Repos4Chem - criteria for acquisition - for suggestion by NFDI4Chem for data providers

A list of criteria for the selection of repositories relevant to the chemistry community has been developed. Repositories will be assessed on their maturity with respect to secure and sustainable operations, reliable data publication and long-term data preservation and accessibility. They should meet the selected requirements and recommendations and, if they do not yet do so, declare their willingness to meet the other missing requirements. These repositories will be recommended in an article to be published in the NFDI4Chem Knowledge Base: the article will serve as a guide for repository managers to understand what criteria need to be met in order to be included in the NFDI4Chem repository federation. The criteria were categorised into mandatory, recommended and optional.

Mandatory Criteria

The repository

- is suitable for the deposition of **molecule related data**
- is **not exclusively an institutional** repository (e.g. the repository is free to use and provides open access to datasets and their metadata)
- uses **Persistent Identifiers** (PIDs) for datasets
- uses community accepted **metadata standards** (e.g. provide DOI with metadata in DataCite format).
- requires **registration** to deposit data **or** data are **curated** by experts
- software is **documented** and **accessible** for transparency
- **is registered** in re3data, FAIRsharing and/or Identifiers.org
- offers well **defined interfaces** (API, REST, UI) (e.g. to connect ELN, Viewer, Data-import/enrichment, -export/retrieval)
- has a **fail-safe hosting** plan, provides policies on data retention and has a plan for long-term management of data.
    - Redundant servers; Load balancing; Data replication; Backup systems; Monitoring and alerts; Disaster recovery plans
- is **hosted** at a **data centre** (to ensure reliable operation, data security and features such as backups, redundancy and fast network connectivity)

Recommended Criteria

The repository

- also uses **Norm Data Identifiers** for entities in metadata (such as for persons e.g. ORCID iDs, devices e.g. PIDINST, Institutions e.g. ROR).
- uses a contributor/researcher **persistent digital identifier** system (e.g. ORCID)

- software architecture is **modularised**, well **documented** and **accessible**
- uses **standards ontologies** (e.g. CHMO, RXNO)
- software is ideally **open source**
- the MD-harvesting over **OAI-PMH** or similar mechanisms (e.g. Schema.org embedded in dataset landing pages) is available
- contains reusable **data** and/oran ecosystem of software tools or libraries (such as viewers, editors or analytical tools) that fulfil the needs of the NFDI4Chem community
- employs well described **operational processes**
    - Repository Creation; Version Control; Access Control; Backup and Recovery; Repository Organization; Documentation; Continuous Integration/Continuous Deployment (CI/CD); Security Measures; Maintenance and Performance; Auditing and Compliance
- provides adequate **file and container formats**
- landing pages for the repository itself and also for the datasets are **machine readable**
- receives regularly/periodically **updates** and new features
- is **certified** (e.g. CTS)
- offers an **Authentication** and **Authorization Infrastructure** (AAI) solution
- provides defined **access rights**, **licence information**, and optional **embargos**
- provides **data acceptance criteria**
- defines **clear roles** such as administrator, curators and editors
- has a defined/dedicated **operational team**

In addition, we list optional criteria for repository managers to develop a plan for further features, depending on the scope of the repository:

Optional Criteria

- Spectral viewers
- Structure editors
- Data converters
- Tools for data processing/analysis
- Automated plausibility and quality checks/peer review
- Further services/functionalities
- Data archiving
- Versioning Dataset
- Versioning MD-Schema

Workflows:

- Long-term preservation
- Ingest
- Backups
- Data archiving
- Redundancy
- Resilience
- Monitoring
- Exit strategy
- Versioning Dataset
- Versioning MD-Schema
- Handling of orphaned records
- Identifier recycling

Safety level of hosting:

- Hosting & Operation Security
- DevOps

Interfaces:

- for data import/enrichment
- for data export/retrieval
- others/further
- export formats