

OPEN
ARTICLE

Linking Research Data with Physically Preserved Research Materials in Chemistry

Chia-Lin Lin¹, Pei-Chi Huang¹, Simone Gräßle¹, Christoph Grathwol¹, Pierre Tremouilhac¹, Sylvia Vanderheiden¹, Patrick Hodapp², Sonja Herres-Pawlis³, Alexander Hoffmann³, Fabian Fink³, Georg Manolikakes⁴, Till Opatz⁵, Andreas Link⁶, M. Manuel B. Marques⁷, Lena J. Daumann⁸, Manuel Tsotsalas⁹, Frank Biedermann¹⁰, Hatice Mutlu¹¹, Eric Täuscher¹², Felix Bach¹³, Tim Drees¹⁴, Steffen Neumann¹⁵, Shashank S. Hariviyasi¹, Nicole Jung^{1,16} ✉ & Stefan Bräse^{1,17} ✉

Results of scientific work in chemistry can usually be obtained in the form of materials and data. A big step towards transparency and reproducibility of the scientific work can be gained if scientists publish their data in research data repositories in a FAIR manner. Nevertheless, in order to make chemistry a sustainable discipline, obtaining FAIR data is insufficient and a comprehensive concept that includes preservation of materials is needed. In order to offer a comprehensive infrastructure to find and access data and materials that were generated in chemistry projects, we combined the infrastructure Chemotion repository with an archive for chemical compounds. Samples play a key role in this concept: we describe how FAIR metadata of a virtual sample representation can be used to refer to a physically available sample in a materials' archive and to link it with the FAIR research data gained using the said sample. We further describe the measures to make the physically available samples not only FAIR through their metadata but also findable, accessible and reusable.

¹Institute of Biological and Chemical Systems - Functional Molecular Systems (IBCS-FMS), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ²Institute for Biological Interfaces 3 - Soft Matter Laboratory (IBG 3 - SML), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ³RWTH Aachen University, Institute of Inorganic Chemistry, Landoltweg 1a, 52074, Aachen, Germany. ⁴RPTU Kaiserslautern-Landau, Department Chemie, Erwin-Schrödinger-Str. Geb. 54, 67663, Kaiserslautern, Germany. ⁵JGU Mainz, Department Chemie, Duesbergweg 10-14, 55128, Mainz, Germany. ⁶Universität Greifswald, Institut für Pharmazie, Friedrich-Ludwig-Jahn-Str. 17, 17489, Greifswald, Germany. ⁷LAQV-REQUIMTE, Department of Chemistry, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516, Caparica, Portugal. ⁸Chair of Bioinorganic Chemistry, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 13, 40225, Düsseldorf, Germany. ⁹Institute of Functional Interfaces (IFG), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ¹⁰Institute of Nanotechnology (INT), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ¹¹Institut de Science des Matériaux de Mulhouse UMR 7361 CNRS/Université de Haute Alsace 15 rue Jean Starcky, Mulhouse Cedex, 68057, France. ¹²Technische Universität Ilmenau, Institut für Chemie und Biotechnik, Weimarer Straße 25, 98693, Ilmenau, Germany. ¹³FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH, Hermann-von-Helmholtz-Platz 1, 76344, Eggenstein-Leopoldshafen, Germany. ¹⁴Legal Affairs, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ¹⁵Leibniz Institute of Plant Biochemistry, Computational Plant Biochemistry group, Halle, Germany. ¹⁶Karlsruhe Nano Micro Facility (KNMF), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ¹⁷Institute of Organic Chemistry (IOC), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131, Karlsruhe, Germany. ✉e-mail: nicole.jung@kit.edu; stefan.braese@kit.edu

Introduction

Sustainable work and provisioning of research results for others is an essential criterion for efficiency in scientific research. Only when results are accessible to the entire scientific community, and thereby reusable, can scientific progress be accelerated in a targeted manner. Since the publication of the FAIR data principles¹, more and more scientists and associated stakeholders such as funding agencies and journals/publishers support generation and provision of FAIR data across disciplines. Accordingly, data should be Findable, Accessible, Interoperable and Reusable (FAIR), especially the research data that form the basis of publications. In chemistry and materials science, a variety of initiatives promote provision of FAIR data or provide assistance with implementation of FAIR data measures. Established stakeholders that supported the concepts of FAIR data even before their explicit publication include IUPAC², RDA³ and CODATA⁴. These have been joined by other important groups such as EOSC⁵ and the National Research Data Infrastructure (NFDI)⁶ in Germany, in particular its consortium for Chemistry (NFDI4Chem)⁷. In future, NFDI4Chem plans to strengthen the chemical community by providing an infrastructure for generating and provisioning FAIR data. Adhering to the FAIR data principles and aligning research processes to obtain FAIR data can form the basis of substantial improvements in data availability and quality. Through transparency, FAIR data can also strengthen trust in research results and promote subsequent usage that is systematic and frictionless.

In synthetic chemistry, the FAIR principles must be extended beyond data and descriptive metadata i.e. synthetic chemists can be encouraged to provide more than data for documentation and subsequent use. In this domain, it is often possible for chemists to substantiate results of reactions in the form of products, and thus provide physical evidence for the research work and its quality. Where the reaction products obtained are stable, they can be collected, stored and registered so that, when suitable, the result can be used directly for further studies. Possible examples of such direct reuse are independent reproduction of experiments, use of these deposited chemical samples for reactions, and analysis of the samples by characterization or screening techniques. Numerous further scenarios are conceivable in which such stored substances could accelerate knowledge gain, especially when they are unambiguously linked to other research output like journal publications and research data. Additionally, such a reusable collection of samples would promote transparency.

Initiatives that provide access to scientific physical collections of materials exist already in other disciplines, such as Geosciences, Microbiology, Botanical Science and Natural History⁸, where collecting and archiving samples for further reuse is widely accepted as an important aspect of scientific work. Examples of such collections of physical samples include the scientific collections of the U.S. Geological Survey⁹ and the collections of drilling cores at the IODP (International Ocean Discovery Program) Core Repository¹⁰. Some of the physical collections aim to make their samples FAIR, using different concepts to define and establish suitable identifiers that were developed in the past. While many of the currently suggested procedures include the concept of International Generic Sample Number (IGSN)^{11,12} into their processes, others find related solutions to enable a straightforward implementation^{13,14}. These different initiatives have developed recommendations that may be a helpful guide to design physical archives that describe materials in a FAIR manner^{15,16}.

In chemistry, a few centers worldwide are making efforts to collect and store chemical substances for subsequent use. Exemplary well-known initiatives for the systematic collection of mostly commercial but also partly academic substances are the Compounds Australia¹⁷, EU-OPENSOURCE¹⁸, Chimiotheque National (ChemBioFrance)¹⁹, and the Boston University Center for Molecular Discovery (BU-CMD)²⁰. The known initiatives collect and register chemical substances for medical or pharmaceutical application purposes but, as far as we know, do not offer open, general subsequent usage of various kinds.

Thus, the work described in this article was started with the aim to provide for concepts and infrastructure that enable a sustainable model to collect, archive, and reuse the physical results of chemical research work and to connect them with existing research data infrastructure in chemistry.

Results

Principles and concept design. Scientific outcome in chemistry consists of data and materials and further studies also often depend on availability of both. Therefore, we suggest complementing the FAIR (meta)data principles with a concept for materials that address sustainable access to physical objects¹¹ such as chemical samples, allowing for reuse, wherever possible. The concept should make it possible to secure research results in their physical form, verify the results obtained and increase reusability of the materials in addition to the reuse of already well-established data. To this aim, chemical compounds or, more precisely, samples of chemical compounds – which are the starting point of analytical studies or the outcome of synthetic studies in chemistry – should be preserved and made available. As a first step, the metadata of samples would need to be Findable, Accessible, Interoperable, and Reusable (=FAIR metadata for samples). As a following step, the samples would need to be registered and stored in a materials' archive to be Findable, Accessible, and made available under suitable access policies and rules to be physically Reusable (=FAR material for samples). This concept is further referred to as “FAIR-FAR samples” (Fig. 1).

The FAIR-FAR sample concept in detail. Metadata for a sample can be seen as a virtual representation of the sample or, in other words, a digital object to which a globally unique and persistent identifier can be assigned. Metadata that are part of such a virtual sample representation include information on the sample's provenance, components/content, and properties. Most of the metadata assigned to such a virtual sample representation are relevant for a FAIR data approach if it comes to a comprehensive description of analytical measurements of chemical samples. This is because chemistry samples can usually be described with standardized metadata that is referred to by both, the physical sample as well as the data obtained by analyzing it. Efficient concepts that consider the deposition of FAIR research data and the deposit of FAIR-FAR samples could, therefore, closely link both of them. A design to implement this is depicted in Fig. 2, showing a virtual sample representation described

Sustainable data and material handling in chemistry

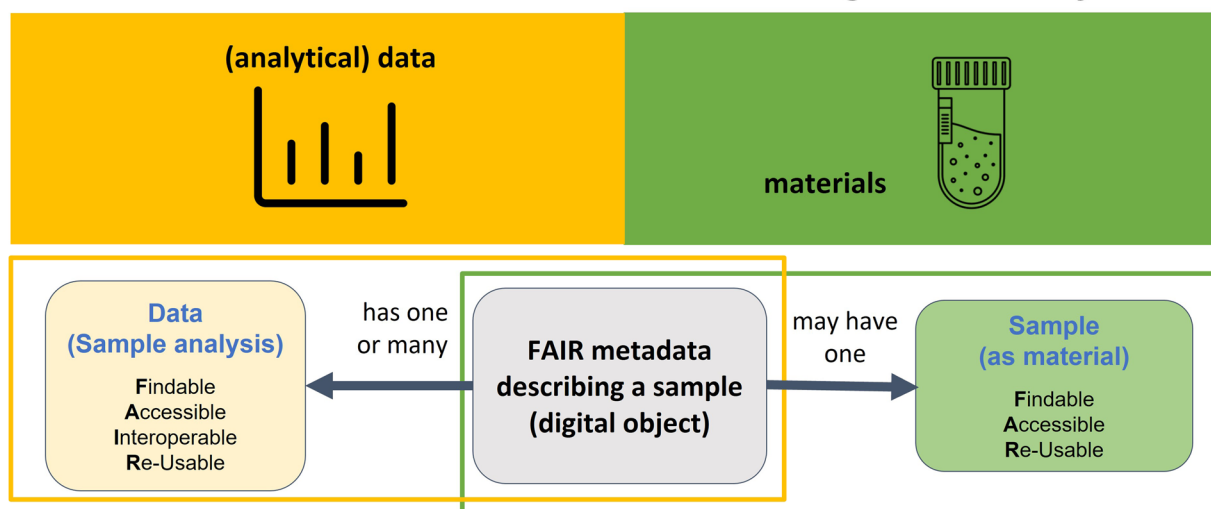


Fig. 1 Extending the principles of FAIR data (usually considered in chemistry is the left part in yellow including FAIR metadata for a sample) with a concept for findable, accessible and reusable samples consisting of the same FAIR metadata of a sample and additional measures meeting the requirements of materials' collection, storage and provision (right part in green). [ref Icons 2: Tube: *Digithrust from the noun project*, Data: *Popular from the noun project*].

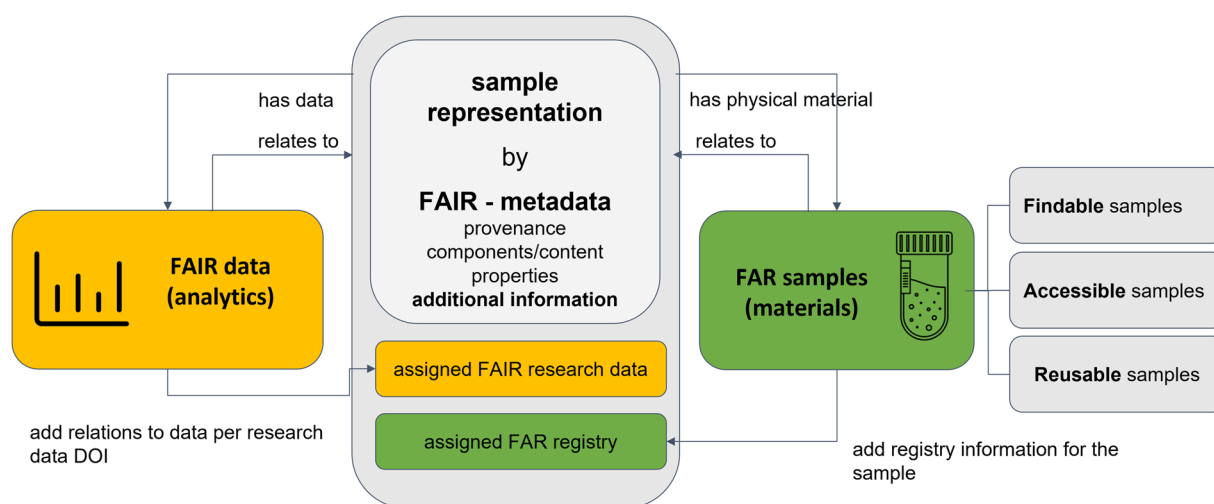


Fig. 2 Design of the infrastructure towards more sustainable scientific work in chemistry: The concepts of FAIR data, e.g. analytical measurements, are combined with the concept of findable, accessible and reusable samples through the virtual sample representation given by a sample's FAIR metadata. [ref Icons 2: Tube: *Digithrust from the noun project*, Data: *Popular from the noun project*].

by rich standardized metadata, which is complemented by (1) information on further relevant data assigned to the sample (Fig. 2, left), and (2) information on available samples' location and its unique registry or reference number (Fig. 2, right). The information on the samples' location must come from the register of samples made available *via* an archive for materials. A standardized process for the stockpiling of the samples, with a suitable validation of the materials and mechanisms to assign unique identifiers, is needed to gain such a deposition and registration of samples. Further, it means that the delivery and sample-sharing process (if applicable) is well described and standardized, including rules that might control access to the materials. Therefore, a sustainable materials approach should include information on the usage conditions of samples, such as legal agreements, and if available, safety information.

Implementation in the form of infrastructure. The combined approach for research data and materials/samples makes sense due to the obvious methodical connection of both *via* the virtual sample representation (Fig. 2) and the often observed issue that chemical samples are only efficiently reusable by others if they are described

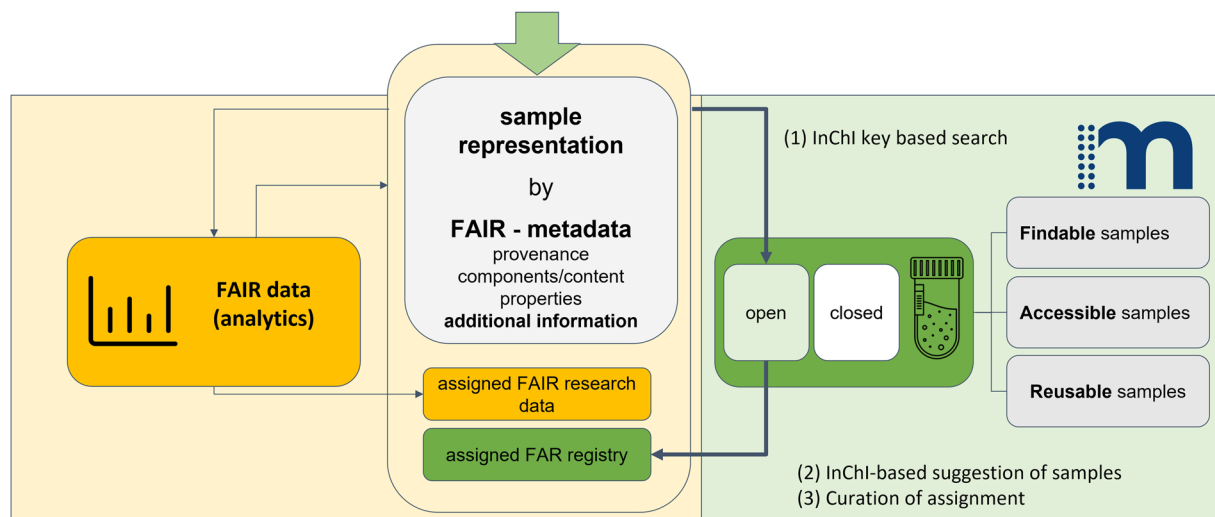


Fig. 3 Technical implementation of the concept described in Fig. 2. The Chemotion repository is used as an infrastructure component to make research data and samples findable and accessible through one search entry, which is built on the sample representation. [ref Icons 3: Tube: *Digithrust from the noun project*, Data: *Popular from the noun project*].

sufficiently, including the full set of available research data or analytical data. Therefore, methodical aspects and scientific reasons favor combining of or linking of a sample archival system and a research data repository. Such a combination was realized by using the Chemotion repository and the Molecule Archive, both located at Karlsruhe Institute of Technology (KIT). Chemotion repository is a repository for research data related to chemistry, particularly experimental chemistry. Scientists can either upload data directly or interoperably transfer it from an electronic lab notebook to the repository without losing information. The data are peer-reviewed in combination with automatic checks and can then be disclosed with persistent identifiers²¹. Especially for organic chemistry, the repository offers additional functions for direct viewing and analysis of the stored data by discipline-specific research software²² and thus enables an easy way of reusing the data. The repository follows an open access policy and is part of the strategy of NFDI4Chem in Germany.

The Molecule Archive is a facility of the KIT which enables the registration, validation, and collection of chemical substances. The substances are preserved for documentation and reuse purposes; therefore, strategies for sharing of the material and its provision have been developed. The use of the services of the Molecule Archive and the Chemotion repository is free of charge.

In context of the concept described herein, we make use of the infrastructure afforded by Chemotion repository (as a research data repository) to publish data and metadata of samples as the primary data entity, including generation of associated DOIs, to obtain a virtual sample representation that makes physical samples findable. This virtual sample representation is matched with the physical samples in Molecule Archive by exchanging information between it and the repository via a defined protocol. This protocol checks if samples' virtual representation in the Chemotion repository can be linked with registered physical samples in the Molecule Archive (Fig. 3, step (1)). The current request is based on the InChI key of a molecule, which is one of the most precise structural descriptors for chemical compounds. As a result, samples in the Molecule Archive, which have a sample representation in the Chemotion repository, are identified and information on their availability can be added to the information in the Chemotion repository. To this very general concept, a few saliences have been added: Since not all samples provided to the Molecule Archive are intended to be publicly visible, the visibility of samples depends on the assignment of samples to an open sample collection in the Molecule Archive. Only samples within the open sample collection of the Molecule Archive are queried through the Chemotion repository (Fig. 3, query to "open" collection given in green, right panel) and then visible through the repository's graphical user interface (GUI) (Fig. 3, step (2)). As the query is based on the InChI key of the molecule, the query may result in different suggested samples matching the InChI key. Therefore, linking of a sample in Chemotion repository to its physical counterpart is additionally curated by the Chemotion repository team (Fig. 3, step (3)).

Application and use of the infrastructure. The infrastructure as described was established at KIT and is used by different scientists dealing with the synthesis of chemical compounds and, therefore, producing chemical samples. The established process was applied to more than 1400 chemical compounds, including examples from organic chemistry to inorganic compounds^{23–26}, and metal-organic frameworks (MOFs)²⁷. Each compound is a well-characterized product of a chemical reaction that was conducted for a specific project or research aim. Most of the compounds were published as part of a chemistry study before or after their submission to the Molecule Archive and the deposition of the corresponding data to the Chemotion repository. Scientists from different institutions tested the robustness of the infrastructure with respect to digital transfer and deposition of data, as well as physical transfer and deposition of samples (Fig. 4). They contributed to adapting of the infrastructure

| Formula | Provided by Group | ID | Embargo | Analyses |
|--------------------------|--|-----------|----------------|----------|
| $C_{14}H_{12}N_2O_6$ | Timo Sehn Soft Matter Lab, KIT Karlsruhe | CRS-14928 | TGS_2020-10-21 | 7 |
| $C_{12}H_{11}ClN_2O_2S$ | Maria Manuel Marques Maria Manuel Marques Group | CRS-28273 | CWG_2022-11-23 | 2 |
| $C_{12}H_{14}N_6O_2$ | Andreas Link Andreas Link Group | CRS-33437 | Embargo | 4 |
| $C_{17}H_{16}N_2O_5$ | Rachel Janßen Lena Daumann Group | CRS-25869 | RAJ_2022-08-25 | 4 |
| $C_{15}H_{14}Cl_2O_2S_2$ | Patrick Hodapp Stefan Bräse Group | CRS-29767 | PH2_2023-01-27 | 3 |
| $C_{19}H_{20}N_2S$ | Fabian Fink Sonja Herres-Pawlis Group | CRS-17304 | Embargo | 7 |

Fig. 4 Overview of available material summarized in the table *physical samples* as given in the GUI of Chemotion repository in the section *Molecule Archive - physical samples*. The examples were collected and reorganized for this figure to reference different contributions (the entries were obtained in the repository Chemotion in a different order).

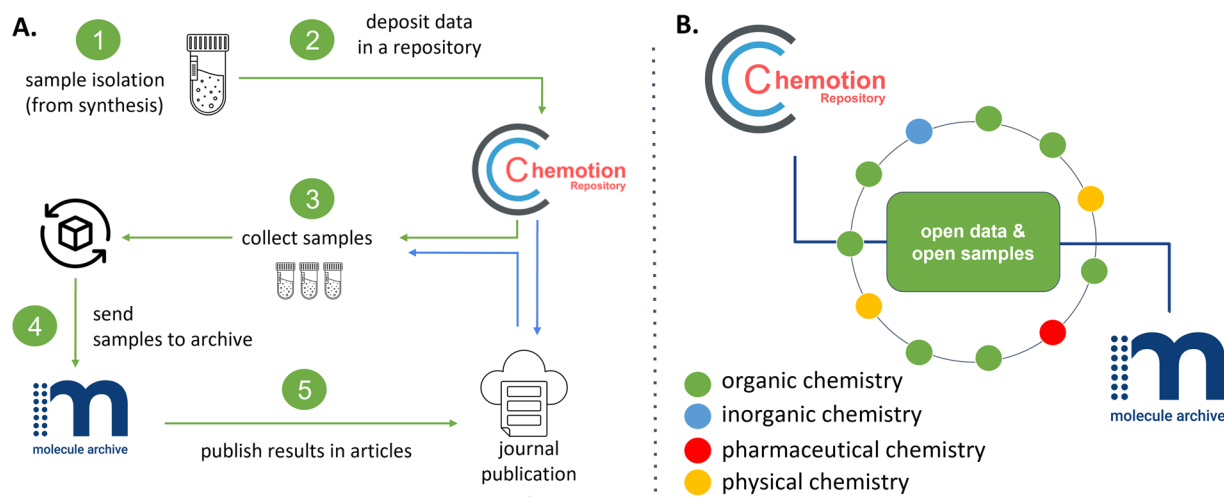


Fig. 5 (A) Proposed workflow to gain FAIR-FAR samples along with the publication of scientific results and data. The workflow consists of five steps, from preparing samples and their characterization to depositing data and physical material, which can be done according to the suggested order (green) or other alternatives (one example given in blue). Following the green workflow offers the option to publish scientific results with reference to the research data and the reference to research materials. (B) The scientific network that contributed to the workflow and infrastructure design as described in this article with examples from different subdisciplines in chemistry and requirements from different groups and sites (details given in the SI, File 2). [ref Icons 4: parcel: *Adrien Coquet*, tube: *Digithrust from the noun project*, publication: *Vectors Lab*].

components to their scientific needs. The proposed process to obtain FAIR-FAR samples includes five steps: (1) Isolation of samples from a chemical reaction or a natural product isolation and their analytical characterization, (2) entering of the samples' metadata and the deposition of the analytical data in a research data repository, (3) collection of samples and (4) shipping them to the location of the materials' archive (in our example: Molecule Archive at KIT) including registration of the samples in the archive by the archive's staff, and finally (5) publication of the results in scientific journals (Fig. 5A). The partner sites and research groups (circles) that helped establish the described FAIR-FAR samples model by providing the first use cases for open data and open materials in different subdisciplines in chemistry and requirements from different groups and sites (details given in the SI, File 2). The implemented process and the order of accompanying proposed steps is aligned with the demands of the research system in Germany,

including new requirements to openly provide research data alongside scientific publications^{28,29}. According to this requirement, the research data should be submitted to a repository before or at least in parallel with the publication of manuscripts in scientific journals in order to allow access to data during the review of publications. Therefore, step 2 needs to be finished before step 5 - even though data might be restricted to reviewers at this stage. The order of publication and materials' provision can vary flexibly depending on the amount of material available and the intended publication strategy. Archiving the material before scientific results are published (green workflow in Fig. 5A) could strengthen the publication, as the materials' archive can confirm the provision of the sample and provide information on the quality/purity of the samples. Further, it then becomes possible to refer to the compounds in the associated publication(s) - very similar to the referencing of data deposition in repositories. On the other hand, it may be preferable to deposit materials after publication of scientific results, i.e. once review is finished, especially when only a limited quantity of material is available (blue arrows in Fig. 5A). This has the advantage that materials are available at hand if reviewers need additional data from the authors.

The scientific network described in Fig. 5B helped fine-tune the proposed workflow of Fig. 5A and demonstrated that central availability of data and materials enables sharing of results with other scientists. We now discuss the impact of infrastructure and workflows described here on open and sustainable science in terms of findability, accessibility, interoperability and reusability of the materials metadata given in the sample representation as well as the findability, accessibility and reusability of the physical material itself.

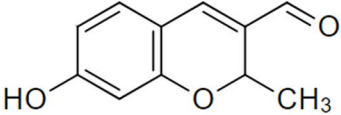
Findable sample metadata. The materials available within the Molecule Archive become manually and machine searchable through the samples' metadata generated in the Chemotion repository. The repository's GUI provides options for searching text and chemical structure, allowing users to find the available metadata based on the samples' description which consists of provenance information, properties and the components given as molecular descriptors. Along with the samples' virtual representation, the associated analytical research data, the physical location and IDs of the materials are findable. The sample representation is assigned a DOI, and its metadata includes this DOI. The metadata scheme (adopted from DataCite) also contains additional provenance information, physical descriptors characterizing the sample, and the information on the ID of the physical sample as registered in the Molecule Archive. Also included are the DOI to the chemical reaction data that generated the sample (if available) and the DOIs to associated analytical details. Metadata in the scheme are assigned to terminologies of established ontologies such as CHMO^{30,31}, CHEMINF³², OBI³³, and ChEBI³⁴ wherever possible. In addition to being downloadable through the user interface of the repository, the metadata is also accessible via an Application Programming Interface (API) using a metadata harvesting protocol (OAI-PMH)³⁵. Two examples of typical metadata schemas have been included with the Supplemental Information (SI, Fig S1 and Fig S2).

The OAI-PMH protocol is used because of its broad usage and well-defined procedure. However, it also has some disadvantages such as the need to run and manage a dedicated OAI-PMH endpoint resulting in increased operational complexity. In addition, OAI-PMH is less scalable for very large datasets or frequent metadata updates as it relies on repeated harvesting. Therefore, in the future, accessibility to the metadata will be improved by including JSON-LD metadata schemas. The benefit of using JSON-LD with community-agreed schemas is that they are better at providing semantically rich domain-specific metadata and the linking of property values to defined terms from ontologies is more explicit. JSON-LD is also simpler to implement than OAI-PMH because it can be embedded directly in web pages and can be accessed via HTTP, eliminating the need for a specialized server or protocol making it faster and more scalable in the long run³⁶. A first draft of such a JSON-LD implementation for samples was embedded in the samples' representation in the Chemotion repository (Fig. 6, part 3; further information can be gained from SI section 5) and will be the subject of further discussion and improvement through the community of NFDI4Chem³⁷.

Accessible sample metadata. The Chemotion repository supports the OAI-PMH protocol, which is a widely used protocol for exposing metadata records in a standardized way that can then be harvested and aggregated by other systems. OAI-PMH provides access to metadata records at various levels, including individual and sets of records. For example, by using the "ListRecords" verb and applying filters such as the metadata prefix "oai_dc" and a date range users can obtain a list of complete records in Dublin Core format from the repository. Similarly, by using the "GetRecord" verb and applying filters such as the metadata prefix "oai_DataCite" and specifying an identifier like a sample's DOI, users can retrieve a specific sample record, and therefore, access specific metadata from the Chemotion Repository in DataCite format. The protocol supports multiple metadata formats, such as Dublin Core and DataCite which allow easy interoperability of the Chemotion repository with other systems such as the NFDI4Chem search service³⁸. Once the user finds the required sample information, all available metadata are directly accessible by their DOI identifier without further authentication and authorization processes. The metadata remains accessible, even when the materials are no longer available, and allow for a constant link to the analytical data.

Interoperable sample metadata. Interoperable metadata are needed in particular to compare the materials with other available materials and to query additional databases. The sample metadata in the Chemotion repository include standardized molecule descriptors following domain-specific standards such as InChI, InChI keys and canonical SMILES strings wherever applicable. The terminologies used in the description of the samples are chosen from domain relevant ontologies and are assigned to the identifiers of the mentioned ontologies.

Reusable sample metadata. The metadata for the sample in the Chemotion repository are described as comprehensively as possible, including description of the components characterizing the samples, their properties, and additional references to analytical data and synthetic origin of the sample (chemical reaction). In particular, the reference to analytical data and reactions allows reuse of the data since this information is needed



IUPAC Name: 7-hydroxy-2-methyl-2H-chromene-3-carbaldehyde (C₁₁H₁₀O₃)


Canonical SMILES: O=CC1=Cc2ccc(cc2OC1C)O



InChI: InChI=1S/C11H10O3/c1-7-9(6-12)4-8-2-3-10(13)5-11(8)14-7/h2-7,13H,1H3


InChIKey: USJQBBWAVFLWBI-UHFFFAOYSA-N

Exact Mass: 190.062994 g·mol⁻¹


A physical sample of this molecule was registered to the Molecule Archive of the Compound Platform

Crosslinks: 

 **Sample Published on** 2023-02-07 



Contributor:  Simone Gräßle


1. Institute of Organic Chemistry, Karlsruhe Institute of Technology, Germany


Author:  Simone Gräßle^{1,2}

1. Institute of Organic Chemistry, Karlsruhe Institute of Technology, Germany

Sample type: Consists of molecule with defined structure



Sample DOI: 10.14272/USJQBBWAVFLWBI-UHFFFAOYSA-N.1   **JSON-LD**


Sample ID: CRS-22414 



Relations of this sample: Is Product of a reaction , has analytical data, has a record as physically available material



Reference in the Literature:



Physical Properties:
Melting point: 153.8 - 166.5
Boiling point:


Material  

Sample Registration Number in Molecule Archive: ComP-3384
Request a sample: 

Analyses  **1H NMR, 13C NMR, DEPT, DEPT, HSQC, HMBC, COSY, MS, IR** 

1H nuclear magnetic resonance spectroscopy (1H NMR)  

Analysis DOI: 10.14272/USJQBBWAVFLWBI-UHFFFAOYSA-N/CHMO0000593  

Reaction ID: CRD-22405 

¹H NMR (400 MHz, Acetone-d₆ [2.05 ppm], ppm) δ = 9.48 (s, 1H), 9.23 (s, 1H), 7.38 (s, 1H), 7.21 (d, J = 8.3 Hz, 1H), 6.50 (dd, J = 2.3 Hz, J = 8.3 Hz, 1H), 6.36 (dd, J = 0.6 Hz, J = 2.3 Hz, 1H), 5.28 (q, J = 6.5 Hz, 1H), 1.27 (d, J = 6.6 Hz, 3H). Impurities: spectrum contains ethyl acetate (4.07 ppm, 1.97 ppm and 1.22 ppm) and water (3.06 ppm).



Datasets
1H NMR  

Fig. 6 Explanation of the main parts that are used to describe a sample in the publication view of Chemotion repository: (1) Formal description of the sample's virtual representation by information on the molecule which is part of the sample; (2) general publication metadata; (3) sample's identifiers in Chemotion repository, (4) Selection of physical properties, (5) Access to the sample by contacting the team of the Molecule Archive and sample's ID in the Molecule Archive; (6) Links to the analytical data that were gained with the sample. The Figure was created from a screenshot from the Chemotion repository and cut/changed to obtain an image that uses less space for a better readability of the article (see SI section 6 for the original screenshot as obtained).

for chemists to reproduce the experiments. A plurality of additional data that further describes the sample can be gained from the references to analytical data and the synthetic origin of the sample.

To gain FAIR-FAR samples, the FAIR metadata for samples is supplemented with measures to obtain findable, accessible and reusable materials/samples:

Findable materials. Labeling materials and cataloging of these labels in an accessible database are two essential steps to make materials findable. A clear and unique label or code (such as a barcode or a QR code) makes materials findable in a certain storage location by humans. As a second step, the material needs to be cataloged in a local database that keeps a certain minimum set of metadata about the material, including at least its label and the location where the material is stored. In our case, the registration is achieved by entering the samples' metadata into the database of the Molecule Archive, making the samples findable by machines. As soon as the samples are physically deposited and registered in the Molecule Archive, the external findability of the available samples is further managed via the website of the Chemotion repository which allows the linkage of the unique label in the Molecule Archive with the metadata of the samples' FAIR metadata which are visible globally.

Accessible materials. Scientists interested in reusing the samples can place their request for sample directly through the repository's interface (Fig. 6, details added to the SI, chapter 2). A contact form has been set up for each available sample to obtain the chemical compound in the form of a part of the sample. The query automatically transfers the identifier of the sample. A key difference in the notion of "accessible" for data and materials is the following: While access to data can be granted to all interested persons without disadvantage in each case, prioritization must be made with respect to access of materials. Since the amount of an available chemical compound archived per sample is usually very limited, there must be a consideration of the purposes for which the material should be released. For the Molecule Archive, the decision on whether to release the materials or not is made either based on a material transfer agreement (MTA³⁹) of compound providers with the Molecule Archive (see chapter "reusability") or is decided by the compound providers (and managed by the operators of the repository). The decision is sent to the interested reusers of the chemical compounds and if the material can be sent, the details and conditions for such reuse are clarified.

Reusable materials. The reuse of samples available in the Molecule Archive is supported by quality assurance measures, framework agreements, consulting, and support in handling. For quality assurance, the registered samples are checked for purity and identity. The registration of the materials is not completed until the substances are physically available at the Molecule Archive and their identity and purity are checked. In contrast to sharing of data which can be managed by suitable licenses, the sharing of material needs material transfer agreements that clarify the role and rights of materials' providers, materials' reusers, and the Molecule Archive. The Molecule Archive supports the reuse of samples under a legal framework by providing a standardized material transfer agreement that was agreed on by KIT with different exemplarily chosen universities and published as a reference for further collaborators³⁹. Further, the Molecule Archive organizes the communication between the participating scientists as needed. After clarification of all technical and legal issues, the provision of the material for subsequent use is initiated by preparing for sample transfer. The materials are then shipped according to common standards for chemical compounds.

Discussion

Using examples from the subdisciplines of organic, pharmaceutical, inorganic, and experimental physical chemistry, a model was designed to provide data and materials of research results. The described approach is intended to be a possible first step towards more sustainable scientific work in chemistry, but currently, some challenges remain unsolved or are insolvable and impose a permanent limitation on the endeavor:

- (1) In principle, samples that do not tend to decompose under ambient conditions and are not volatile can be registered and recorded by the Molecular Archive. Currently, unstable metal-organic compounds in particular cannot be introduced as openly accessible and reusable samples. Storage under an inert gas atmosphere could expand the model's applicability to include a wider range of substances; however, this is not done in the present setup.
- (2) So far, the method has only been used to provide data and materials for samples with a defined chemical structure. Therefore, mixtures or natural extracts have not been submitted as FAIR-FAR samples yet. This is, in principle, possible but the Chemotion repository is currently geared towards pure substances and will be adapted in the medium term.
- (3) While the registration and also the analysis of more complicated compound classes such as Metal-Organic-Frameworks (MOFs) is possible and the described infrastructure can be used to store and preserve MOF materials and data, the infrastructure is not well-suited for related samples such as SURMOFs (Surface-anchored MOFs). The infrastructure and concepts can still be used also for samples beyond their current scope - such as SURMOFs - but the physical archival and the evaluation of the gained material would need further solutions.
- (4) In some areas of chemistry, samples are generated that cannot be unambiguously described by any unique chemical structure. In such cases, the FAIR-FAR sample process described here can be used without restrictions, making the samples discoverable by DOIs. However, samples in these cases cannot be searched as efficiently as possible for samples with unique chemical structures because they would lack structural descriptors in the metadata.
Other obvious hurdles for accessible and reusable samples in such cases arise from their limited availability, the resulting need for coordination to release the samples, and additional legal, technical, and security aspects.
- (5) While FAIR data can be reused almost indefinitely by granting a license and choosing the appropriate data infrastructure with regard to the number of reusers, various interests may have to be weighed against each other for the reuse of material - and the resources may be exhausted even if there is a need for further reuse. Currently, the Molecular Archive cannot provide a universal solution to this problem

because the reuse scenarios are different, and decisions must be made on a case-by-case basis to achieve the highest benefit for the available substances. A well-organized distribution system must ensure that a certain amount of reference substance is kept, even if a high request for the compounds (samples) exists. This allows a residual amount to be available as analytical evidence independent of subsequent users of the substances. Still, the sample is linked to the synthesis protocol in Chemotion, as last aid in the case of depletion.

- (6) The provision of substances for reuse purposes requires time for preparation, at least at the time of the first provision of materials, due to the usually required MTA between the partners involved. The provision of samples therefore also depends on the processing time by the respective organizational units of the partners involved, if a legal basis for material reuse is to be created as it is done when licensing research data.

Despite the obvious challenges of providing materials, the combination of FAIR data and FAIR-FAR materials reveals enormous potential for more transparent and sustainable research. In particular, experimental disciplines in natural sciences may benefit from a concept introducing models for a systematic provision of samples. If samples are submitted before the publication of results, the benefits could be increased as the sample deposit can be directly linked to the publication. This strengthens the trust in the research work and allows direct linking of publication(s) to all its results: research data as well as the physical material.

Knowing that the implementation of the FAIR data principles is still far from completion, the additional push for FAIR-FAR samples might seem to be challenging. Nevertheless, the establishment of a standardized process for FAIR-FAR samples can be done very easily once the concept of FAIR data is adopted: the infrastructure supporting the access and reuse of chemical samples already exists, and the effort for a single scientist as the material's producer is low if the initial agreements between the partner sites and the Molecule Archive are already in place. While the deposition of FAIR data currently lacks incentives for the providing scientists, the provision of materials offers special scientific advantages such as publications with materials' reusers - this makes materials storage and provision an attractive aim that could foster the broad application of the FAIR-FAR concept. The provision of samples to the Molecule Archive has been the origin of many publications that were done in collaboration with compound providers and reusers - proving that the provisioning can result directly in more visibility and impact for the scientists sharing their research results⁴⁰⁻⁴⁶.

Methods

Software. The infrastructure described in this article is built with the use of open source software that was developed at KIT and has been described in previous articles. Both the Chemotion repository and the software behind the Molecule Archive were developed based on components of the source code of Chemotion ELN^{47,48}. Further extensions of Chemotion ELN, including submission and reviewing workflow, provide the necessary functionality to operate the Chemotion research data repository^{21,49}. The source code for the Chemotion repository can be obtained from GitHub⁵⁰. For the operation of the Molecule Archive, Chemotion ELN was also used and adapted with a plugin⁵¹ to keep additional information on the sample information as provided by the owner and was extended with Foreign Data Wrappers (FDW)⁵² for smooth integration of data from the Chemotion repository and the Molecule Archive. An archived version of the source code of the Chemotion repository, as used for the work described in this article, can be obtained from Zenodo⁵³.

Submission of data to the Chemotion repository. The submission of data to the Chemotion repository enables the generation of the FAIR metadata of the samples' virtual representation and the storage of FAIR research data. The data uploaded to the repository includes a request to the user to add information on the sample that was used for the measurement of the data. Usually this is the chemical structure of the compound assigned to the sample and additional information on the purity of the sample and other characteristics. This information is used by the software to automatically generate further sample metadata that can be used to identify the sample and to search for the sample. The combination of information entered by the user and system-generated information directly forms the virtual metadata representation of the sample and defines the digital object. With the information available about the sample, the submission of research data can be started. The data has to be prepared according to discipline-specific standards, described in detail in the online documentation of the Chemotion repository⁵⁴. As soon as the virtual sample representation is visible (along with the data) after the publication of the submission, the correlation of research data and materials can be started by the team of the repository.

Setting up a legal framework for sharing materials. Sharing of materials in a scientific environment can be done in two ways: either via a donation of the material from one scientific group to another one, or *via* the transfer of material under certain negotiated conditions. The FAIR-FAR samples concept supports both ways of sharing materials. Establishing the transfer of materials under an MTA costs more time and effort at the beginning of the sharing process - but enables the provision and reuse of compounds under clearly defined rules and is therefore the preferred way of sharing materials. Together with five exemplarily chosen partner sites, KIT created a standard MTA several years ago, which is now used as a routine process to introduce new partner sites to provide materials to the Molecule Archive³⁹. The MTA outlines the rights and obligations among the material providers, the entity managing ComPlat and the reuser. It regulates the handling of compounds as well as data, and the publication of results in good scientific practice. Scientific credit is in particular important if the reusers of the materials gain scientific results with the provided compounds, and these results should be published together. The MTA emphasizes that reuse shall aim at publication and noncommercial purpose. A notification and non-disclosure retention period allows the preparation of publications and, exceptionally, the patenting of inventions that may involve the material or a related process. Where patentable inventions of a provider are inseparably linked to the scientific results of the reuser, the latter shall contractually perceive a contractual license

option against conditions customary on the market. This compensates for the tax-payer funded prior-efforts, in accordance with state aid law. Altogether, more than three dozen research groups at universities and other noncommercial research entities are already partners of the Molecule Archive network, and scientists working at these sites can work under the existing MTA. Other scientists who work at institutions that do not have an agreement yet can request to start the MTA generation process with their institution.

Submission of samples to the molecule archive. The submission of samples to the Molecule Archive works *via* a simple workflow: The Molecule Archive provides suitable standard vessels that are sent to the compound provider. The vessels are calibrated and carry a unique number for their later identification. A table sheet is provided to the users, which is required to register the sample in the database of the Molecule Archive. The sample providers need to give brief information on the chemical structure assigned to the materials in the form of the corresponding SMILES code, the code of the used vessel, the internal laboratory ID and properties such as the approximate purity of the material and the filled mass (an example is added to the SI, section 4). The filled vessels are then sent back to the Molecule Archive using the provided packaging material and the digital upload form is transmitted *via* email.

Registration of samples in the molecule archive. The registration of the samples in the database of the Molecule Archive is done by the team of the Molecule Archive after the arrival of the material and works *via* upload of the table sheet to the database of the Molecule Archive. The Molecule Archive team checks the identity and purity of the compounds *via* LCMS (liquid chromatography coupled with mass spectrometry) and other techniques if required. If the data correspond to the provided structure of the sample-associated compound, the sample registration is finished, and the provider receives documentation about the submission and the results of the quality control. If the sample provider decides to make the sample openly accessible, the database entry for the sample is assigned to the collection of open samples within the database. The collection can then be accessed through the Chemotion repository and the sample is visible in the user interface of the repository along with the virtual sample representation (as depicted in Fig. 6 and the SI).

Data availability

The Supplemental Information (SI, file 1) includes two examples as representative metadata schema of a virtual sample representation available in the Chemotion repository (chapter 1), and a detailed description of how the samples visible in the Chemotion repository website can be assessed for further reuse through the provided request form (chapter 2). Further, SI - part 1 covers further additional information on the processes of the Molecule Archive (chapter 3) including a template of a data upload form that is used to register samples in the Molecule Archive (chapter 4), an example for the currently implemented JSON-LD description (chapter 5), and the complete images of Fig. 6 in this article (chapter 6). The SI (file 2) contains an exemplarily collected and non-comprehensive list of partners and their contributions to the FAIR-AR samples concept. To give examples for the work described here, 170 examples out of ca. 1400 FAIR-FAR open samples (accessed on June, 5 2023, <https://www.chemotion-repository.net/welcome>) are cited in file 2 to allow direct access to some examples.

Code availability

The code of the software that was adapted for this project is available as open source: Chemotion repository: <https://doi.org/10.5281/zenodo.8028033>; Molecule Archive functionality in Chemotion ELN code: <https://gitlab.kit.edu/kit/complat/xvial>.

Received: 20 September 2023; Accepted: 3 January 2025;

Published online: 22 January 2025

References

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
2. IUPAC | International Union of Pure and Applied Chemistry. <https://iupac.org/who-we-are/>. Date accessed: June 11, 2023.
3. RDA | research data alliance. <https://www.rd-alliance.org/>. Date accessed: June 11, 2023.
4. CODATA | Materials Data, Infrastructure & Interoperability Interest group in research data alliance (RDA). <https://www.rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html>. Date accessed: June 11, 2023.
5. EOSC | European Open Science Cloud, <https://eosc-portal.eu/>. <https://eosc-portal.eu/>. Date accessed: June 11, 2023.
6. Hartl, N., Wössner, E. & Sure-Vetter, Y. Nationale Forschungsdateninfrastruktur (NFDI). *Informatik Spektrum* **44**, 370–373 (2021).
7. Steinbeck, C. *et al.* NFDI4Chem—A research data network for international chemistry. *Chem. Int.* **45**, 8–13 (2023).
8. Hardisty, A. *et al.* The Specimen Data Refinery: A canonical workflow framework and FAIR digital object approach to speeding up digital mobilisation of natural history collections. *Data Intell.* **4**, 320–341 (2022).
9. National Research Council, Division on Earth and Life Studies, Board on Earth Sciences and Resources, Committee on Earth Resources & Committee on the Preservation of Geoscience Data and Collections. *Geoscience Data and Collections: National Resources in Peril*. <https://doi.org/10.17226/10348> (National Academies Press, 2002).
10. IODP | International Ocean Discovery Program - Bremen Core Repository. <https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html>. Date accessed: June 11, 2023.
11. Klump, J. *et al.* Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Sci. J.* **20** (2021).
12. Damerow, J. E. *et al.* Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. *Data Sci. J.* **20**, 11 (2021).
13. Hardisty, A. *et al.* A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). *Res. Ideas Outcomes* **7** (2021).
14. Deck, J. *et al.* The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol.* **15**, e2002925 (2017).
15. Davies, N. *et al.* Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *Gigascience* **10** (2021).

16. Thessen, A. E. *et al.* Proper attribution for curation and maintenance of research collections: Metadata recommendations of the RDA/TDWG working group. *Data Sci. J.* **18**, 54 (2019).
17. Simpson, M. & Poulsen, S.-A. An overview of Australia's compound management facility: the Queensland Compound Library. *ACS Chem. Biol.* **9**, 28–33 (2014).
18. Brennecke, P. *et al.* EU-OPENSREEN: A Novel Collaborative Approach to Facilitate Chemical Biology. *SLAS Discov* **24**, 398–413 (2019).
19. Mahuteau-Betzer, F. Chimiothèque Nationale - Avancées et perspectives. *médecine/sciences* **31**, 417–422 (2015).
20. Center for Molecular Discovery at Boston University. <https://www.bu.edu/articles/2016/center-for-molecular-discovery/>. Date accessed: June 11, 2023.
21. Tremouilhac, P., Huang, P. C. & Lin, C. L. Chemotion repository, a curated repository for reaction information and analytical data. *Chemistry - Methods* **1**, 8–11 (2021).
22. Huang, Y.-C., Tremouilhac, P., Nguyen, A., Jung, N. & Bräse, S. ChemSpectra: a web-based spectra editor for analytical data. *J. Cheminform.* **13**, 8 (2021).
23. Fink, F. Dichlorocopper;2,2-di(pyrazol-1-yl)ethanamine (C8H11Cl2CuN5), Chemotion repository, <https://doi.org/10.14272/AMQWWHVPPZUOKP-UHFFFAOYSA-L.1.chemotion.net> (2020).
24. Kalyakina, A. C49H36EuF3N4O6, Chemotion repository, <https://doi.org/10.14272/XSDMQMVIDCRHRR-UHFFFAOYSA-K.1.chemotion.net> (2022).
25. Holzhauer, L. [2-(4-Butyl-1H-1,2,3-triazol-1-yl)-3-phenylquinoxaline]bromotricarbonylrhenium(I), Chemotion repository, <https://doi.org/10.14272/BKFBOTQHKAJKOY-UHFFFAOYSA-M.1.chemotion.net> (2022).
26. Schissler, C. C76H47N9NiZn, Chemotion repository, <https://doi.org/10.14272/JVGPXCSDWSVRU-HWNMUZRGSA-N.1.chemotion.net> (2022).
27. Pilz, L. Dicopper;benzene-1,3,5-tricarboxylate, Chemotion repository, <https://doi.org/10.14272/JMHFTDFPRQWUAN-UHFFFAOYSA-H.22.chemotion.net> (2022).
28. *Guidelines for Safeguarding Good Research Practice: Code of Conduct.* (Deutsche Forschungsgemeinschaft, 2019).
29. D Forschungsgemeinschaft. Guidelines for Safeguarding Good Research Practice. Code of Conduct, <https://doi.org/10.5281/zenodo.6472827> (2022).
30. McEwen, L. R. & Buntrock, R. E. *The Future of the History of Chemical Information (ACS Symposium, Band 1164)*. (Am Chem Soc, 2015).
31. Strömert, P., Hunold, J., Castro, A., Neumann, S. & Koepler, O. Ontologies4Chem: the landscape of ontologies in chemistry. *J. Macromol. Sci. Part A Pure Appl. Chem.* **94**, 605–622 (2022).
32. Hastings, J. *et al.* The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One* **6**, e25513 (2011).
33. Bandrowski, A. *et al.* The Ontology for Biomedical Investigations. *PLoS One* **11**, e0154556 (2016).
34. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–9 (2016).
35. Lagoze, C. & Van de Sompel, H. The Open Archives Initiative: Building a low-barrier interoperability framework, <https://www.openarchives.org/documents/jcdl2001-oi.pdf>. Date accessed: June 11, 2023.
36. Klump, J. *et al.* Scaling identifiers and their metadata to gigascale: An architecture to tackle the challenges of volume and variety. *Data Sci. J.* **22** (2023).
37. NFDI4Chem | National Research Data Infrastructure for Chemistry, summary of the network of NFDI4Chem: <https://www.nfdi4chem.de/index.php/network/>. Date accessed: June 11, 2023.
38. NFDI4Chem Search Service, <https://search.nfdi4chem.de/>. Date accessed: Jun 20, 2023.
39. Karlsruhe Institute of Technology, legal affairs unit. Agreement on the transfer of materials via the Compound Platform (ComPlat). Preprint at <https://doi.org/10.35097/1022> (2023).
40. Macara, J. *et al.* Practical synthesis and biological screening of sulfonyl hydrazides. *Org. Biomol. Chem.* **21**, 2118–2126 (2023).
41. Apweiler, M. *et al.* Functional Selectivity of Coumarin Derivates Acting via GPR55 in Neuroinflammation. *Int. J. Mol. Sci.* **23** (2022).
42. Frei, A. *et al.* Metal Complexes as Antifungals? From a Crowd-Sourced Compound Library to the First *In Vivo* Experiments. *JACS Au* **2**, 2277–2294 (2022).
43. Lei, W. *et al.* Droplet microarray as a powerful platform for seeking new antibiotics against multidrug-resistant bacteria. *Adv Biol (Weinh)* e2200166 <https://doi.org/10.1002/adbi.202200166> (2022).
44. Hofmann, D. *et al.* A small molecule screen identifies novel inhibitors of mechanosensory nematocyst discharge in Hydra. *Sci. Rep.* **11**, 20627 (2021).
45. König, G. *et al.* Rational prioritization strategy allows the design of macrolide derivatives that overcome antibiotic resistance. *Proc. Natl. Acad. Sci. USA.* **118** (2021).
46. Raudszus, R. *et al.* Pharmacological inhibition of TRPV2 attenuates phagocytosis and lipopolysaccharide-induced migration of primary macrophages. *Br. J. Pharmacol.* <https://doi.org/10.1111/bph.16154> (2023).
47. Tremouilhac, P. *et al.* Chemotion ELN: an Open Source electronic lab notebook for chemists in academia. *J. Cheminform.* **9**, 54 (2017).
48. Kotov, S., Tremouilhac, P., Jung, N. & Bräse, S. Chemotion-ELN part 2: adaption of an embedded Ketcher editor to advanced research applications. *J. Cheminform.* **10**, 38 (2018).
49. Tremouilhac, P. *et al.* The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry*. *Angew. Chem. Int. Ed Engl.* **59**, 22771–22778 (2020).
50. *GitHub Reference for Chemotion REPO: A Repository Based on Chemotion ELN*, https://github.com/ComPlat/chemotion_REPO. (Github).
51. GitLab reference for a Module to enable X-Vial listing in chemotion ELN, <https://git.scc.kit.edu/ComPlat/Xvial>. *GitLab*.
52. Chapter 56. Writing a foreign data wrapper. *PostgreSQL Documentation* <https://www.postgresql.org/docs/12/fdwhandler.html> (2023).
53. Lin, C.-L., Huang, P. C., Tremouilhac, P., Jung, N. & Le, L. *ComPlat/chemotion_REPO: Chemotion Repository 1.1.0* <https://doi.org/10.5281/zenodo.8028033> (2023).
54. Documentation for Chemotion Repository, <https://www.chemotion.net/docs/repo>. Date accessed: June 11, 2023.

Acknowledgements

The results of this project could be gained due to the support of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for the projects Chemotion ELN (project number: 266379491), DFG core facilities Compound Platform (project number: 284178167) and the NFDI4Chem (project number: 441958208). The project group was further supported by the project ELN ElCh of the BMBF cluster ETOS (03ZU1205OA). We are very thankful to the members of the Stefan Bräse group who contributed to the establishment of the repository and the Molecule Archive. We thank the following scientists who provided samples for the Molecule Archive and the Chemotion repository to improve the workflows described in this publication: scientists from

Karlsruhe Institute of Technology, KIT, Germany (department at KIT): Changming Hu (INT), Timo Sehn (IOC), Lena Pilz (IFG), Ilona Wagner (IFG); scientists from other universities: Violeta Vetsova, Rachel Janssen (both LMU, Munich, Germany), Sylvain Grosjean (Université de Franche-Comté – UFC, Besancon, France), Miro Hałaczkiwicz (RPTU Kaiserslautern-Landau, Germany), Robert Forster, Rainer Wiechert (both JGU Mainz, Germany), Felix Potlitz (University of Greifswald, Germany) and Fabian Thomas, Regina Schmidt, Christian Conrads (RWTH Aachen University). We thank Noura Rayya (FSU Jena, Germany), Tillmann Fischer (IPB Halle, Germany) and Philip Strömert (TIB Hannover, Germany) for helpful advice referring to metadata and schemas. Likewise, we are thankful for the support of the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK Baden-Württemberg) through the project MoMaF, which facilitated the hosting of the Chemotion repository as part of the developments within the Science Data Center of the MWK. We further acknowledge the support of the Helmholtz research field information and the Karlsruhe Nano Micro Facility, which support the maintenance of the software Chemotion ELN.

Author contributions

C.L.L., P.C.H. and P.T. designed the technical processes and adapted the software Chemotion repository and Molecule Archive to meet the needs of this work. S.G., S.V., N.J., P.H. and C.G. elaborated the necessary steps for operating the Molecule Archive in its current form. S.H.P., G.M., T.O., A.L., M.M.B.M., L.J.D., M.T., F.B., H.M., E.T. are PIs that acted as early adopters of the herein presented FAIR AR samples concept and contributed with materials and data to establish the system as described herein, F.B., S.N., A.H. and F.F. contributed with ideas and suggestions for the improvement of the overall process. T.D. elaborated (with colleagues) the legal basis of the work of the Molecule Archive. N.J. and S.B. established the herein described infrastructures Chemotion repository and Molecule Archive and developed the basic concepts to build the FAIRAR infrastructure. All authors edited the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-04404-2>.

Correspondence and requests for materials should be addressed to N.J. or S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025