (Repositories, TA3) aims to establish a **virtual environment of federated repositories** for collecting, storing, processing, analysing, disclosing and re-using research data as a part of the NFDI4Chem infrastructure. The federation will be realised by implementing standards regarding metadata and Application Programming Interface (API) defined in TA2 (Smart Lab) and TA4 (Standards) as well as vocabularies and ontologies addressed in TA6 (Synergies). This will result in the integration within the Smart Lab environments including Electronic Laboratory Notebook (ELN) in TA2.

In order to identify those repositories which can form the nucleus for the envisioned virtual federation, the TA3-team developed a list of criteria which is reported below:

- The repository is suitable for the deposition of molecule related data
- The repository contains reusable data or functionality that covers the needs of the NFDI4Chem community
- The repository software is open source
- The operators of the repositories have declared their willingness to adapt their services to the standards developed by NFDI4Chem including the FAIR principles
- The repository operators can be funded in accordance with the funding guidelines of the NFDI (i.e., the main operator is based in Germany and is a non-profit organisation)

The application of these criteria led to the following list of “selected repositories” in 2019

- Chemotion Repository
- nmrshiftdb2
- MassBank EU
- VibSpecDB
- Suprabank
- NOMAD
- STRENDA DB

Most repositories do not yet meet all of the above mentioned criteria needed to be reliably integrated into the NFDI4Chem service platform. However, some of them already play an important role in chemistry research.

TA3 will try to support further databases and data repositories on interoperability issues and encourage them to participate in the development of NFDI4Chem standards and interfaces.

Since they have already stated their commitment to comply with the NFDI4Chem standards, TA3 identified **CSD** and **ICSD** as “associated repositories''. In addition, two other generic/multidisciplinary data repositories, **bwDataArchive** and **RADAR**, were selected because they fulfil three of the mentioned criteria and can play an important role both as catch-all repositories and data archiving services. Additionally, they can offer long-term archival functionality for the other selected repositories where needed.

Aiming both to identify gaps in the coverage regarding data types or disciplines and to close them through adjustments or, if necessary, new developments, the TA3-team performed a gap analysis of the existing relevant repositories by means of individual interviews with the repository leaders. The interview consisted of a series of questions ranging from general information up to metadata standards and ontology, data contents, technical information about Authorisation and Authentication Infrastructure (AAI), API, services and functionality, operating environment as well as software architecture and workflows. The interviews will serve to establish the current degree of maturity as well as the operational fitness of the selected repositories and to derive suitable recommendations aiming to fulfil the yet missing requirements.

In the next section (“description of the work”) the collected information for each repository (referred to the beginning of 2021) is reported in the form of repository profiles.

Description of Work

This section describes the current status with regard to organisational, functional and technical features of the repositories addressed.

Repositories based on process and analysis data

Chemotion - Repository for molecules and research data

The Chemotion Repository, created in 2014, covers research data that is assigned to molecules, their properties and identification as well as reactions and experimental investigations and is hosted at Karlsruhe Institute for Technology (KIT). It is productively used at several locations in Europe. Scientists in the domains of chemistry, molecular chemistry, or organic molecular chemistry are supported in their efforts to handle data in a FAIR manner: the project data is stored along with molecule and reaction specific identifiers and Digital Object Identifier (DOI)-assigned data files are given with distinct ontology-supported (Chemical Information Ontology (CHEMINF), Chemical Methods Ontology (CHMO) and Name Reaction Ontology (RXNO)) metadata. The findability of the data is achieved by a text and structure search and its availability via PubChem. The repository is interoperable with the Chemotion ELN with respect to data transfer and offers export schemes to other systems. Data is curated by automatic checks and a peer reviewing process. The integration of data stored in the repository in publications was shown with several examples and its usage is currently recommended by “Chemistry methods”. Authors can be referenced by Open Researcher and Contributor iD (ORCID iD). Chemists, materials scientists, and biologists as their target audience can publish data for open access (data view) and registered access (dataset contribution and download). Stored data can be searched by chemical structure, author, dataset type, status, or identifier. The current AAI solution is based on an internal user administration (administrator, anonymous and registered user, curator). Metadata according to DataCite is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) scheme. The Chemotion Repository offers an internal substance register, a spectra viewer (ChemSpectra), a structure editor (Ketcher) and their own data converter.

Quick stats:

- **Number of datasets:** 2442 Samples, 734 Reactions, 7338 Analyses, 7338 of them are published
- **Listed in:** re3data: r3d100010748, FAIRsharing: FAIRsharing.iagXcR, bio.tools: chemotion, Wikidata: Q98391181, PubChem: Data Source ID: 1195
- **Accepted data types:** DX, JCAMP, MzMl, MzXML (open, visualisable and processable), RAW (processed and converted in JCAMP)
- **Used standards/ontologies:** DataCite Metadata Schema, InChi, SMILES, CHEMINF Ontology, CHMO Ontology, RXNO Ontology
- **Access rights/licence information/embargo:** CC0, CCBY, CCBY-SA; embargo possible (unlimited)
- **Used Software:** https://github.com/ComPlat/chemotion_REPO

Future activities for improvements contain the following points: the operators of Chemotion wish to enrich the repository by gaining the Core Trust Seal (CTS) certification as well as to be recommended by all relevant journals within their scientific domains. A general API and, especially, a machine learning interface for data export and retrieval are of particular interest. In addition, their tools portfolio is planned to be extended by NMRium (as a viewer), by Marvin (as a structure editor) and by an integrated data converter. On top of that, the improvement of their data archive by a version control system is desired.

Spectroscopy-based repositories

nmrshiftdb2

nmrshiftdb2 was created in 2010 and is hosted at University of Cologne (UzK). It contains datasets covering more than 50,000 molecules. Those molecules' structures, chemical shifts (1D and 2D NMR spectra, so far mainly for nuclei 1H and 13C) and assignments can be accessed for search (e.g. dereplication, similarity, fragments, structure, spectra, signals) or prediction. The data is curated via QuickCheck based on CSEARCH-Robot-Referee. The repository is targeted at scientists, students, and commercial end-users in the fields of chemistry, molecular chemistry, natural sciences, physical and theoretical chemistry, and bioinformatics. Researchers can publish data for open access (data view) and registered access (dataset contribution). More recently, deposition of raw data and electronic assignments in NMReDATA format were included and entries receive DOI's to facilitate citation in publications. Data can be exported in various formats, including as an NMReDATA file. The nmrshiftdb2 database can be installed as a local repository in NMR laboratories to improve integrating the workflow of academic chemistry groups via optional Laboratory Information Management System (LIMS) functionality. The provided tool suite consists of a viewer, a structure editor, a data converter, a data processing and analysis tool, and a prediction tool (preditorc.jar for carbon shifts, predictorh.jar for hydrogen shifts). Additional tools were developed to assist and evaluate spectra assignment to further reduce the barrier of electronic data processing. It is part of the DFG-funded project "Initiative to Improve on Data Quality in NMR Spectroscopy" (IDNMR). Their current AAI solution is based on an internal user administration (data provider, curator, reviewer, administrator). They offer a web interface and a Web Service API.

Quick stats:

- **User documentation:** https://nmrshiftdb.nmr.uni-koeln.de/portal/js_pane/P-Help
- **Number of datasets:** 53842, all are published
- **Listed in:** re3data: r3d100010316 (nmrschiftdb), FAIRsharing: FAIRsharing.nYaZ1N, Wikidata: Q24170714, PubChem: Data Source ID: NMRShiftDB
- **Accepted data types:** NMReData, Bruker
- **Used standards/ontologies:** DataCite, internal NMReDATA tag format
- **Access rights/licence information/embargo:** nmrshiftdbs database licence based on ODC Open Database License (ODbl); no embargo

- **Used Software:** https://sourceforge.net/p/nmrshiftdb2/code/HEAD/tree/trunk/nmrshiftdb2/doc/developerguide.rtf?force=True

The software base of nmrshiftdb2 is rather old and difficult to maintain. Therefore a complete software re-engineering is planned. A new repository called nmrXiv will replace nmrshiftdb2 in the future.

The results of the interview identified the following desired developments: as in the case of Chemotion repository, the operators of nmrshiftdb work in the direction of a CTS repository certification as well as towards an increasing support by all relevant journals in their target domains to recommend their repository. Funding support to attain these goals is planned through NFDI4Chem. Topics of improvements are represented by the minimum information standard (from TA4) as the metadata standard (nmrCV ontology) and ORCID ID as the contributor identification system. They intend to expand their target user group specifically by organic chemists and NMR spectroscopists. Concerning AAI solutions, nmrshiftdb operators plan to follow the NFDI4Chem recommendations/requirements, combined with the possibility of embargos, the “external reviewer” as a new user role, and validation as part of the review process. The intention is to work on the extension of their offered export formats and to include more viewers for structures, spectra, and metadata. Future plans envision nmrXiv as the main part for data processing and analysis (NMRium based) as well as to have the nmrXiv steering committee as the technical project lead. Additional goals are the digital preservation of data from the ‘Digitale Bibliothek Thühringen’ and the establishment of a version control system for this very data.

MassBank EU - High Quality Mass Spectral Database

MassBank EU was created 2010 based on a collaboration between Germany and Japan, is hosted at Helmholtz-Centre for Environmental Research (UFZ) Leipzig, and is the first public repository of mass spectral data for sharing them among the scientific research community. Their target user groups in the domains of chemistry and life sciences are analytical chemists, metabolomics, biochemists, and bioinformaticians. MassBank EU spectral data of known, unknown and provisionally identified substances are useful for the chemical identification and structure elucidation of chemical compounds detected by mass spectrometry. Their data sets from community users and projects are openly accessible and they represent the official database of the Mass Spectroscopy Society of Japan. They use GitHub as their current AAI environment (open read access, limited write access) and GitHub issues as their curation tracking system. The curation itself is performed by the MassBank record validator. The data sets can be searched for compound and mass spectrometry information and peaks and they use MassBank Record ID (Accession) and USI (Universal Spectrum Identifier) as persistent identifier systems. The MassBank EU spectral data is hosted in a revision control system with all spectral data and the corresponding metadata in a human-readable record format, and continuous integration (CI) checking record integrity for each change. Instances of the web interface are hosted at UFZ and Leibniz Institute of Plant Biochemistry (IPB) (Halle) and can be installed locally as well. They offer interfaces for data import via Git (MassBank record format) and for data export

(JSON-LD) and a REST API. On top of that they provide RMassBank as a separated data processing/analysis tool.

Quick stats:

- **Number of datasets:** https://massbank.eu/MassBank/RecordIndex
- **Listed in:** re3data: r3d100011839**,** FAIRsharing: FAIRsharing.dk451a, bio.tools: massbank, Wikidata: Property:P6689 Q24088019 , Identifiers.org: massbank, PubChem: Data Source ID: 23999
- **Accepted data types:** MassBank format
- **Used standards/ontologies:** Internal MassBank Record Format
- **Access rights/licence information/embargo:** Copyright; individual licences based on creative commons; no embargo
- **Used Software:** https://massbank.github.io/MassBank-documentation/developer_documentation.htmlhttps://github.com/MassBank/MassBank-web

Future activities are represented by gaining the CTS certification and by enabling the extension of their internal record format by using the Dublin Core specification and the Mass spectrometry Ontology (MS). The operators support the use of DOI as an identification system for citations and see mirroring of their Git repository as a possible solution for backups.

VibSpecDB

VibSpecDB is a currently internally utilised database for vibrational spectra (Raman and IR spectra) that is hosted at Friedrich Schiller University Jena (FSU). In the course of the project, VibSpecDB will be integrated into the NFDI4Chem spectroscopy concept and converted to full open access. Currently, the database itself features APIs for programming languages like Python or R, but no Graphical User Interface (GUI) based import routines, web-interface or viewers. These functionalities will be developed in the course of NFDI4Chem and a licence as well as access-right management system will be added to the database forming a repository for vibrational spectra.

Their goal is to be listed in public catalogues, such as re3data and to use CHMO and BioAssay Ontology (BAO) as ontologies. DOI should be used in the future as the identification system for citations and ORCID iD as the identification system for authors. The operators aim at a version-controlled data system that is searchable by properties such as experimental vs. sample data, data ownership, metadata or open user projects. Users should be able to share their data with other registered users. The implementation of the DFN-AAI and the establishment of the following user roles are intended developments: registered users and admin users on various access levels. The access itself should be organised via two-factor authorisation or access tokens (single user, user groups). Their wish is to realise a comprehensive Kubernetes setup including database backups on an external server, redundancy via Kubernetes replica pods with load balancers and a monitoring system in order to have a fail-safe repository. Import, export, and ingest are intended to be managed via a web User Interface (UI) component where the supported exports formats ought to be CSV, JSON and ISA. They also want to support APIs for Raman analysis tools, basic spectra analysers, format converters, and ELNs and connect viewers for spectra, data structures and metadata.

Subdomain-specific repositories

SupraBank

SupraBank has been hosted at KIT (Karlsruhe) since 2019 and is a curated database that provides project data on intermolecular interactions of molecular systems and supramolecular interactions which are not available in other repositories or databases. SupraBank is mainly aimed at supramolecular and physical chemists or biologists in the domain of organic chemistry who deal with binding, assembly, and interaction phenomena. Molecular properties are retrieved from PubChem, allowing the correlation of intermolecular interactions parameters to molecular properties of the interacting components. All molecules, solvents, and additives are searchable by their chemical identifiers. At present, the Suprabank stores more than 3500 curated data sets of intermolecular interaction parameters. The data has open access for viewing and registered access for data download and contribution. It can be searched for experiments and related components, molecule interactions, and publications while being curated by non-judgemental plausibility checks. The current implementation of AAI consists of internal user administration (anonymous and non-anonymous user, data provider, administrator). DOI is used as the identification system for citations and ORCID iD as the identification system for authors. The repository is implemented in Ruby on rails and uses a PostgreSQL database. Its web interface offers file format compatibility with CSV, JSON, BibTex, RIS, and Endnote and its tool suite contains molecule representations as pictures, a structure editor and a simulation modeller tool.

Quick stats:

- **Number of datasets:** 4000, 3700 are curated
- **Listed in:** re3data: r3d100013265 **,** FAIRsharing: bsg-d001818
- **Accepted data types:** JSON (DataCite), CDX (for 2D/3D molecule structure), PNG, proprietary formats
- **Used standards/ontologies:** DataCite 4.0, Dublin Core for metadata tags
- **Access rights/licence information/embargo:** CC licences (CC0, BY, BY-SA), embargo possible (unlimited)

Focus of the repository operators is a strategy to increase the daily number of accesses (from 40 to 1,000) as well as of data sets (from 4,000 to 20,000). Their data are intended to be published with DOI and their repository to be CTS-certified and subsequently recommended by relevant journals in their field. The operators plan to implement the usage of established ontologies and are currently waiting for NFDI4Chem recommended AAI solutions. An import- (e.g. JSON-LD) as well as an export-API (e.g. OAI-PMH via DataCite Metadata (MD) Store) belong to the wishlist. Furthermore, a version control system following DataCite standards is part of the future focus. Planned is also the transfer of the repository to the Steinbuch Centre for Computing (SCC) at KIT to secure the digital preservation and redundancy.

NOMAD Repository and Archive - Novel Materials Discovery

The NOMAD repository was created in 2006 and is hosted at Fritz Haber Institute of the Max Planck Society (FHI) Berlin. It enables the confirmatory analysis of calculated materials data, their reuse, and repurposing. It facilitates research groups to share and exchange their results, inside a single group or between two or more. All data are available in their raw format as well as in a common and machine-processable data format (NOMAD Archive), which is based on the internal data schema NOMAD Metainfo. The data can be used under the CC-BY 4.0 licence and published with a 3-year embargo. There is also the possibility to assign DOIs for citing the data. Users can search for the simulated system, the code and method used, and properties related to the data archived there. The data can be viewed and downloaded free of charge, but registration is required to contribute data. The repository provides a Restful HTTP API (NOMAD API) and a Python package for accessing the NOMAD archive as well as services such as the NOMAD Encyclopedia and the NOMAD Analytics Artificial-Intelligence (AI) Toolkit.

Quick stats:

- **Number of datasets:** 12.091.376 entries, 2.984.283 materials, 991 datasets
- **Listed in:** re3data: r3d100011583, FAIRsharing: FAIRsharing.aq20qn
- **Accepted data types:** RAW as ZIP or GZ (converted into NOMAD Archive by uploading)
- **Used standards/ontologies:** NOMAD Metainfo
- **Access rights, licence information, embargo:** CC-BY 4.0, embargo possible (up to 3 years)
- **User documentation:** https://nomad-lab.eu/prod/rae/docs/introduction.html

STRENDA DB - Standards for Reporting Enzymology Data

STRENDA DB is a repository operated since 2016 at Beilstein Institute (BI) Frankfurt for enzymology data providing the means to ensure that data sets are complete and valid before scientists submit them as part of a publication. Their target audience are biochemists, systems biologists, and biocatalysts in the fields of life sciences, chemistry, biological and food chemistry, and molecular chemistry. The typical data contained in this repository consists of functional enzymology data (kinetic and experimental data) from manuscripts and publications. Data manually entered in the STRENDA DB are automatically checked (according to STRENDA Guidelines), allowing users to receive notifications for necessary but missing information. The successful formal compliance is confirmed by notification, the assignment of a STRENDA Registry Number (SRN) and DOI and is documented in a fact sheet (PDF) containing all input data that can be submitted with the manuscript to the journal. Datasets assigned with a DOI allow reference and tracking of the data. The data become publicly available in the database only after the corresponding article has been peer-reviewed and published in a journal. Currently, more than 55 international biochemistry journals already include the STRENDA guidelines in their instructions for authors. DOI is used as the identification system for citations and ORCID iD as the identification system for authors. Data viewing is possible via open access and data contribution is possible after a required registration where the current AAI is provided through an internal user administration (user, administrator). The repository uses an Oracle database and is organised in a multilayer architecture that realises access for the user via a graphical HTML interface.

Quick stats:

- **Number of datasets:** 150
- **Listed in:** re3data: r3d100012329 , FAIRsharing: FAIRsharing.ekj9zx, Wikidata: Q58034053
- **Accepted data types:** Currently none, EnzymeML (in development)
- **Used standards/ontologies:** DataCite, InChI, EnzymeML (in development)
- **Access rights/licence information/embargo:** Creative Commons Attribution 4.0 International (CC BY 4.0), Term of Use; no embargo
- **User documentation:** https://www.beilstein-strenda-db.org/strenda/help/STRENDA_DB_UserGuide_v0.92.pdf

One of their future goals is to substantially increase the number of daily accesses (from 10 to 100) and end up with about 100 published data sets per year. The wish to provide users with an API and plan functional extensions addressing systems biology and applied biocatalysis needs. EnzymeML is their preferred data format and implementation for import, export and ingest. The operators would like to establish two new user roles such as the “reviewer” and the “editor”.

Multi-disciplinary repositories

RADAR - Research Data Repository

RADAR was created in 2017 and is hosted by Leibniz Institute for Information Infrastructure (FIZ) Karlsruhe. It serves as a catch-all repository and provides a cross-disciplinary data archiving and publishing (via DOI) service for any research data from completed scientific studies and projects. RADAR is available in different variants (RADAR Cloud, RADAR Hybrid and RADAR Local). It also serves as the basis for RADAR4KIT, a research data repository of the KIT launched in December 2020. RADAR aims to ensure access to and long-term availability of archived and published datasets according to FAIR criteria. RADAR is intended as a generic infrastructure component in several NFDI consortia. For the purpose of interoperability, it therefore takes into account data types recommended by the NFDI and supports discipline-specific metadata. The data within the repository are made findable via the repository's own search portal (metadata and DOI are searchable). Browsing the data is free of charge, but the archive and publication service are chargeable. A repository's own rights and role model is used for authorisation management; authentication is supported via DFN-AAI (Shibboleth) and OAuth 2.0. Author identification is ensured by using ORCID iD and funder identification by CrossRef Open Funder Registry. Metadata are recorded using the internal RADAR Metadata Schema (based on DataCite Metadata Schema 4.0). Data can be published with an embargo period (between 1 and 12 months) and must be provided with a user licence. Curation of data can be performed using RADAR's own configurable curation workflow, which includes an optional peer review step. Further functionalities on offer include a web interface for data ingest and download, a REST API, a DataCenter API for integration with data archive systems (Secure File Transfer Protocol (SFTP)), a REST interface to DataCite (Fabrica) and an OAI-Provider for metadata harvesting. The frontend is currently implemented in Groovy on Grails and the backend in Java. Their persistence layer is based on a combination of the Apache Cassandra database and Elasticsearch as a search engine. Archival is realised via HPPS and Spectrum Project, whose configuration is dependent on specific users and operation models.

Quick stats:

- **Number of datasets:** 110 published datasets (plus archived datasets not available for public access)
- **Listed in:** re3data: r3d100012330
- **Accepted data types:** All data types/formats (format recommendations exist)
- **Used standards/ontologies:** RADAR Metadata Schema (based on DataCite Metadata Schema 4.0), Dublin Core, schema.org
- **Access rights/licence information/embargo:** Terms and conditions for both data providers and data users, mandatory licences for datasets (e.g. Creative Commons), embargo period (1-12 months)
- **User documentation:** https://radar.products.fiz-karlsruhe.de/en/

bwDataArchive

bwDataArchive was created between 2013 and 2016 and is hosted by KIT Karlsruhe. Like RADAR, it serves as a catch-all repository and provides a cross-disciplinary long-term data archiving service for any scientific data (and stores RADAR's data). The access requires a user registration and a contract with KIT for the affiliated organisation. Their AAI implementation is based on the bwIDM and the DFN-AAI. Persistent identifiers are managed internally and authors can be identified via ORCID iD. The repository uses a custom backend built on IBM High Performance Storage System (HPSS). The user access regarding data transfers is performed via SFTP. A web-based user frontend is written in PHP and allows users to manage personal data in their account.

Quick stats:

- **Number of dataset:** file based, > 800.000.000 files
- **Accepted data types:** any
- **Used standards/ontologies:** none
- **Access rights/licence information/embargo:** none (no public access)

Analysis of the repositories

The analysis provides an overview of identified gaps and designates recommendations to close these gaps. The analysis is based on inquiries on the repositories' homepages, supplemented by the results of an initial repository workshop, the aim of which was to introduce each repository and identify topics considered in need of optimisation. Further information was obtained through interviews with the repository operators, which addressed topics such as content covered, metadata standards and ontologies, as well as technical aspects such as operating environment, software, interfaces, authorisation and authentication, and additional services/functionalities. In a similar manner, user profiles have been recorded by TA5 that could complement our analysis in the future but are not part of the current document.

The following topics were identified as relevant for the gap analysis:

Authorisation and Authentication Infrastructure (AAI)

One prerequisite for the planned virtual environment of federated repositories is the establishment of a common AAI to regulate user authorisation and authentication. Four (Chemotion, nmrshiftdb2, SupraBank and STRENDA DB) out of six repository operators interviewed stated that they do not currently have such a federated permission system in use but do use an internal user management system through which different roles can be assigned, such as anonymous and registered user, administrator or curator. They also stated that they are currently waiting for joint solutions by NFDI in this regard. One repository (MassBank EU) uses GitHub or rather GitHub accounts for rights (grants read access uncontrolled and for roles with write access a GitHub account is needed) and role (anonymous and registered user, administrator) management. VibSpecDB uses a standard built-in Laravel authentication system to manage access rights for registered users and administrators. With regard to AAI, it should be noted that there is currently no common solution for the entire NFDI. This has yet to be established. Possible solutions are being discussed within the NFDI section “Common Infrastructures'' and as an important cross-cutting topic, a common AAI may be delivered by a potential future basic-service consortium.

Regarding the AAI/SSO (authentication and authorisation infrastructure/single-sign-on) services, promising candidates were identified: a cooperation with the German National

Research and Education Network (DFN) is considered necessary. Beyond that, other various services such as ELIXIR- and the Helmholtz-AAI are also considered of interest.

Worth mentioning are also activities such as the ”Survey on AAI in the NFDI service landscape” which aims to provide an overview of existing experiences, current and future requirements regarding an AAI for services in and around the NFDI. The results of the questionnaire will support, coordinate and prioritise AAI activities within the NFDI.

Nevertheless, as claimed in the “Sektionskonzept Common Infrastructures zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V.” the NFDI infrastructure will rely on the integration of the preliminary work accomplished in 2021 by a subgroup of the NFDI Task Force Tools.

Interfaces

In addition to a common AAI, APIs are needed for various purposes (see Figure 1). Among other things, interfaces that enable data/metadata -import and -export are of importance, but also an interface between ELN “Chemotion” and federated repositories to ensure the data flow. Furthermore, an interface to the terminology service must be provided to ensure access to the terms listed there, as well as interfaces from each service component (e.g., storage, long-term preservation “LZA”) and search systems or software applications (e.g. as ELN, editors or viewers), and to the AAI. Five (Chemotion, nmrshiftdb2, MassBank EU, SupraBank and STRENDA DB) out of six repository operators interviewed indicated that they provide an interface (via HTML interface or Git) for data import. VibSpecDB plans to implement a corresponding interface. Three (Chemotion, nmrshiftdb2 and MassBank EU) of the six repository operators interviewed provide an interface (via HTML interface or OAI-PMH) for data/metadata export. The remaining repositories (SupraBank, STRENDA DB and VibSpec) plan to implement such an interface. Chemotion has additional interfaces to the Chemotion ELN, to the PubChem database (RESTful API), to an internal substance register and to the ChemSpectra viewer. MassBank EU also offers a Restful API, which is also planned for SupraBank. Nmrshiftdb2 additionally offers a Web Service API whereas VibSpecDB plans to implement APIs for Raman analysis tools, format converters and ELN.

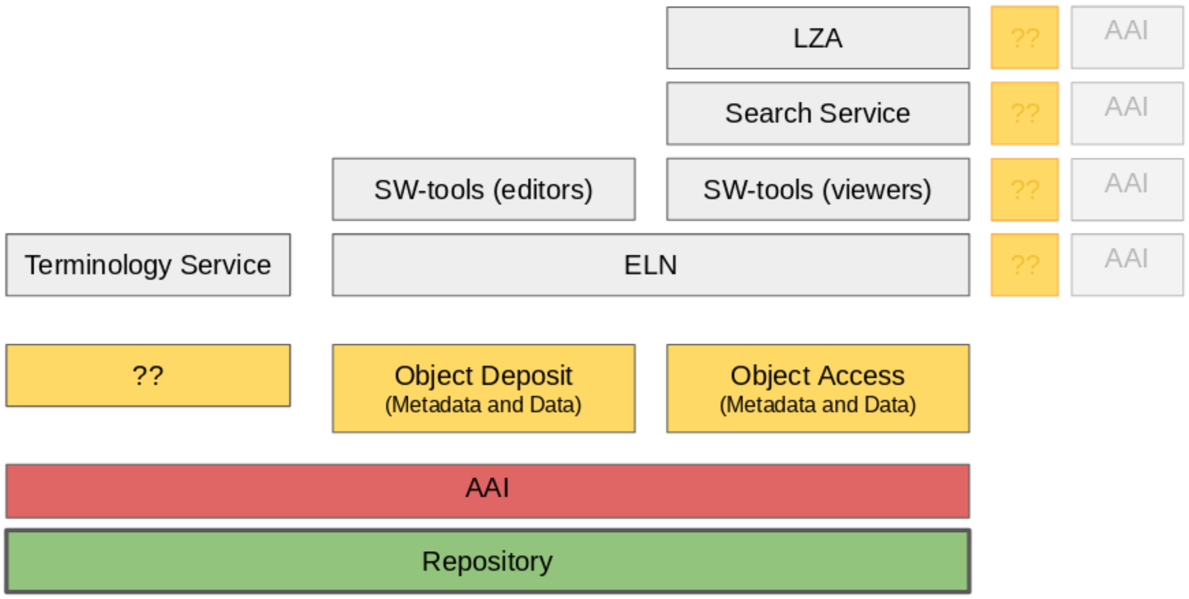

Figure 1. Identified interfaces (yellow) that are required to connect repositories and services.

We recommend the establishment of common interfaces for:

- the exchange of data and metadata between the federated repositories
- the exchange of data and metadata between the federated repositories and the Chemotion ELN
- the exchange of metadata between the federated repositories and the NFDI4Chem Terminology Service
- the exchange of metadata between the federated repositories and the NFDI4Chem Search Service (i.e. for metadata harvesting)
- connecting to a common AAI service
- connecting to a common LZA service (provided e.g. by generic services like RADAR4Chem or bwDataArchive)
- connecting to additional services such as viewers, editors and analysis tools

Metadata standards/Ontologies

The use of common metadata standards and ontologies - referred to as Minimum Information (MI) Standards in the context of NFDI4Chem - is crucial for the exchange of metadata between the federated repositories and the central search service. Therefore, the federated repositories must implement these MI Standards for content that is to be made searchable by the search service. Furthermore, the use of norm data such as ORCID iD is another prerequisite for achieving cross-searching and interoperability beyond the boundaries of NFDI4Chem. Four (Chemotion, nmrshiftdb2, SupraBank and STRENDA DB) out of six repository operators interviewed use the generally accepted metadata standard DataCite for collecting descriptive metadata. One (Chemotion Repository) out of six repositories also uses standards such as InChI and SMILES as well as the CHMO and RXNO ontology for capturing subject-specific metadata. The InChi standard is also used by another repository (STRENDA DB), which also has the EzymeML specification in the implementation phase. Two out of six repositories use internal description formats for annotating their data (nmrshiftdb2: NMReDATA Tag Format) and sdf-File (chemical structure in MOL format); MassBank EU: MassBank Record Format). VibSpecDB is still in the process of developing a corresponding description standard. Regarding the provision of a personal identification system, it can be confirmed that three (Chemotion, SupraBank and STRENDA DB) of six repositories provide ORCID iD, two (nmrshiftdb2 and VibSpecDB) plan to do so and one (MassBank EU) has no contributor identification system in place. SupraBank additionally uses the Research Organisation Registry ID (ROR ID) for identifying affiliations.

We recommend the reuse of the MI (among other things DataCite) defined in NFDI4Chem for data annotation. Related to this aspect, a possible approach could be the development of an overarching modular description standard that covers all (meta)data information relevant for the NFDI4Chem community and ensures their context-dependent compilation through the given modularisation within the repositories. In addition, the use of authority data such as ORCID iD guarantees interoperability across NFDI4Chem boundaries.

Digital Preservation

A long-term preservation service offered within the federated repositories enables the long-term and secure preservation and usability of the data available there. To this end, it is necessary to ensure the preservation of data at the technical level (bitstream preservation: preserving at the physical level and ensuring data integrity by monitoring and exchanging storage media), at the level of data formats (local preservation: preserving technical interpretability and data authenticity by migrating data formats) and at the level of content (semantic preservation: preserving interpretability and discoverability of data by describing data using metadata). Furthermore, it is important to ensure the possibility of referencing data via permanent identification systems such as DOI. Of six repository operators interviewed, four (Chemotion, nmrshiftdb2, MassBank EU and STRENDA DB) have committed to long-term archiving, SupraBank is planning to do so and VibSpecDB has not decided yet. Chemotion Repository ensures long-term preservation by connecting to bwDataArchive and nmrshiftdb2 via the digital library Thüringen. MassBank EU uses GitHub and Zenodo whereas SupraBank is planning a connection to SCC in this regard. Concerning the provision of persistent identifiers (PIDs), it can be confirmed that three (Chemotion, nmrshiftdb2 and STRENDA DB) of six repositories provide DOI for citing data, two (SupraBank and VibSpecDB) are planning to do so and one (MassBank EU) uses internal identifiers (MassBank Record ID (Accession), USI (Universal Spectrum Identifier)). In nmrshiftdb2, in addition to DOI, a Handle can also be selected.

We recommend the long-term preservation for data worthy of archiving to ensure their long-term availability in a generally understandable format. A common solution can be perhaps realised via RADAR4Chem. Also of importance is the use of a PID system, e.g. as an established standard such as DOI, to ensure the permanent citability of the data within the federated repositories.

Public Relations

The outreach is of central importance in order to generate and increase the awareness of the repositories in the community. Six repositories (Chemotion Repository, MassBank EU, NOMAD, SupraBank, STRENDA DB and RADAR) use social media channels such as Twitter or YouTube to reach their target audience. Four repositories (Chemotion Repository, NOMAD, STRENDA DB and RADAR) offer tutorial videos on their homepage or via their YouTube channel to increase their user-friendliness. Further dissemination activities conducted by the repositories include provision of publications on the repository and newsletter, networking, as well as participation in conference talks and workshops.

We recommend creating synergies in outreach, e.g. by developing joint strategies for website design (e.g. in terms of user-friendly design by providing a user guide and glossary, offering newsletter or producing promotional or tutorial videos) or social media channels (e.g. using NFDI4Chem's or Chemotion's Twitter or YouTube channels), and involving TA5 for further support. In addition, it may be beneficial to highlight the association to the NFDI4Chem initiative and/or to use the NFDI4Chem branding to promote one's repository and its functionalities.

Our analysis has shown that the selected repositories are in good shape in terms of the quality of their software development and processes. It should be noted that of the analysed repositories, one (nmrshiftdb2) is currently being revised due its outdated status and one (VibSpecDB) is currently still in the development phase. The greatest need to catch up is in the area of authorisation and authentication, but this may also be due to the fact that the repository operators are waiting for a common solution proposed by the NFDI. The support of PID systems for persistent data citation and personal and organisational identification, the use of community-approved metadata standards and ontologies for describing the data, and the provision of the data and associated metadata via an API are aspects where there is room for improvement, but which are already planned or being implemented by the repository operators. In this regard, it is important to ensure that the compatibility of the repositories with each other and with community-approved standards is guaranteed.

Next steps

Future efforts should also focus on community needs to avoid the development of undesired functionalities. Taking into account current activities in the analysis and comparison of user profiles will play a crucial role in this.

In the next months, in order to create synergies within the consortium, the TA3-team will focus on promoting the exchange among the operators of the relevant repositories but also with the other TAs. To this end, the TA3-team will also organise regular meetings and will establish working groups for a direct exchange that will help to close identified gaps more quickly. If required, we will also support the operators of the repositories in implementing functionalities that have been missing so far (e.g. AAI or APIs). With this strategy, we believe that the repository profiles will be completed and the virtual federation environment will be operational as planned.

References

The references are embedded and linked directly in the text.

Appendix

Interview Quick-reference

AAI, infrastructure for authentication and authorisation:

4 of 6 not using any permission system, instead using internal user management systems

1 of 6 using GitHub (via GitHub accounts)

1 of 6 using Laravel permission system

Universal standards are missing at this point (for the whole NFDI)

Interfaces:

5 of 6 offering an import interface (HTML interface, Git)

1 of 6 planning an import interface

3 of 6 offering an export interface (OAI-PMH, HTML interface)

3 of 6 planning an export interface

Digital preservation:

4 of 6 offering digital preservation services

1 of 6 planning

1 of 6 unknown

Identification systems/ standard data:

3 of 6 using DOI

2 of 6 planning

1 of 6 using internal identifier

3 of 6 using ORCID

2 of 6 planning

1 of 6 not using any

Metadata standards:

4 of 6 using a metadata standard (DataCite)

1 of 6 using internal description format

1 of 6 planning