

Supporting Sustainability of Chemistry by Linking Research Data with Physically Preserved Research Materials

*Chia-Lin Lin^a, Pei-Chi Huang^a, Simone Gräßle^a, Christoph Grathwol^a, Pierre Tremouilhac^a,
Sylvia Vanderheiden^a, Patrick Hodapp^b, Sonja Herres-Pawlis^c, Alexander Hoffmann^c, Fabian
Fink^c, Georg Manolikakes^d, Till Opatz^e, Andreas Link^f, M. Manuel B. Marques^g, Lena J.
Daumann^h, Manuel Tsotsalasⁱ, Frank Biedermann^j, Hatice Mutlu^k, Eric Täuscher^l, Felix Bach^m,
Tim Dreesⁿ, Steffen Neumann^o, Nicole Jung^{*a,p}, Stefan Bräse^{*a,q}*

Email: nicole.jung@kit.edu; stefan.braese@kit.edu

^aInstitute of Biological and Chemical Systems - Functional Molecular Systems (IBCS-FMS),
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-
Leopoldshafen, Germany; ^bInstitute for Biological Interfaces 3 - Soft Matter Laboratory (IBG 3 -
SML), Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-
Leopoldshafen, Germany; ^cRWTH Aachen University, Institute of Inorganic Chemistry,
Landoltweg 1a, 52074 Aachen, Germany; ^dRPTU Kaiserslautern-Landau, Department Chemie,
Erwin-Schrödinger-Str. Geb. 54, 67663 Kaiserslautern, Germany; ^eJGU Mainz, Department
Chemie, Duesbergweg 10-14, 55128 Mainz, Germany; ^fUniversität Greifswald, Institut für
Pharmazie, Friedrich-Ludwig-Jahn-Str. 17, 17489 Greifswald, Germany; ^gLAQV-REQUIMTE,

Department of Chemistry, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal; ^hDepartment of Chemistry, Ludwig-Maximilians-Universität München, Butenandtstr. 5-13, 81377 München, Germany; ⁱInstitute of Functional Interfaces (IFG), Karlsruhe Institute of Technology, Hermann von Helmholtz Platz 1, 76131 Karlsruhe, Germany; ^jInstitute of Nanotechnology (INT), Karlsruhe Institute of Technology, Hermann von Helmholtz Platz 1, 76131 Karlsruhe, Germany; ^kInstitut de Science des Matériaux de MulhouseUMR 7361 CNRS/Université de Haute Alsace15 rue Jean Starcky, Mulhouse Cedex 68057, France; ^lTechnische Universität Ilmenau, Institut für Chemie und Biotechnik, Weimarer Straße 25, 98693 Ilmenau; ^mFIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany; ⁿLegal Affairs, Karlsruhe Institute of Technology, Kaiser Strasse 12, 76131 Karlsruhe, Germany; ^oLeibniz Institute of Plant Biochemistry, Computational Plant Biochemistry group, Halle, Germany; ^pKarlsruhe Nano Micro Facility (KNMFi), Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany; ^qInstitute of Organic Chemistry (IOC), Karlsruhe Institute of Technology, Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany.

ABSTRACT

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 a FAIR (Findable, Accessible, Interoperable, Reusable) manner in research data repositories. Nevertheless, in order to make chemistry as a discipline sustainable, obtaining FAIR data is insufficient and a comprehensive concept including the preservation of materials is needed. We describe in this article how we combined two infrastructures, a repository for research data (Chemotion repository) and an archive for chemical compounds (Molecule Archive), in order to offer a comprehensive infrastructure to find and access data and materials that were generated in chemistry projects. Samples play a key role in this concept: we describe how FAIR metadata of a virtual sample representation can be used to refer to the physically available sample stored in a materials' archive and to link FAIR research data gained with the sample. We further describe the measures to make the physically available samples not only FAIR through the sample's metadata but also accessible and reusable in the form of their material for others.

Keywords: FAIR data, FAIR materials, Repositories, Sustainability, Open Science, Chemistry

INTRODUCTION

Sustainable work and the sustainable provision of research results for others is an essential criterion for efficient work in science. Only if results are accessible to the entire scientific community and thereby reusable, scientific progress can be accelerated in a targeted manner. Since the publication of the FAIR data principles¹, more and more scientists and stakeholders such as

funding agencies and journals/publishers support the generation and provision of FAIR data, regardless of the discipline. Accordingly, data should be Findable, Accessible, Interoperable and Reusable (FAIR), which is particularly relevant for providing research data that form the basis of publications. In chemistry and materials science, a variety of initiatives promote the provision of FAIR data or provide assistance for the implementation of FAIR data measures. Established stakeholders that supported the concepts of FAIR data even before their explicit publication are e.g. IUPAC², RDA³ or CODATA⁴. These have been joined by other important groups such as EOSC⁵, and in Germany by the National Research Data Infrastructure (NFDI)⁶, in particular the consortium for Chemistry (NFDI4Chem)⁷. In the future, NFDI4Chem will strengthen the chemical community by providing infrastructure for the generation and provision of FAIR data. Adhering to the FAIR data principles and aligning research processes to obtain FAIR data can be the basis for substantial improvement in data availability and quality. FAIR data can strengthen trust in research results through transparency and promote their systemic and barrier-free subsequent use. In synthetic chemistry, the FAIR principles must not be reduced to data and descriptive metadata alone. Chemists can provide more than data for documentation and subsequent use. Very often, chemists can substantiate results by synthesis products and thus provide physical evidence of the research work and its quality. Where the reaction products obtained are stable, they can be collected, stored and registered so that, if suitable, the work result can be used directly for further studies. Examples of such direct re-use could be the independent reproduction of experiments, the use of the chemical samples for reactions or the analysis of the samples by characterization or screening techniques. The reusable collection of samples promotes transparency. Numerous scenarios are conceivable in which such stored substances accelerate knowledge gain, especially

when they are unambiguously linked to other research output like journal publications and research data.

Initiatives that provide access to scientific physical collections of materials exist already in other disciplines, such as Geosciences, Microbiology, and Botanical Science, where collecting and archiving samples for further re-use is widely accepted as an important contribution to scientific work. Examples are the scientific collections of the U.S. Geological Survey⁸ and the collections of drilling cores at the IODP (International Ocean Discovery Program) Core Repository⁹. 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 Australia¹⁰, EU-Openscreen¹¹, Chimiotheque National (ChemBioFrance)¹², and the Boston University Center for Molecular Discovery (BU-CMD)¹³. The known initiatives collect and register the chemical substances for medical or pharmaceutical application purposes but, as far as we know, not for a general, open subsequent use of various kinds.

The starting point of the work described in this article was the aim for concepts and infrastructure that enable such a sustainable model to collect, archive, and re-use the physical results of chemical research work and connect such a concept with existing research data infrastructure in chemistry.

RESULTS

Principles and concept design: Scientific outcome in chemistry consists of data and materials and new studies also depend on both. Therefore, we suggest complementing the FAIR (meta)data principles with a concept for materials that address sustainable access to and re-use of physical objects such as chemical samples wherever possible. The concept should make it possible to secure

physical research results, verify the gained results and increase the re-usability of the materials in addition to the re-use 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 **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (= **FAIR** metadata for samples). In the second step, the samples would need to be registered and stored in a materials archive to be **A**ccessible, and made available under suitable access policies and rules to be physically **R**eusable (= **AR** material for samples). This concept is further referred to as the FAIR-AR samples concept (Figure 1).

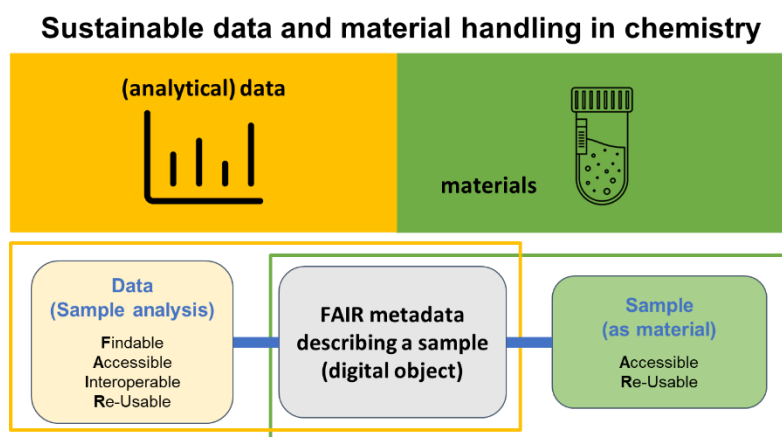


Figure 1. Extending the principles of FAIR data (usually considered in chemistry is the left part in yellow including FAIR metadata for a sample) with a concept for accessible and reusable samples consisting of the same FAIR metadata of a sample and additional measures meeting the requirements of materials'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

The FAIR-AR 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, to which both, the physical sample and the gained data refer. Efficient concepts that intend to consider the FAIR deposition of research data and the deposit of FAIR-AR samples could, therefore, closely link research data deposition to sample deposition. A design to this is depicted in Figure 2, showing a virtual sample representation described by rich standardized metadata, which is complemented by (1) information on further relevant data assigned to the sample (Figure 2, left), and (2) information on available samples' location and its unique registry or reference number (Figure 2, right). The information on the samples' location must come from the registration 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.

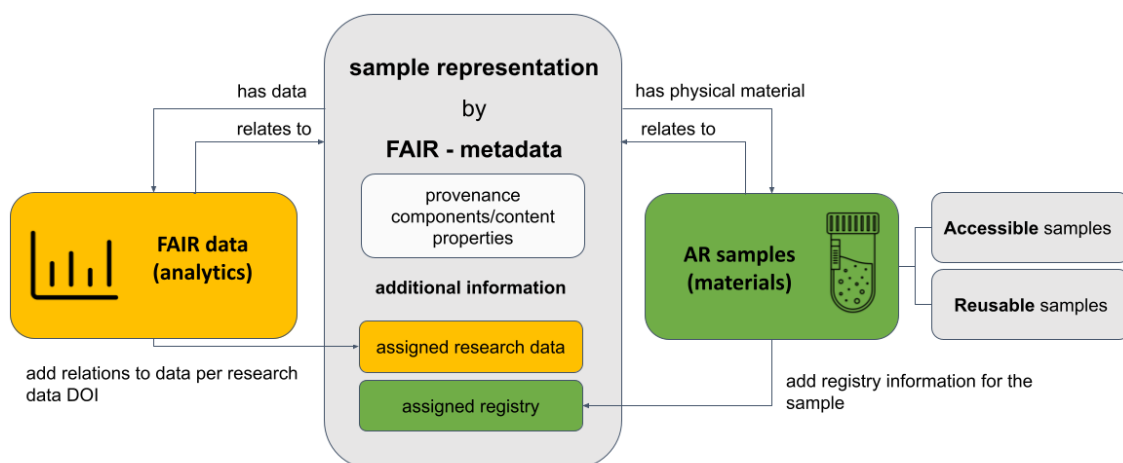


Figure 2. Design of the infrastructure towards more sustainable scientific work in chemistry: The concepts of FAIR data, e.g. analytical measurements, are combined with the concept of Accessible and reusable samples through the virtual sample representation given by a sample's FAIR metadata. [ref Icons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

The 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 (Figure 2) and the often observed issue that chemical samples are only efficiently reusable by others if they are described sufficiently - which also includes the full set of available research data or analytical data. Therefore, methodical aspects and scientific reasons can speak in favor of a combination of a sample archival system and a research data repository. Such a combination was realized with the Chemotion repository and the Molecule Archive at KIT (Karlsruhe Institute of Technology). The 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 identifiers.¹⁴ Especially for organic chemistry, the repository offers additional functions for direct viewing and analysis of the stored data by discipline-specific research software¹⁵ and thus enables an easy way of re-using 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 Karlsruhe Institute of Technology (KIT) which enables the registration, validation, and collection of chemical substances. The substances are preserved for documentation and re-use 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 repository Chemotion is free of charge.

As the infrastructure of the Chemotion repository is already in place to publish data and metadata of samples as a main data entity also including the generation of DOIs, the research data repository is used in the herein described concept for the generation of the virtual sample representation to make physical samples findable. Both systems have to exchange information *via* a defined protocol to enable the additional matching of samples' virtual representation with the content of the Molecule Archive. This protocol checks if sample representations in the Chemotion repository match registered samples in the Molecule Archive (Figure 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 dependencies 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 those samples within the Molecule Archive that may also be publicly listed are queried through the Chemotion repository (Figure 3, query to “open” collection given in green, right panel) and then visible through the repository’s graphical user interface (GUI) (Figure 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 (Figure 3, step (3)).

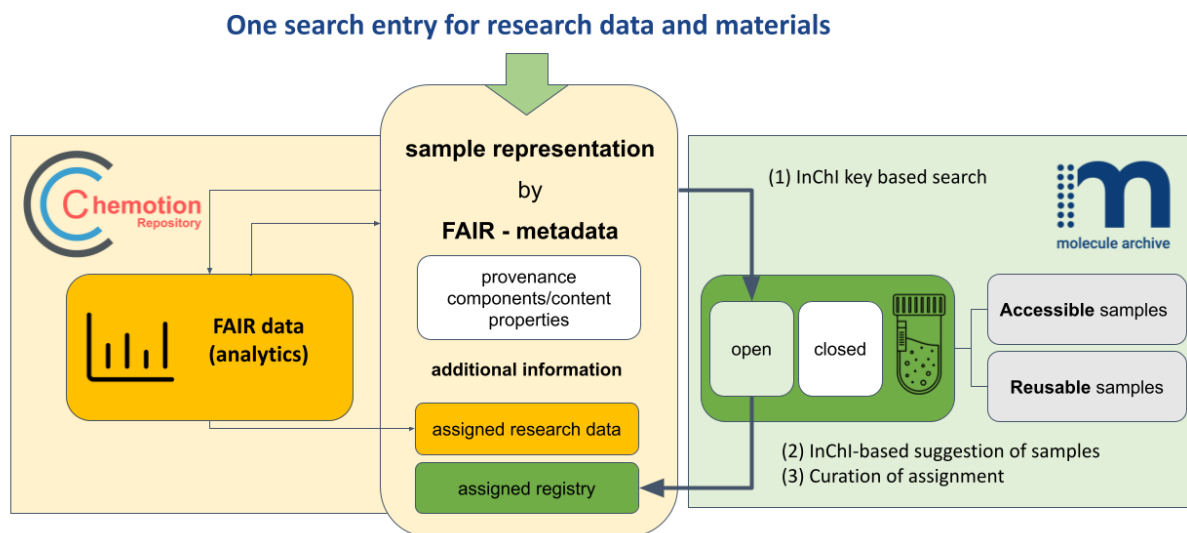


Figure 3. Technical implementation of the concept described in Figure 2. The Chemotion repository is used as an infrastructure component to make research data and samples findable and accessible through one search entry, which is built on the sample representation. [ref Icons 3: Tube: *Digithrust from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

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,^{16, 17, 18, 19} and metal-organic frameworks (MOFs)²⁰. Each of them 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 chemical study before or after their submission to the Molecule Archive and the deposition of the data to the Chemotion repository. Scientists from different institutions tested the robustness of the infrastructure with respect to the digital transfer and deposition of data, as well as the physical transfer and deposition of samples (Figure 4). They contributed to adapting the infrastructure components to their scientific needs. The proposed process to gain FAIR-AR samples includes five main steps: (1) The isolation of the samples from a chemical reaction or a natural product isolation and their analytical characterization, (2) the entering of the samples' metadata and the deposition of the analytical data in a research data repository, (3) the collection of samples and the (4) shipping to the location of the materials' archive (in our example: Molecule Archive at KIT) including the registration of the samples to the archive by the archive's staff, and finally (5) the publication of the results in scientific journals (Figure 5A). The partner sites and research groups (red circles) that established the described FAIR-AR samples model and provided the first use cases for open data and open materials in different subdisciplines in chemistry are represented schematically in Figure 5B. The implemented process and the order of the proposed steps considers the requirements of the current research system in Germany, including new requirements to openly provide research data along with scientific publications^{21,22}. According to this, the research data

should be submitted to a repository before or at least in parallel with the publication of manuscripts in scientific journals 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 accessible only 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 Figure 5A) could strengthen the publication, as the materials archive can confirm the provision of the sample and hand out information on the quality/purity of the samples. Suppose compounds are already available in the Molecule Archive when the referring work is to be published. In that case the information on the availability of the materials can be added to the publication - very similar to the referencing of data deposition in repositories. The sending of materials to an archive after the publication of scientific results (blue arrows in Figure 5A) has the advantage that materials are quickly available if reviewers need additional data from the authors, in cases where only limited material is available.

Chemotion-Repository	My DB	Data publ	Molecule Archive	Embargoed Publications	Newsroom	How-To	Simone Gräßle
	Formula C ₁₄ H ₁₂ N ₂ O ₆	Provided by Timo Sehn Soft Matter Lab, KIT Karlsruhe	ID CRS-14928	Embargo TGS_2020-10-21	Analyses 7		Formula C ₁₂ H ₁₁ ClN ₂ O ₂ S
	Formula C ₁₂ H ₁₄ N ₆ O ₂	Provided by Andreas Link Andreas Link Group	ID CRS-33437	Embargo	Analyses 4		Formula C ₁₇ H ₁₈ N ₂ O ₅
	Formula C ₁₅ H ₁₄ Cl ₂ O ₂ S ₂	Provided by Patrick Hodapp Stefan Bräse Group	ID CRS-29767	Embargo PH2_2023-01-27	Analyses 3		Formula C ₁₉ H ₂₀ N ₂ S
							Provided by Maria Manuel Marques Maria Manuel Marques Group
							ID CRS-28273
				Embargo CWG_2022-11-23	Analyses 2		
							Provided by Rachel Janßen Lena Daumann Group
							ID CRS-25869
				Embargo RAJ_2022-08-25	Analyses 4		
							Provided by Fabian Fink Sonja Herres-Pawlis Group
							ID CRS-17304
				Embargo	Analyses 7		

Figure 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).

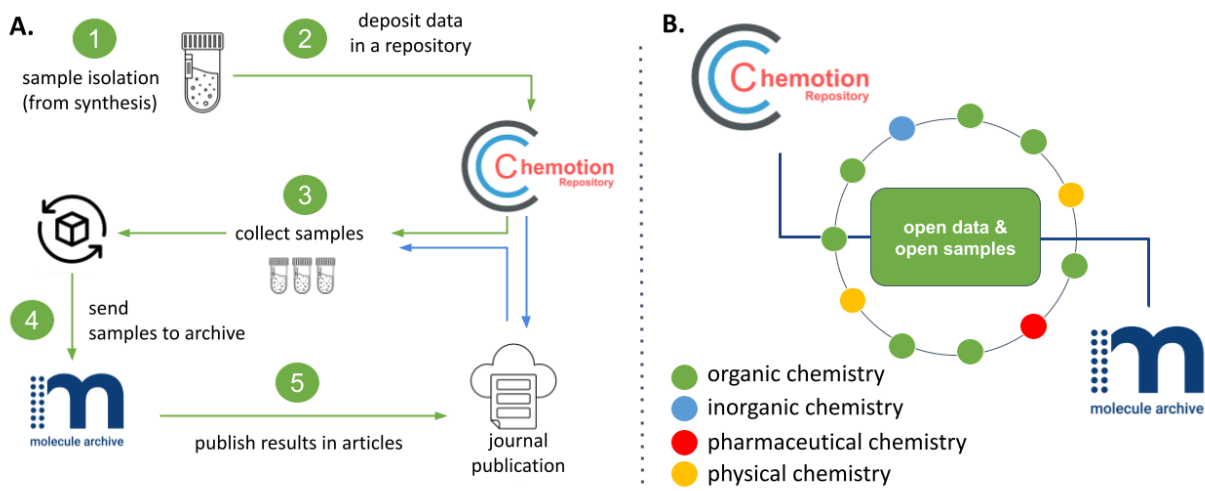


Figure 5. A. Proposed workflow to gain FAIR-AR samples along with the publication of scientific results and data. The workflow consists of five steps, from preparing samples and their characterization to depositing data and physical material, which can be done according to the suggested order (green) 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 sub-disciplines 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*]

As demonstrated with the example sites shown in Figure 5B, the proposed workflow of Figure 5A leads to centrally available data and materials, enabling sharing of results with other scientists. The impact of the established infrastructure and workflows on open and sustainable science is

described as an assessment according to the findability, accessibility, interoperability and re-usability of the materials metadata given in the sample representation and the accessibility and re-usability of the physical material.

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 text and structure searching, 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 analytical research data and the 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 CHMO²³,²⁴, CHEMINF²⁵, OBI²⁶, and ChEBI²⁷ wherever possible. A protocol for metadata harvesting (OAI PMH)²⁸ allows the retrieval of the provided metadata per API, the metadata information is available through DataCite and the metadata can be downloaded through the GUI of the repository. Two examples of typical metadata schemas have been included with the Supplemental Information (SI, Fig S1 and Fig S2).

The currently established GUI and API-based metadata concepts will be extended by JSON-LD metadata schemas in the future. The benefit of using JSON-LD with community-agreed schemas is that they are better suited to provide semantically rich and domain-specific metadata, and linking

of property values to defined terms from ontologies is more explicit. A first draft of such an implementation for samples was embedded in the samples' representation in the Chemotion repository (Figure 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 NFDI4Chem²⁹.

Accessible sample metadata: The repository Chemotion supports the OAI-PMH protocol, which is a widely used protocol for exposing metadata records in a standardized way and which can 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 Chemotion¹. Similarly, by using the "GetRecord" verb and applying filters such as the metadata prefix "oai_DataCite" and specifying an identifier like a sample 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 service³⁰. 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 are accessible, even when the materials are no longer available, and allow the 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

¹ applicable after April 1, 2023

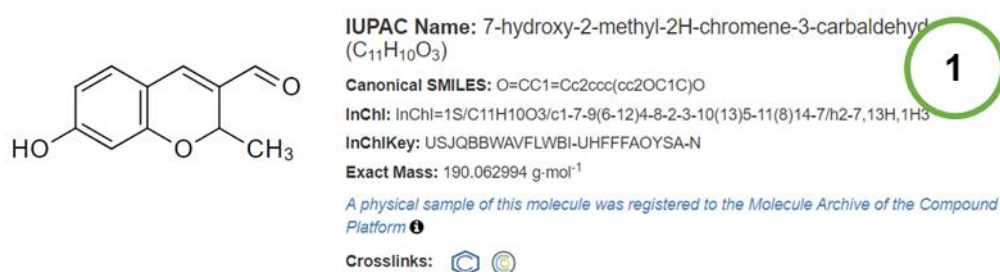
standards such as InChI, InChI keys and canonical SMILES strings wherever applicable. The terminologies used in the description of the samples are chosen from domain relevant ontologies and are assigned to the identifiers of the mentioned ontologies.



Reusable sample metadata: The metadata for the sample in the Chemotion repository are described as comprehensively as possible, including the description of the components characterizing the samples, their properties, and additional references to analytical data and the synthetic origin of the sample (chemical reaction). In particular, the reference to the analytical data and the reactions allows the re-use of the data, as this information is needed for chemists to reproduce the experiments. A plurality of additional data describing the sample further can be gained from the latter two references.


To gain FAIR-AR samples, the FAIR metadata for samples are completed with measures to gain accessible and reusable materials/samples:

Accessible materials: The first essential step to enable access to the samples is their registration in the Molecule Archive. As soon as the samples are physically deposited and registered in the Molecule Archive, the visibility of the available samples is managed *via* the website of the Chemotion repository. Scientists interested in re-using the samples can place their sample request directly through the repository's interface (Figure 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 for access to the 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 (MTA³¹) 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 re-users of the chemical compounds and if the material can be sent, the details and conditions for such re-use are clarified.



 **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](https://doi.org/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](https://doi.org/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](#)





 

Figure 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 (left) was created from a screenshot from the Chemotion repository and changed regarding the layout for better readability (see SI section 6 for the original screenshot as obtained).

Reusable materials: The re-use of samples available in the Molecule Archive is supported by quality assurance measures, framework agreements, consulting, and support in handling of the re-use. 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. Compared to the 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' re-users, and the Molecule Archive. The Molecule Archive supports the re-use 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 collaborators.³¹ Further, the Molecule Archive organizes the communication between the participating scientists as needed. The provision of the material for subsequent use is initiated by the preparation of the sample

transfer. After clarification of all technical and legal issues, the materials are shipped according to common standards for chemical compounds.

DISCUSSION

Using examples from the sub-disciplines of organic, pharmaceutical, inorganic, and experimental physical chemistry, a model was demonstrated 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 still need to be solved or are not solvable and will permanently represent a limitation of 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 current setup.

(2) So far, the method has only been used to provide data and materials related to a defined chemical structure. Therefore, no substance mixtures or natural extracts have been submitted as FAIR-AR samples yet. This is possible in principle, but the Chemotion repository is currently geared towards pure substances; this 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). Infrastructure and concepts can still be used also

for samples beyond the current scope - such as SURMOFs - but the physical archival and the evaluation of the gained material would need other solutions.

(4) In some areas of chemistry, samples are generated that cannot be unambiguously described by any unique chemical structure. In these cases, the FAIR-AR 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 due to the lack of structural descriptors in the metadata.

Other obvious hurdles for accessible and reusable samples 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 re-used almost indefinitely by granting a license and choosing the appropriate data infrastructure with regard to the number of re-users, various interests may have to be weighed against each other for the re-use of material - and the resources may be exhausted even if there is a need for further re-use. Currently, the Molecular Archive cannot provide a universal solution to this problem because the re-use 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 post-use purposes requires - at least at the time of the first provision of materials - time for preparation due to the usually required preparation of an MTA by 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 re-use is to be created comparable to the licensing of research data.

Despite the obvious challenges of providing materials, the combination of FAIR data and FAIR-AR 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 work and allows a direct link of a publication to all gained results – the research data and the available material.

Knowing that the implementation of the FAIR data principles has not been achieved among all scientists, the additional claim for FAIR-AR samples might seem to be challenging. Nevertheless, the establishment of a standardized process for FAIR-AR samples can be very easy once the concept of FAIR data is adopted: the infrastructure supporting the access and re-use of chemical samples already exists, and the effort for the single scientist as the material producer is low if the initial agreements between the partner sites and the Molecule Archive are done. 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' re-users - this makes materials storage and provision an attractive aim that could foster the broad application of the FAIR-AR 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 re-users - approving that the provision of data can result directly in more visibility and impact for the scientists sharing their research results.^{32, 33, 34, 35, 36, 37, 38}

ONLINE METHODS

Software: The infrastructure described in this article is built with the use of open source software that was developed at KIT and was 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.^{39,40} Further extensions of Chemotion ELN, including submission and reviewing workflow, provide the necessary functionality to operate the research data repository Chemotion.^{14,41} The source code for the Chemotion repository can be obtained from GitHub⁴². For the operation of the Molecule Archive, Chemotion ELN was also used and adapted with a plugin⁴³ to keep additional information on the sample information as provided by the owner and was extended with Foreign Data Wrappers (FDW)⁴⁴ 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 Zenodo.⁴⁵

Submission of data to the Chemotion repository: The submission of data to the Chemotion repository is – in the examples described in this article – the next step towards FAIR data and FAIR-AR samples. It enables the generation of the FAIR metadata of the samples' virtual representation and the storage of FAIR research data. The data uploaded can be managed through either the GUI of the repository or a transfer of data from an ELN to the repository. Both ways to get data available in the repository are established and were used by the groups who contributed to this work. Both processes include 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 automatic

generation of, e.g., molecular descriptors facilitate the work of the scientists as the most suitable descriptors for a search for chemical compounds are generated by the software without additional efforts. 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 available information on 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 repository.⁴⁶ The better the data is prepared according to the given best practice examples, the FAIRer is the data that is accessible in the repository in the end. After submitting the dataset, the data is reviewed for plausibility and completeness, before it is open and accessible. As soon as the virtual sample representation is visible (along with the data), 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-AR 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 re-use of compounds under clearly defined rules and is therefore the preferred way of sharing materials. Together with five exemplarily chosen partner sites, KIT elaborated a standard MTA several years ago, which is used as a routine process to introduce new partner sites to provide materials to the Molecule Archive.³¹ The MTA defines the rights of the material providers, it regulates intellectual property ownership, the handling of obtained or transmitted data, and the publication of results. This is in particular important if the re-users of the materials gain scientific results with the provided compounds, and these results should be published together.

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 be accessed through the Chemotion repository, and the sample is visible in the GUI of the repository along with the virtual sample representation (as depicted in Figure 6 and the SI).

Declarations

Availability of data and materials

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 re-use through the provided request form (chapter 2). The 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). Finally, 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-AR open samples (accessed on June, 5 2023) are cited in file 2.

Funding

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).

Acknowledgements

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 facilitates 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.

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. 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*. (National Academies Press, 2002). doi:10.17226/10348.
9. IODP | International Ocean Discovery Program - Bremen Core Repository. <https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html>. Date accessed: June 11, 2023.
10. Simpson, M. & Poulsen, S.-A. An overview of Australia's compound management facility: the Queensland Compound Library. *ACS Chem. Biol.* **9**, 28–33 (2014).
11. Brennecke, P. *et al.* EU-OPENSREEN: A Novel Collaborative Approach to Facilitate Chemical Biology. *SLAS Discov* **24**, 398–413 (2019).
12. Mahuteau-Betzer, F. Chimiothèque Nationale - Avancées et perspectives. *médecine/sciences* **31**, 417–422 (2015).
13. Center for Molecular Discovery at Boston University. <https://www.bu.edu/articles/2016/center-for-molecular-discovery/>. Date accessed: June 11, 2023.
14. 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).
15. 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).
 16. Fink, F. Dichlorocopper;2,2-di(pyrazol-1-yl)ethanamine (C₈H₁₁Cl₂CuN₅), Chemotion repository, <http://dx.doi.org/10.14272/AMQWWHVPPZUOKP-UHFFFAOYSA-L.1>. (2020).
 17. Kalyakina, A. C₄₉H₃₆EuF₃N₄O₆, Chemotion repository, <http://dx.doi.org/10.14272/XSDMQMVIDCRHRR-UHFFFAOYSA-K.1>. (2022).
 18. Holzhauer, L. [2-(4-Butyl-1H-1,2,3-triazol-1-yl)-3-phenylquinoxaline]bromotricarbonylrhenium(I), Chemotion repository, <http://dx.doi.org/10.14272/BKFBOTQHKAJKOY-UHFFFAOYSA-M.1>. (2022).
 19. Schissler, C. C₇₆H₄₇N₉NiZn, Chemotion repository, <http://dx.doi.org/10.14272/JVGPCXSCDWSVRU-HWNMUZRGSA-N.1>. (2022).
 20. Pilz, L. Dicopper;benzene-1,3,5-tricarboxylate, Chemotion repository, <http://dx.doi.org/10.14272/JMHFTDFPRQWUAN-UHFFFAOYSA-H.22>. (2022).
 21. *Guidelines for Safeguarding Good Research Practice: Code of Conduct*. (Deutsche Forschungsgemeinschaft, 2019).
 22. Deutsche Forschungsgemeinschaft. Guidelines for Safeguarding Good Research Practice. Code of Conduct, <http://dx.doi.org/10.5281/zenodo.6472827>. <https://zenodo.org/record/6472827> (2022) doi:10.5281/zenodo.6472827.
 23. McEwen, L. R. & Buntrock, R. E. *The Future of the History of Chemical Information (ACS Symposium, Band 1164)*. (Am Chem Soc, 2015).
 24. 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).
 25. Hastings, J. *et al.* The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One* **6**, e25513 (2011).
 26. Bandrowski, A. *et al.* The Ontology for Biomedical Investigations. *PLoS One* **11**, e0154556 (2016).

27. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–9 (2016).
28. Lagoze, C. & Van de Sompel, H. The Open Archives Initiative: Building a low-barrier interoperability framework, <https://www.openarchives.org/documents/jcdl2001-oai.pdf>. Date accessed: June 11, 2023.
29. 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.
30. NFDI4Chem Search Service, <https://search.nfdi4chem.de/>. Date accessed: Jun 20, 2023.
31. 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).
32. Macara, J. *et al.* Practical synthesis and biological screening of sulfonyl hydrazides. *Org. Biomol. Chem.* **21**, 2118–2126 (2023).
33. Apweiler, M. *et al.* Functional Selectivity of Coumarin Derivates Acting via GPR55 in Neuroinflammation. *Int. J. Mol. Sci.* **23**, (2022).
34. 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).
35. Lei, W. *et al.* Droplet microarray as a powerful platform for seeking new antibiotics against multidrug-resistant bacteria. *Adv Biol (Weinh)* e2200166 (2022) doi:10.1002/adbi.202200166.
36. Hofmann, D. *et al.* A small molecule screen identifies novel inhibitors of mechanosensory nematocyst discharge in Hydra. *Sci. Rep.* **11**, 20627 (2021).
37. König, G. *et al.* Rational prioritization strategy allows the design of macrolide derivatives that overcome antibiotic resistance. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
38. Raudszus, R. *et al.* Pharmacological inhibition of TRPV2 attenuates phagocytosis and lipopolysaccharide-induced migration of primary macrophages. *Br. J. Pharmacol.* (2023) doi:10.1111/bph.16154.
39. Tremouilhac, P. *et al.* Chemotion ELN: an Open Source electronic lab notebook for chemists in

- academia. *J. Cheminform.* **9**, 54 (2017).
40. 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).
 41. Tremouilhac, P. *et al.* The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry*. *Angew. Chem. Int. Ed Engl.* **59**, 22771–22778 (2020).
 42. *Github reference for chemotion REPO: A Repository based on chemotion ELN*, https://github.com/ComPlat/chemotion_REPO. (Github).
 43. *GitLab reference for a Module to enable X-Vial listing in chemotion ELN*, <https://git.scc.kit.edu/ComPlat/Xvial>. *GitLab*.
 44. Chapter 56. Writing a foreign data wrapper. *PostgreSQL Documentation* <https://www.postgresql.org/docs/12/fdwhandler.html> (2023).
 45. Lin, C.-L., Huang, P. C., Tremouilhac, P., Jung, N. & Le, L. *ComPlat/chemotion_REPO: Chemotion Repository 1.1.0*. (2023). doi:10.5281/zenodo.8028033.
 46. Documentation for Chemotion Repository, <https://www.chemotion.net/docs/repo>. Date accessed: June 11, 2023.