The LOTUS Initiative for Open Natural Products Research: Knowledge Management through Wikidata

This manuscript (permalink) was automatically generated from lotusnprod/lotus-manuscript@bb47ea0 on May 27, 2021.

Authors

- **Adriano Rutz**
  0000-0003-0443-9902 · adafede · adafede
  School of Pharmaceutical Sciences, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland; Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland

- **Maria Sorokina**
  0000-0001-9359-7149 · mSorok · ms_sorok
  Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07732 Jena, Germany

- **Jakub Galgonek**
  0000-0002-7038-544X · galgonek · JGalgonek
  Institute of Organic Chemistry and Biochemistry of the CAS, Flemingovo náměstí 2, 166 10, Prague 6, Czech Republic

- **Daniel Mietchen**
  0000-0001-9488-1870 · Daniel-Mietchen · EvoMRI
  School of Data Science, University of Virginia, Dell 1 Building, Charlottesville, Virginia 22904, United States

- **Egon Willighagen**
  0000-0001-7542-0286 · egonw · egonwillighagen
  Dept of Bioinformatics-BiGCaT, NUTRIM, Maastricht University, Universiteitssingel 50, NL-6229 ER, Maastricht, The Netherlands

- **Arnaud Gaudry**
  0000-0002-3648-7362 · ArnaudGaudry
  School of Pharmaceutical Sciences, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland; Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland

- **James G. Graham**
  0000-0002-7114-8921
  Center for Natural Product Technologies and WHO Collaborating Centre for Traditional Medicine (WHO CC/TRM), Pharmacognosy Institute; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States; Department of Pharmaceutical Sciences; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States

- **Ralf Stephan**
  0000-0002-4650-631X · rwst
  Ontario Institute for Cancer Research (OICR), 661 University Ave Suite 510, Toronto, Canada

- **Roderic Page**
  0000-0002-7101-9767 · rdmpage · rdmpage
  IBAHCM, MVLS, University of Glasgow, Glasgow, United Kingdom

- **Jiří Vondrášek**
  0000-0002-6066-973X
  Institute of Organic Chemistry and Biochemistry of the CAS, Flemingovo náměstí 2, 166 10, Prague 6, Czech Republic

- **Christoph Steinbeck**
  0000-0001-6966-0814 · steinbeck · csteinbeck
  Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07732 Jena, Germany

- **Guido F. Pauli**
  0000-0003-1022-4326
  Center for Natural Product Technologies and WHO Collaborating Centre for Traditional Medicine (WHO CC/TRM), Pharmacognosy Institute; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States; Department of Pharmaceutical Sciences; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States

- **Jean-Luc Wolfender**
  0000-0002-0125-952X
  School of Pharmaceutical Sciences, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland; Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland

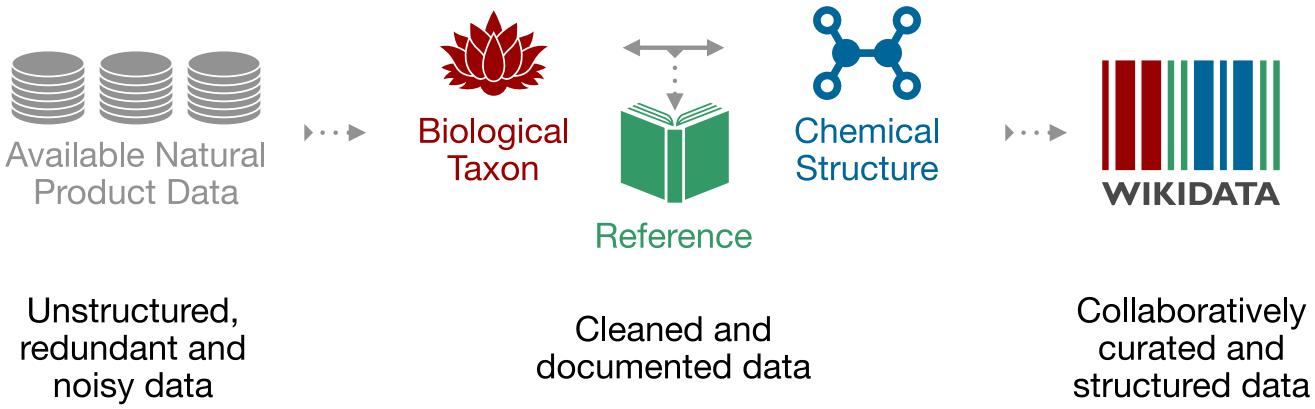
- **Jonathan Bisson** ✉
  0000-0003-1640-9989 · bjonnh · Bjonnh
  Center for Natural Product Technologies and WHO Collaborating Centre for Traditional Medicine (WHO CC/TRM), Pharmacognosy Institute; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States; Department of Pharmaceutical Sciences; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States

- **Pierre-Marie Allard** ✉
  0000-0003-3389-2191 · oolonek · NatprodCbn
  School of Pharmaceutical Sciences, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland; Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland; Department of Biology, University of Fribourg, Chemin du Musée 10, 1700 Fribourg, Switzerland

✉ — technical correspondence preferred *via* GitLab Issues. Otherwise, address correspondence to research@bjonnh.net, and pierre-marie.allard@unifr.ch.

Abstract

Contemporary bioinformatic and chemoinformatic capabilities hold promise to reshape knowledge management, analysis and interpretation of data in natural products research. Currently, reliance on a disparate set of non-standardized, insular, and specialized databases presents a series of challenges to data access, either within the discipline or to integration and interoperability between related domains. The fundamental elements of exchange are referenced structure-organism pairs that establish relationships between distinct molecular structures and the living organisms from which they were identified. Consolidating and sharing such information *via* an open platform has strong transformative potential for natural products research and beyond. This is the ultimate goal of the newly established LOTUS initiative, which has now completed the first steps toward the harmonization, curation, validation and open dissemination of 700,000+ referenced structure-organism pairs. LOTUS data is hosted on Wikidata and regularly mirrored on https://lotus.naturalproducts.net. Data sharing within the Wikidata framework broadens data access and interoperability, opening new possibilities for community curation and evolving publication models. Furthermore, embedding LOTUS data into the vast Wikidata knowledge graph will facilitate new biological and chemical insights. The LOTUS initiative represents an important advancement in the design and deployment of a comprehensive and collaborative natural products knowledge base.



Introduction

Evolution of Electronic Natural Products Resources

Natural Products (NP) research is a transdisciplinary field with wide-ranging interests: from fundamental structural aspects of naturally-occurring molecular entities to their effects on living organisms and extending to the study of chemically-mediated interactions within entire ecosystems. Despite the ambiguous definition of "natural" ("All natural," 2007), the basis of our definition of a NP as a chemical entity *found in* a living organism is predicated on the identification of the explicit relationship between a naturally-occurring chemical entity and its source organism. A third fundamental element of a structure-organism pair is a reference to the experimental evidence that establishes the linkages between a chemical structure and a biologicl organism and a future-oriented electronic NP resource should contain only fully-referenced structure-organism pairs.

Reliance on data from the NP literature presents many challenges. The assembly and integration of NP occurrences into an inter-operative platform relies primarily on access to a heterogeneous set of databases (DB) whose content and maintenance status are critical factors in this dependency (Tsugawa, 2018). A tertiary inter-operative NP platform is thus dependent on a secondary set of data that has been selectively annotated into a DB from primary literature sources. The experimental data itself reflects a complex process involving collection or sourcing of natural material (and establishment of its identity), a series of material transformation and separation steps and ultimately the chemical or spectral elucidation of isolates. The specter of human error and the potential for the introduction of biases are present at every phase of this journey. These include publication biases (Lee et al., 2013), such as emphasis on novel and/or bioactive structures in the review process, or, in DB assembly stages, with selective focus on a specific compound class or a given taxonomic range, or disregard for annotation of other relevant evidence that may have been presented in primary sources. Temporal biases also exist: a technological "state-of-the-art" when published can eventually be recast as anachronistic.

The advancement of NP research has always relied on the development of new technologies. In the past century alone, the rate at which unambiguous identification of new NP entities from biological matrices can be achieved has been reduced from years to days and in the past few decades, the scale at which new NP discoveries are being reported has increased exponentially. Without a means to access and process these disparate NP data points, information is fragmented and scientific progress is impaired (Balietti et al., 2015). To this extent, contemporary bioinformatic tools enable the (re-)interpretation and (re-)annotation of (existing) datasets documenting molecular aspects of biodiversity (Jarmusch et al., 2020; Mongia and Mohimani, 2021).

While large, well-structured and freely accessible DB exist, they are often concerned primarily with chemical structures (e.g. PubChem (Kim et al., 2019), with over 100M entries) or biological organisms (e.g. GBIF ("GBIF.org," 2020), with over 1,400M entries), but scarce interlinkages limit their application for documentation of NP occurrence(s). Currently, no open, cross-kingdom, comprehensive, computer-interpretable electronic NP resource links NP and their producing organisms, along with referral to the underlying experimental work. This shortcoming breaks the crucial evidentiary link required for tracing information back to the original data and assessing its quality. Even valuable commercially available efforts for compiling NP data, such as the Dictionary of Natural Products (DNP), can lack proper documentation of these critical links.

Pioneering efforts to address such challenges led to the establishment of KNApSAck (Shinbo et al., 2006), which is likely the first public, curated electronic NP resource of referenced structure-organism pairs. KNApSAck currently contains 50,000+ structures and 100,000+ structure-organism pairs. However, the organism field is not standardized and access to the data is not straightforward. One of the earliest-established electronic NP resources is the NAPRALERT dataset (Graham and Farnsworth, 2010), which was compiled over five decades from the NP literature, gathering and annotating data derived from over 200,000 primary literature sources. The dataset contains 200,000+ distinct compound names and structural elements, along with 500,000+ records of distinct, fully-cited structure-organism pairs. In total, NAPRALERT contains over 900,000 such records, due to equivalent structure-organism pairs reported in different citations. NAPRALERT is not an open platform, employing an access model that provides only limited free searches of the dataset. Finally, the NPAtlas (van Santen et al., 2019) is a more recent project aimed at complying with the FAIR (Findability, Accessibility, Interoperability and Reuse) guidelines for digital assets (Wilkinson et al., 2016) and offering web access. While the NPAtlas encourages submission of new compounds with their biological source, it focuses on microbial NP and ignores a wide range of biosynthetically active organisms found in the Archaeplastida.

Building on experience with the recently published COlleCtion of Open NatUral producTs (COCONUT) (Sorokina et al., 2021), the LOTUS initiative seeks to address the aforementioned shortcomings. At its current stage of development, LOTUS disseminates 700,000+ referenced structure-organism pairs. After extensive data curation and harmonization, each pair was standardized at the chemical, biological and reference levels. These efforts and experiences represent an intensive preliminary curatorial phase and the first major step towards providing a high quality, computer-interpretable knowledge base capable of transforming NP research data management from a classical (siloed) database approach to an optimally-shared resource.

Accommodating Principles of FAIRness and TRUSTworthiness for Natural Products Knowledge Management

In awareness of the multi-faceted pitfalls associated with implementing, using and maintaining classical scientific DBs (Helmy et al., 2016), and to enhance current and future sharing options, the LOTUS initiative selected the Wikidata platform for disseminating its resources. Since its creation, Wikidata has focused on cross-disciplinary and multilingual support. Wikidata is curated and governed collaboratively by a global community of volunteers, about 20,000 of which are contributing monthly. Wikidata currently contains more than 1 billion statements in the form of subject-predicate-object triples. Triples are machine-interpretable and can be enriched with qualifiers and references. Within Wikidata, data triples correspond to approximately 90 million entries, which can be grouped into classes as diverse as countries, songs, disasters, or chemical compounds. The statements are closely integrated with Wikipedia and serve as the source for its infoboxes. Various workflows have been established for reporting such classes, particularly those of interest to life sciences, such as genes, proteins, diseases, drugs, or biological taxa (Waagmeester et al., 2020).

Building on the principles and experiences described above, the present report introduces the development and implementation of the LOTUS workflow for NP occurrences' curation and dissemination, which applies both FAIR and TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles (Lin et al., 2020). LOTUS' data upload and retrieval procedures ensure optimal accessibility by the research community, allowing any researcher to contribute, edit and reuse the data with a clear and open CC0 license (Creative Commons 0).

Despite many advantages, Wikidata hosting has some notable, yet manageable drawbacks. While its SPARQL query language offers a powerful way to query available data, it can also appear intimidating to the less experienced user. Furthermore, some typical queries of molecular electronic NP resources such as structural or spectral searches are not yet available in Wikidata. To bridge this gap, LOTUS is hosted in parallel at https://lotus.naturalproducts.net (LNPN) within the naturalproducts.net ecosystem. The Natural Products Online is a portal for open-source and open-data resources for NP research. In addition to the generalistic COCONUT and LNPN databases, the portal will enable hosting of arbitrary and skinned collections, themed in particular by species or taxonomic clade, by geographic location or by institution, together with a range of cheminformatics tools for NP research. LNPN is periodically updated with the latest LOTUS data. This dual hosting provides an integrated, community-curated and vast knowledge base (*via* Wikidata), as well as a NP community-oriented product with tailored search modes (*via* LNPN).

The LOTUS initiative and its multiple data interaction options establish the basis for transparent and sustainable access, sharing and creation of knowledge on NP occurrence. LOTUS represents an important advancement in the design and deployment of a comprehensive and collaborative NP knowledge base. More broadly, the LOTUS initiative fosters cross-fertilization of the fields of chemistry, biology and associated disciplines.

Results & Discussion

This section is structured as follows: first, we present an overview of the LOTUS initiative at its current stage of development. The central curation and dissemination elements of the LOTUS initiative are then explained in detail. The third section addresses the interaction modes between LOTUS and its end-users, including data retrieval, addition and editing. The final section is dedicated to the interpretation of LOTUS data and illustrates the dimensions and qualities of the current LOTUS dataset from chemical and biological perspectives.

Blueprint of the LOTUS Initiative

Building on the standards established by the related Wikidata project (Chemistry, Taxonomy and Source Metadata), a NP chemistry-oriented subproject was created (Chemistry/Natural products). Its central data consists of three minimal sufficient objects:

- A *chemical structure object*, with associated Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988), International Chemical Identifier (InChI) (Heller et al., 2013) and InChIKey (a hashed version of the InChI).
- A *biological organism object*, with associated taxon name, the taxonomic DB where it was described and the taxon ID in the respective DB.
- A *reference object* describing the structure-organism pair, with the associated article title and a Digital Object Identifier (DOI), a PubMed (PMID), or PubMed Central (PMCID) ID.

As data formats are largely inhomogeneous among existing electronic NP resources, fields related to chemical structure, biological organism and references are variable and essentially not standardized. Therefore, LOTUS implements multiple stages of harmonization, cleaning and validation (Figure 1, stages 1 to 3). LOTUS employs a Single Source of Truth (SSOT, Single source of truth) to ensure data reliability and continuous availability of the latest curated version of LOTUS data in both Wikidata and LNPN (Figure 1, stage 4). The SSOT approach consists of a PostgreSQL DB that structures links and data schemes such that every data element has a single place. The LOTUS processing pipeline is tailored to efficiently include and diffuse novel or curated data directly from new sources or at the Wikidata level. This iterative workflow relies both on data addition and retrieval actions as described in the Data Interaction section. The overall process leading to referenced and curated structure-organisms pairs is illustrated in Figure 1 and detailed below.

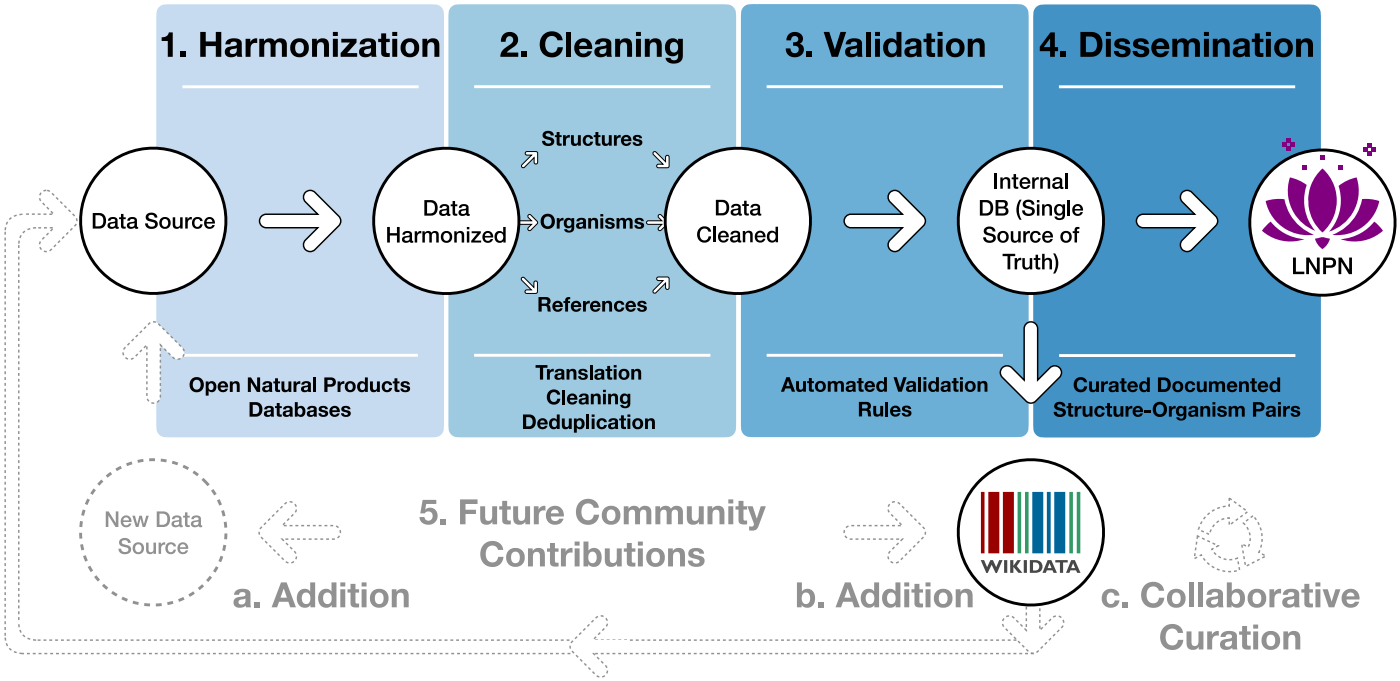


Figure 1: Blueprint of the LOTUS initiative. Data undergo a four-stage process: (1) Harmonization, (2) Cleaning, (3) Validation and (4) Dissemination. The process was designed to incorporate future contributions (5), either by the addition of new data from within Wikidata (a) or new sources (b) or *via* curation of existing data (c). The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_blueprint.svg.

By design, this iterative process fosters community participation, essential to efficiently document NP occurrences. All stages of the workflow are described on the git sites of the LOTUS initiative at https://gitlab.com/lotus7 and https://github.com/mSorok/LOTUSweb. At the time of writing, 700,000+ LOTUS entries contained a curated chemical structure, biological organism and reference and were available on both Wikidata and LNPN. As the LOTUS data volume is expected to increase over time, a frozen (as of 2021-05-23), tabular version of this dataset with its associated metadata is made available at https://osf.io/eydjs/.

Data Curation

Data Harmonization

Multiple data sources were processed as described hereafter. All publicly accessible electronic NP resources included in COCONUT that contain referenced structure-organism pairs were considered as initial input. The data were complemented with COCONUT's own referenced structure-organism pairs (Sorokina and Steinbeck, 2020a), as well as the following additional electronic NP resources: Dr. Duke ("U.S. Department of Agriculture, Agricultural Research Service. Dr. Duke's Phytochemical and Ethnobotanical Databases." 1992–2016), Cyanometdb (Jones et al., 2021), Datawarrior (Sander et al., 2015), a subset of NAPRALERT, Wakankensaku ("WAKANKENSAKU - Main Page," 2020) and DiaNat-DB (Madariaga-Mazón et al., 2021).

The contacts of the electronic NP resources not explicitly licensed as open were individually reached for permission to access and reuse data. A detailed list of data sources and related information is available as SI-1. All necessary scripts for data gathering and harmonization can be found in the lotus-processor repository in the src/1_gathering directory. All subsequent and future iterations that include additional data sources, either updated

information from the same data sources or new data, will involve a comparison of the new with previously gathered data at the SSOT level to ensure that the data is only curated once.

Data Cleaning & Validation

As shown in Figure 1, data curation consisted of three stages: harmonization, cleaning and validation. Thereby, after the harmonization stage, each of the three central objects - chemical compounds, biological organisms and reference - were cleaned. Given the data size (2.5M+ initial entries), manual validation was unfeasible. Curating the references was a particularly challenging part of the process. Whereas organisms are typically reported by at least their vernacular or scientific denomination and chemical structures via their SMILES, InChI, InChIKey or image (not covered in this work), references suffer from largely insufficient reporting standards. Despite relatively poor standardization of the initial reference field, proper referencing remains an indispensable way to establish the validity of structure-organism pairs. Better reporting practices, supported by new tools such as Scholia (Nielsen et al., 2017; Rasberry et al., 2019) and relying on Wikidata, Fatcat, or Semantic Scholar should improve reference-related information retrieval in the future.

In addition to curating the entries during data processing, 420 referenced structure-organism pairs were selected for manual validation. An entry was considered as valid if: *i)* the structure (in the form of any structural descriptor that could be linked to the final sanitized InChIKey) was described in the reference *ii)* the containing organism (as any organism descriptor that could be linked to the accepted canonical name) was described in the reference and *iii)* the reference was describing the occurrence of the chemical structure in the biological organism. This process allowed us to establish rules for automatic filtering and validation of the entries. The filtering was then applied to all entries. To confirm the efficacy of the filtering process, a new subset of 100 diverse, automatically curated and automatically validated entries was manually checked, yielding a rate of 97% of true positives. The detailed results of the two manual validation steps are reported in Supporting Information SI-2. The resulting data is also available in the dataset shared at https://osf.io/eydjs/. Table 1 shows an example of a referenced structure-organism pair before and after curation. This process resolved the structure to an InChIKey, the organism to a valid taxonomic name and the reference to a DOI, thereby completing the essential referenced structure-organism pair.

Table 1: Example of a referenced structure-organism pair before and after curation

	Structure	Organism	Reference
Before curation	Cyathocaline	Stem bark of Cyathocalyx zeylanica CHAMP. ex HOOK. f. & THOMS. (Annonaceae)	Wijeratne E. M. K., de Silva L. B., Kikuchi T., Tezuka Y., Gunatilaka A. A. L., Kingston D. G. I., J. Nat. Prod., 58, 459-462 (1995).
After curation	VFIIVOHWCNHINZ-UHFFFAOYSA-N	Cyathocalyx zeylanicus	10.1021/NP50117A020

Challenging examples encountered during the development of the curation process were compiled in an edge case table (tests/tests.tsv) to allow for automated unit testing. These tests allow a continuous revalidation of any change made to the code, ensuring that corrected errors will not reappear. The alluvial plot in Figure 2 illustrates the individual contribution of each *source* and *original subcategory* that led to the *cleaned categories*: structure, organism and reference.

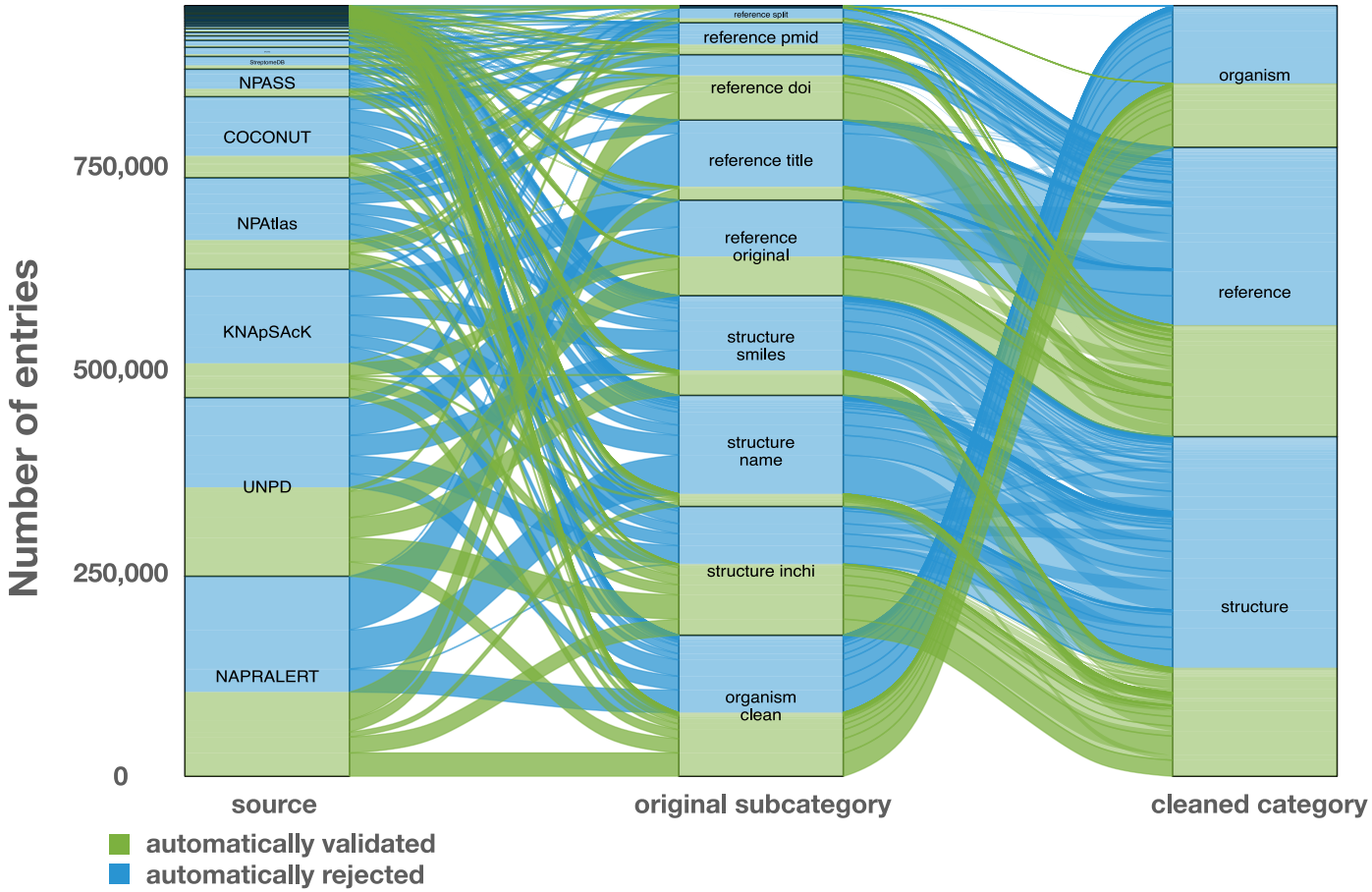


Figure 2: Alluvial plot of the data transformation flow within LOTUS during the automated curation and validation processes. The figure also reflects the relative proportions of the data stream in terms of the contributions from the various sources ("source" block, left), the composition of the harmonized subcategories ("original subcategory" block, middle) and the validated data after curation ("cleaned category" block, right). Automatically validated entries are represented in green, rejected entries in blue. The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_alluvial_plot.svg.

The figure highlights, for example, the essential contribution of the DOI category of references contained in NAPRALERT towards the current set of validated references in LOTUS. The combination of the results of the automated curation pipeline and the manually curated entries led to the establishment of four categories (manually validated, manually rejected, automatically validated and automatically rejected) of the referenced structure-organism pairs that formed the processed part of the SSOT. Out of a total of 2.5M+ pairs, the manual and automatic validation retained 700,000+ pairs (approximately 30%), which were then selected for dissemination on Wikidata. The disseminated data contains 250,000+ unique chemical structures, 30,000+ distinct organisms and 75,000+ references.

Data Dissemination

Research worldwide can benefit the most when all results of published scientific studies are fully accessible immediately upon publication (Agosti and Johnson, 2002). This concept is considered the foundation of scientific investigation and a prerequisite for effectively directing new research efforts based on prior information. To achieve this, research results have to be made publicly available and reusable. As computers are now the main investigation tool for a growing number of scientists, all research data including those in publications should be disseminated in computer-readable format, following the FAIR principles. LOTUS uses Wikidata as a repository for referenced structure-organism pairs, as this allows documented research data to be integrated with a large, pre-existing and extensible body of chemical and biological knowledge. The dynamic nature of Wikidata fosters the continuous curation of deposited data through the user community. Independence from individual and institutional funding represents another major advantage of Wikidata. The Wikidata knowledge base and the option to use elaborate SPARQL queries allow the exploration of the dataset from a sheer unlimited number of angles. The openness of Wikidata also offers unprecedented opportunities for community curation, which will support, if not guarantee, a dynamic and evolving data repository. At the same time, certain limitations of this approach can be anticipated. Despite (or possibly due to) their power, SPARQL queries are complex and often require a relatively in-depth understanding of the models and data structure. This involves a steep learning curve which tends to discourage end-users. Furthermore, traditional ways to query electronic NP resources such as structural or spectral searches are currently not within the scope of Wikidata and, thus, are addressed in LNPN. Using the pre-existing COCONUT template, LNPN hosting allows the user to perform structural searches by drawing a molecule, thereby addressing the current lack of structural search possibilities in Wikidata. Since metabolite profiling by Liquid Chromatography (LC) - Mass Spectrometry (MS) is now routinely used for the chemical composition assessment of natural extracts, future versions of LOTUS and COCONUT are envisioned to be augmented by predicted MS spectra and hosted at https://naturalproducts.net/ to allow mass and spectral-based queries. To facilitate queries focused on specific taxa (e.g., “return all molecules found in the Asteraceae family”), a unified taxonomy is paramount. As the taxonomy of living organisms is a complex and constantly evolving field, all the taxon identifiers from all accepted taxonomic DB for a given taxon name were kept. Initiatives such as the Open Tree of Life (OTL) (Rees and Cranston, 2017) will help to gradually reduce these discrepancies, the Wikidata platform should support such developments. OTL also benefits from regular expert curation and new data. As the taxonomic identifier property for this resource did not exist in Wikidata, its creation was requested and obtained. The property is now available as “Open Tree of Life ID” (P9157).

Following the previously described curation process, all validated entries have been made available through Wikidata and LNPN. LNPN will be regularly mirroring Wikidata LOTUS through the SSOT as described in Figure 1.

User Interaction with LOTUS Data

The possibilities to interact with the LOTUS data are numerous. The following gives examples of how to retrieve, add and edit LOTUS data.

Data Retrieval

LOTUS data can be queried and retrieved either directly in Wikidata or on LNPN, both of which have distinct advantages. While Wikidata offers modularity at the cost of potentially complex access to the data, LNPN has a graphical user interface with capabilities of drawing chemical structure, simplified structural or biological filtering and advanced chemical descriptors, albeit with a more rigid structure. For bulk download, a frozen version of LOTUS data (timestamp of 2021-05-23) is also available at https://osf.io/eydjs/. More refined approaches to the direct interrogation of the up-to-date LOTUS data both in Wikidata and LNPN are detailed in the following.

Wikidata

The easiest way to search for NP occurrence information in Wikidata is by typing the name of a chemical structure directly into the “Search Wikidata” field on the upper right of the Wikidata homepage. For example, by typing “erysodine”, the user will land on the page of this compound (Q27265641). Scrolling down to the “found in taxon” statement will allow the user to view the biological organisms reported to contain this NP (Figure 3). Clicking the reference link under each taxon name links to the publication(s) documenting the occurrence.

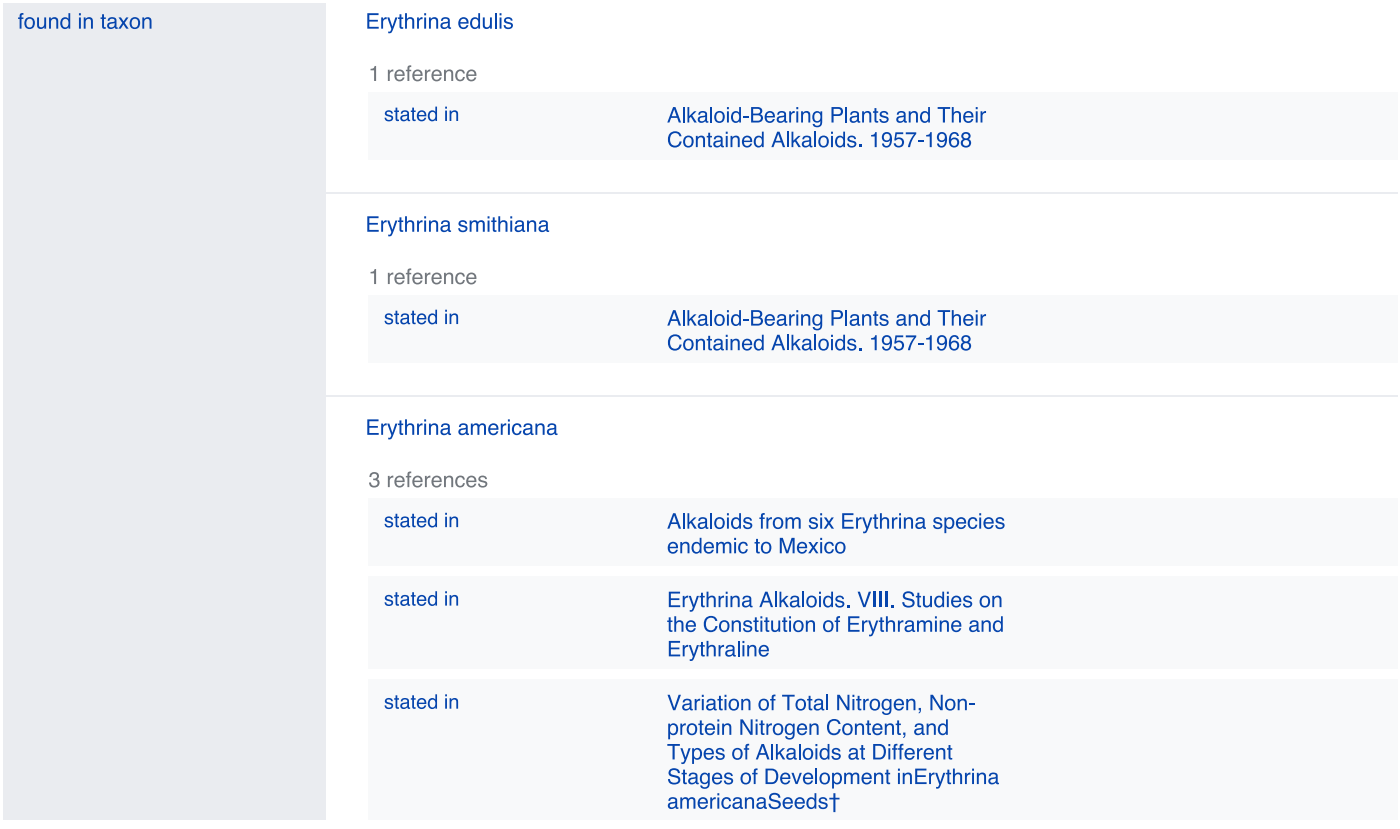


Figure 3: Illustration of the “found in taxon” statement section on the Wikidata page of erysodine Q27265641 showing a selection of containing taxa and the references documenting these occurrences.

The typical approach to more elaborated queries consists in writing SPARQL queries using the Wikidata Query Service or a direct connection to a SPARQL endpoint. Below are some examples from simple to more elaborated queries, demonstrating what can be done using this approach. The full-text queries with explanations are included in SI-3.

Table 2: Potential questions about referenced structure-organism relationships and the corresponding Wikidata SPARQL query that provides an answer.

Question	Wikidata SPARQL query
What are the compounds present in Mouse-ear cress (*Arabidopsis thaliana*)?	https://w.wiki/3HMX
Which organisms are known to contain β-sitosterol?	https://w.wiki/3HLy
Which organisms are known to contain stereoisomers of β-sitosterol?	https://w.wiki/3Jgs
Which pigments are found in which taxa, according to which reference?	https://w.wiki/3H3o
What are examples of organisms where compounds were found in an organism sharing the same parent taxon, but not the organism itself?	https://w.wiki/3HM6
Which *Zephyranthes* species lack compounds known from at least two species in the genus?	https://w.wiki/3HJf
How many compounds are structurally similar to compounds labeled as antibiotics? Results are grouped by the parent taxon of the organism they were found in.	https://w.wiki/3HMA
Which are the available referenced structure-organism pairs? (example limited to 1000 results)	https://w.wiki/3JpE
Which organisms contain indolic scaffolds? Count occurrences, group and order the results by the parent taxon.	https://w.wiki/3HMD
How many structure-organism pairs have been referenced by certain authors? (Here, two senior natural products chemists and co-authors of this paper are compared to the late Ferdinand Bohlmann).	https://w.wiki/3HML

The queries presented in Table 2 are only a few examples and many other ways of interrogating LOTUS can be formulated. Generic queries can be used, for example, for hypothesis generation when starting a research project. For instance, a generic SPARQL query - listed in Table 2 as "Which are the available referenced structure-organism pairs?" - retrieves all structures, identified by their InChIKey (P235), which contain "found in taxon" (P703) statements that are stated in (P248) a bibliographic reference: https://w.wiki/3JpE. Data can then be exported in various formats, such as classical tabular formats, json, or html tables (see Download tab on the lower right of the query frame). At the time of writing (2021-05-05), this query (without the LIMIT 1000) returned 797,123 entries; a frozen query result is available at https://osf.io/thqaw/.

Targeted queries allowing to interrogate LOTUS data from the perspective of one of the three objects forming the referenced structure-organism pairs can be also built. Users can, for example, retrieve a list of all structures reported from a given organism, such as all structures reported from *Arabidopsis thaliana* (Q158695) (https://w.wiki/3HLn). Alternatively, all organisms containing a given chemical can be queried via its structure, such as in the search for all organisms where β-sitosterol (Q121802) was found in (https://w.wiki/3HLy). For programmatic access, the lotus-wikidata-exporter repository also allows data retrieval in RDF format and as TSV tables.

As indicated, certain types of queries that are customary in existing molecular electronic resources, such as structure or similarity searches, are not directly available in Wikidata as SPARQL does not natively support a simple integration of such queries. To address this issue, Galgonek et al. developed an in-house SPARQL engine that allows utilization of Sachem, a high-performance chemical DB cartridge for PostgreSQL for fingerprint-guided substructure and similarity search (Kratochvíl et al., 2018). The engine is used by the Integrated Database of Small Molecules (IDSM) that operates, among other things, several dedicated endpoints allowing structural search in selected small-molecule datasets via SPARQL (Kratochvíl et al., 2019). To allow substructure and similarity searches via SPARQL also on compounds from Wikidata, a dedicated IDSM/Sachem endpoint was created for the LOTUS project. The endpoint indexes isomeric (P2017) and canonical (P233) SMILES code available in Wikidata. To ensure that data is kept up-to-date, SMILES codes are automatically downloaded from Wikidata daily. The endpoint allows users to run federated queries and, thereby, proceed to structure-oriented searches on the LOTUS data hosted at Wikidata. For example, the SPARQL query https://w.wiki/3HMD returns a list of all organisms that produce NP with an indolic scaffold. The output is aggregated at the parent taxa level of the containing organisms and ranked by the number of scaffold occurrences.

Lotus.NaturalProducts.Net (LNPN)

In the search field of the LNPN interface (https://lotus.naturalproducts.net/), simple queries can be achieved by typing the molecule name (e.g., protopine) or pasting a SMILES, InChI, InChIKey string, or a Wikidata identifier. All compounds reported from a given organism can be found by entering the organism name at the species or any higher taxa level (e.g. *Tabernanthe iboga*). Compound search by chemical class is also possible.

Alternatively, a structure can be directly drawn in the structure search interface (https://lotus.naturalproducts.net/search/structure), where the user can also decide on the nature of the structure search (exact, similarity, substructure search). Refined search mode combining multiple search criteria, in particular physicochemical properties, is available in the advanced search interface (https://lotus.naturalproducts.net/search/advanced).

Within LNPN, LOTUS bulk data can be retrieved as SDF or SMILES files, or as a complete MongoDB dump via https://lotus.naturalproducts.net/download. Extensive documentation describing the search possibilities and data entries is available at https://lotus.naturalproducts.net/documentation. LNPN can also be queried via the application programming interface (API) as described in the documentation.

Data Addition and Evolution

One major advantage of the LOTUS architecture is that every user has the option to contribute to the NP occurrences documentation effort by adding new or editing existing data. As all LOTUS data applies the SSOT mechanism, reprocessing of previously treated elements is avoided. However, at the moment, the SSOT channels are not open to the public for direct write access to maintain data coherence and evolution of the SSOT scheme. For now, the users can employ the following approaches to add or modify data in LOTUS.

Sources

LOTUS data management involves regular re-importing of both current and new data sources. New and edited information from these electronic NP resources will be checked against the SSOT. If absent or different, data will be passed through the curation pipeline and subsequently stored in the SSOT. Accordingly, by contributing to external electronic NP resources, any researcher has a means of providing new data for LOTUS, keeping in mind the inevitable delay between data addition and subsequent inclusion into LOTUS.

Wikidata

The currently favored approach to add new data to LOTUS is to edit Wikidata entries directly. Newly edited data will then be imported into the SSOT repository. There are several ways to interact with Wikidata which depend on the technical skills of the user and the volume of data to be imported/modified.

Manual Upload

Any researcher interested in reporting NP occurrences can manually add the data directly in Wikidata, without any particular technical knowledge requirement. The only prerequisite is a Wikidata account and following the general object editing guidelines. Regarding the addition of NP-centered objects (i.e., referenced structure-organisms pairs), users shall refer to the WikiProject Chemistry/Natural products group page.

A tutorial for the manual creation and upload of a referenced structure-organism pair to Wikidata is available in SI-4. While direct Wikidata upload is possible, contributors are encouraged to use the LOTUS curation pipeline as a preliminary step to strengthen the initial data quality. The added data will therefore benefit from the curation and validation stages implemented in the LOTUS processing pipeline.

Batch and Automated Upload

Through the initial curation process described previously, 700,000+ referenced structure-organism pairs were validated for Wikidata upload. To automate this process, a set of programs were written to automatically process the curated outputs, group references, organisms and compounds, check if they are already present in Wikidata (using SPARQL and direct Wikidata querying) and insert or update the entities as needed (i.e., upserting). These scripts can be used for future batch upload of properly curated and referenced structure-organism pairs to Wikidata. Programs for data addition to Wikidata can be found in the repository lotus-wikidata-importer. The following Xtools page offers an overview of the latest activity performed by our NPimporterBot, using those programs.

Data Editing

Even if correct at a given time point, scientific advances can invalidate or update previously uploaded data. Thus, the possibility to continuously edit the data is desirable and guarantees data quality and sustainability. Community-maintained knowledge bases such as Wikidata encourage such a process. Wikidata presents the advantage of allowing both manual and automated correction. Field-specific robots such as SuccuBot, KrBot, Pi bot and ProteinBoxBot or our NPimporterBot went through an approval process. The robots are capable of performing thousands of edits without the need for human input. This automation helps reduce the amount of incorrect data that would otherwise require manual editing. However, manual curation by human experts remains irreplaceable as a standard. Users who value this approach and are interested in contributing are invited to follow the manual curation tutorial in SI-4.

The Scholia platform provides a visual interface to display the links among Wikidata objects such as researchers, topics, species or chemicals. It now provides an interesting way to view the chemical compounds found in a given biological organism (see here for the metabolome view of *Eurycoma longifolia*). If Scholia currently does not offer a direct editing interface for scientific references, it still allows users to proceed to convenient batch editing via Quick Statements. The adaptation of such a framework to edit the referenced structure-pairs in the LOTUS initiative could thus facilitate the capture of future expert curation, especially manual efforts that cannot be replaced by automated scripts.

Data Interpretation

To illustrate the nature and dimensions of the LOTUS dataset, some selected examples of data interpretation are shown. First, the repartition of chemical structures among four important NP reservoirs: plants, fungi, animals and bacteria (Table 3). Then, the distribution of biological organisms according to the number of related chemical structures and likewise the distribution of chemical structures across biological organisms are illustrated (Figure 4). Furthermore, the individual electronic NP resources participation in LOTUS data is resumed using the UpSet plot depiction, which allows the visualization of intersections in data sets (Figure 5). Across these figures we take again the two previous examples, i.e, *β*-sitosterol as chemical structure and *Arabidopsis thaliana* as biological organism because of their well-documented statuses. Finally, a biologically-interpreted chemical tree and a chemically-interpreted biological tree are presented (Figure 6 and Figure 7). The examples illustrate the overall chemical and biological coverage of LOTUS by linking family-specific classes of chemical structures to their taxonomic position. Table 3, Figures 4, 6 and 7 were generated using the frozen data (2021-05-23 timestamp), which is available for download at https://osf.io/eydjs/. Figure 5 required a dataset containing information from DNP and the complete data used for its generation is therefore not available for public distribution. All scripts used for the generation of the figures (including SI-5) are available in the lotus-processor repository in the src/4_visualizing directory for reproducibility.

Repartition of Chemical Structures across reported Biological Organisms in LOTUS

Table 3 summarizes the repartition of chemical structures and their chemical classes (according to NPClassifier (kim et al., 2020)) across the biological organisms reported in LOTUS. For this, biological organisms were grouped into four artificial taxonomic levels (plants, fungi, animals and bacteria). These were built by combining the two highest taxonomic levels in the OTL taxonomy, namely Domain and Kingdom levels. When the chemical structure/class was reported only in one taxonomic grouping, it was counted as “specific”.

Table 3: Repartition and specificity of chemical structures across four important NP reservoirs: plants, fungi, animals and bacteria. When the chemical structure/class appeared only in one group and not the three others, they were counted as “specific”. Chemical classes were attributed with NPClassifier.

Group	Organisms	Structure-Organism Pairs	Chemical Structures	Specific Chemical Structures	Chemical Classes	Specific Chemical Classes
Plantae	26,204	485,018	186,906	181,263 (97%)	510	62 (12%)
Fungi	3,642	54,744	38,695	35,538 (92%)	398	31 (8%)
Animalia	2,524	34,464	24,757	21,084 (85%)	427	16 (4%)
Bacteria	1,424	29,311	23,917	22,124 (93%)	334	38 (11%)

Distributions of Organisms per Structure and Structures per Organism

Readily achievable outcomes from LOTUS show that the depth of exploration of the world of NP is rather limited: as depicted in Figure 4, on average, three organisms are reported per chemical structure and eleven structures per organism. Notably, half of all structures have been reported from a single organism and half of all studied organisms are reported to contain five or fewer structures. Metabolomic studies suggest that these numbers are heavily underrated (Noteborn et al., 2000; Wang et al., 2020) and indicate that a better reporting of the metabolites detected in the course of phytochemical investigations should greatly improve coverage. This incomplete coverage may be partially explained by the habit in classical NP journals to accept only new and/or bioactive chemical structures for publication.

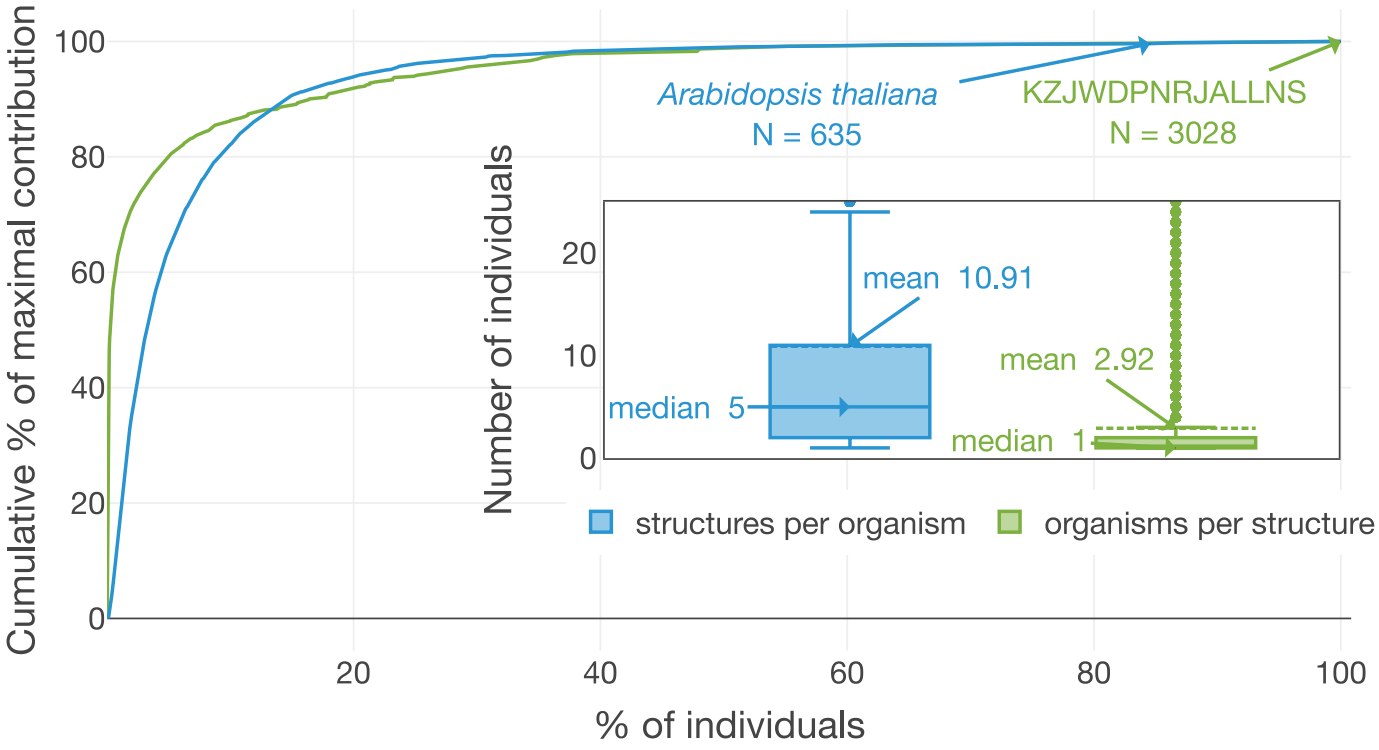


Figure 4: Distribution of "structures per organism" and "organisms per structure". The number of organisms linked to the planar structure of *β*-sitosterol (KZJWDPNRJALLNS) and the number of chemical structures in *Arabidopsis thaliana are* two exemplary highlights. The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_structure_organism_distribution.svg.

Contribution of Individual Electronic NP Resources to LOTUS

The added value of the LOTUS initiative to assemble multiple electronic NP resources is illustrated in Figure 5 : Panel A shows the contributions of the individual electronic NP resources to the ensemble of chemical structures found in one of the most studied vascular plants, *Arabidopsis thaliana* ("Mouse-ear cress"; Q147096). Panel B shows the ensemble of taxa reported to contain the planar structure of the widely occuring triterpenoid *β*-sitosterol (Q121802).

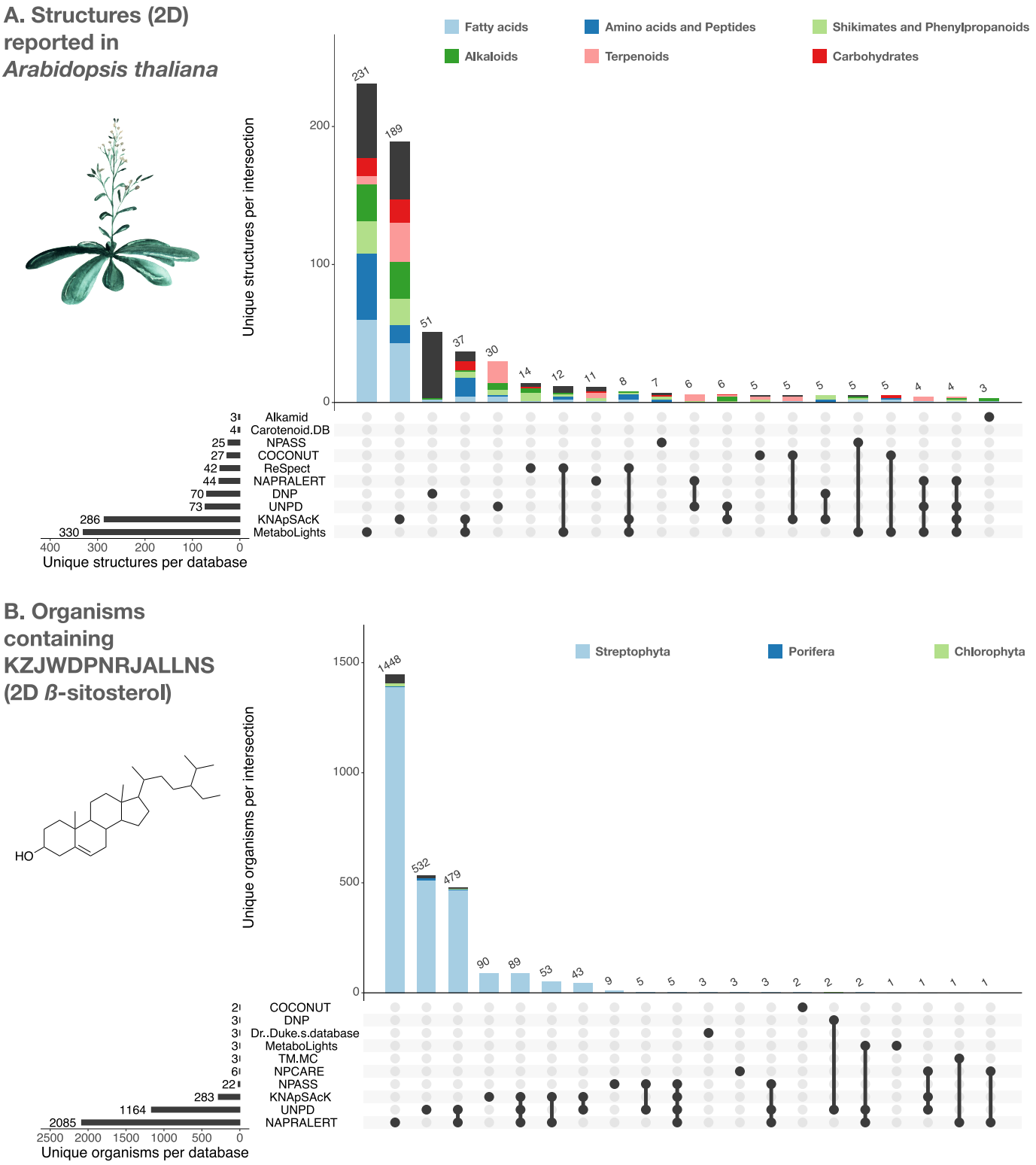


Figure 5: UpSet plots of the individual contribution of electronic NP resources to the planar structures found in *Arabidopsis thaliana* (A) and to organisms reported to contain the planar structure of *β*-sitosterol (KZJWDPNRJALNS) (B). UpSet plots are evolved Venn diagrams, allowing to represent intersections between multiple sets. The horizontal bars on the lower left represent the number of corresponding entries per electronic NP resource. The dots and their connecting line represent the intersection between source and consolidate sets. The vertical bars indicate the number of entries at the intersection. For example, 479 organisms containing the planar structure of *β*-sitosterol are present in both UNPD and NAPRALERT, whereas each of them respectively reports 1,164 and 2,085 organisms containing the planar structure of *β*-sitosterol. The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_upset_plot.svg.

Figure 5 A. also shows that according to NPClassifier, the chemical pathway distribution across electronic NP resources is unconserved. Note that NPClassifier and ClassyFire (Djoumbou Feunang et al., 2016) chemical taxonomies are both available as metadata in the frozen LOTUS export and LNPN. Both classification tools return a chemical taxonomy for individual structures, thus allowing their grouping at higher hierarchical levels, in the same way as it is done for biological taxonomies. The UpSet plot in Figure 5 indicates the poor overlap of preexisting electronic NP resources and the added value of an aggregated dataset. This is particularly well illustrated in Figure 5 B., where the number of organisms for which the planar structure of *β*-sitosterol (KZJWDPNRJALLNS) has been reported is shown for each intersection. NAPRALERT has by far the highest number of entries (2,085 in total), while other electronic NP resources complement this well: e.g., UNPD has 532 reported organisms with *β*-sitosterol that do not overlap with those reported in NAPRALERT. Of note, *β*-sitosterol is documented in only 3 organisms in the DNP, highlighting the importance of a better systematic reporting of ubiquitous metabolites and the interest of multiple data sources agglomeration.

A Biologically-interpreted Chemical Tree

The chemical diversity captured in LOTUS is here displayed using tmap (Figure 6), a visualization library allowing the structural organization of large chemical datasets as a minimum spanning tree (Probst and Reymond, 2020). Using Faerun, an interactive HTML file is generated to display metadata and molecule structures by embedding the SmilesDrawer library (Probst and Reymond, 2018a, 2018b). Planar structures were used for all compounds to generate the TMAP (chemical space tree-map) using MAP4 encoding (Capecchi et al., 2020). As the tree organizes structures according to their molecular fingerprint, an anticipated coherence between the clustering of compounds and the mapped NPClassifier chemical class is observed (Figure 6 A.). For clarity, the eight most represented chemical classes of LOTUS plus the quassinoids and carotenoids (C40, *β-β*) classes are mapped, with examples of a quassinoid (green star) and a carotenoid (yellow star) and their corresponding location in the TMAP.

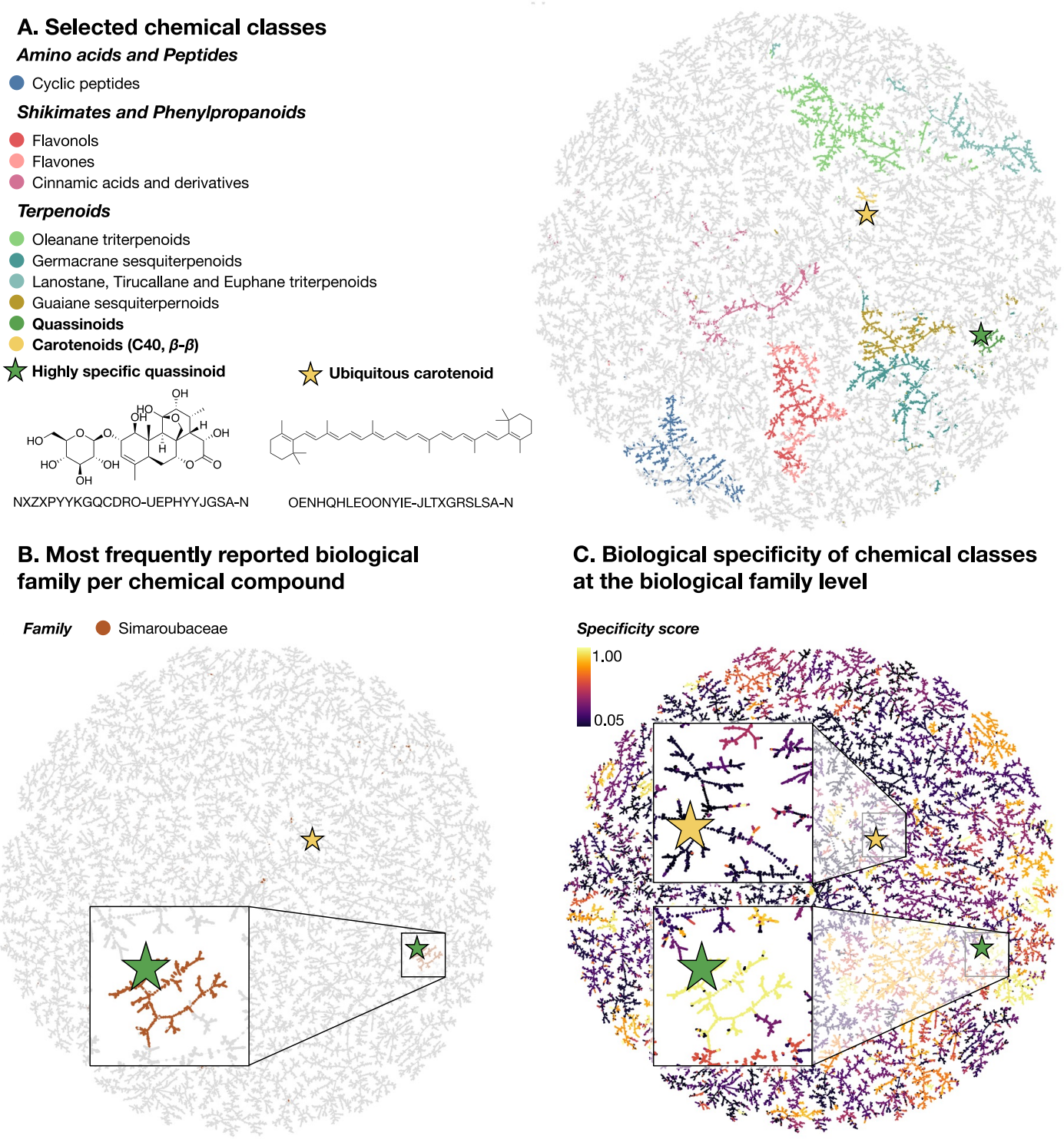


Figure 6: TMAP visualizations of the chemical diversity present in LOTUS. Each dot corresponds to a chemical structure. A highly specific quassinoid (green star) and an ubiquitous carotenoid (yellow star) are mapped as examples in all visualizations. In panel A., compounds (dots) are colored according to the NPClassifier chemical class they belong to. In panel B., compounds which are mostly reported in the Simaroubaceae family are highlighted in red. Finally, in panel C., the compounds are colored according to the specificity score of chemical classes found in biological organisms. This biological specificity score at a given taxonomic level for a given chemical class is calculated as the number of structure-organism pairs within the taxon where the chemical class occurs the most, divided by the total number of pairs in the chemical class. A chemical class biological specificity score of 1 means that compounds in that chemical class were reported in a unique biological family. Zooms on a group of compounds of high biological specificity score (in yellow) and on compounds of low specificity (black) are depicted. An interactive HTML visualization of the LOTUS TMAP is available at https://osf.io/kqa8b/. The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_biologically_interpreted_chemical_tree.svg.

To explore relationships between chemistry and biology, it is possible to map taxonomical information such as the most reported biological family per chemical compound (Figure 6 B.) or the biological specificity of chemical classes (Figure 6 C.) on the TMAP. The biological specificity score at a given

taxonomic level for a given chemical class is calculated as the number of structure-organism pairs within the taxon where the chemical class occurs the most, divided by the total number of pairs. See Equation 1:

$$\text{Specificity score}_{\text{bio}} = \frac{|\text{Pairs in chemical class} \cap \text{Pairs in taxon where chemical class occurs the most}|}{|\text{Pairs in chemical class}|} \quad (1)$$

This visualization allows to highlight chemical classes specific to a given taxon, such as the quassinoids in the Simaroubaceae family. In this case it is striking to see how well the compounds of a given chemical class (quassinoids) (Figure 6 A.) and the most reported plant family per compound (Simaroubaceae) (Figure 6 B.) overlap. This is also evidenced on Figure 6 C. with a chemical class specificity of 0.95 at the biological family level for quassinoids. In this plot, it is also possible to identify chemical classes that are widely spread among living organisms, such as the carotenoids (C40, *β-β*), which exhibit a specificity of 0.12 at the biological family level. This means that among all the carotenoids (C40, *β-β*) - organism pairs, about one tenth belong to the most common family.

A Chemically-interpreted Biological Tree

An alternative view of the biological and chemical diversity covered by LOTUS is illustrated in Figure 7. Here chemical compounds are not organized but biological organisms are placed in their taxonomy. To limit bias due to underreporting in the literature and keep a reasonable display size, only families with at least 50 reported compounds were included. Organisms were classified according to the OTL taxonomy and structures according to NPClassifier. The tips were labeled according to the biological family and colored according to their biological kingdom. The bars represent structure specificity of the most characteristic chemical class of the given biological family (the higher the more specific). This specificity score was calculated as in Equation 2:

$$\text{Specificity score}_{\text{chem}} = \frac{|\text{Structures in chemical class} \cap \text{Structures in taxon}|^2}{|\text{Structures in chemical class}| \cdot |\text{Structures in taxon}|} \quad (2)$$

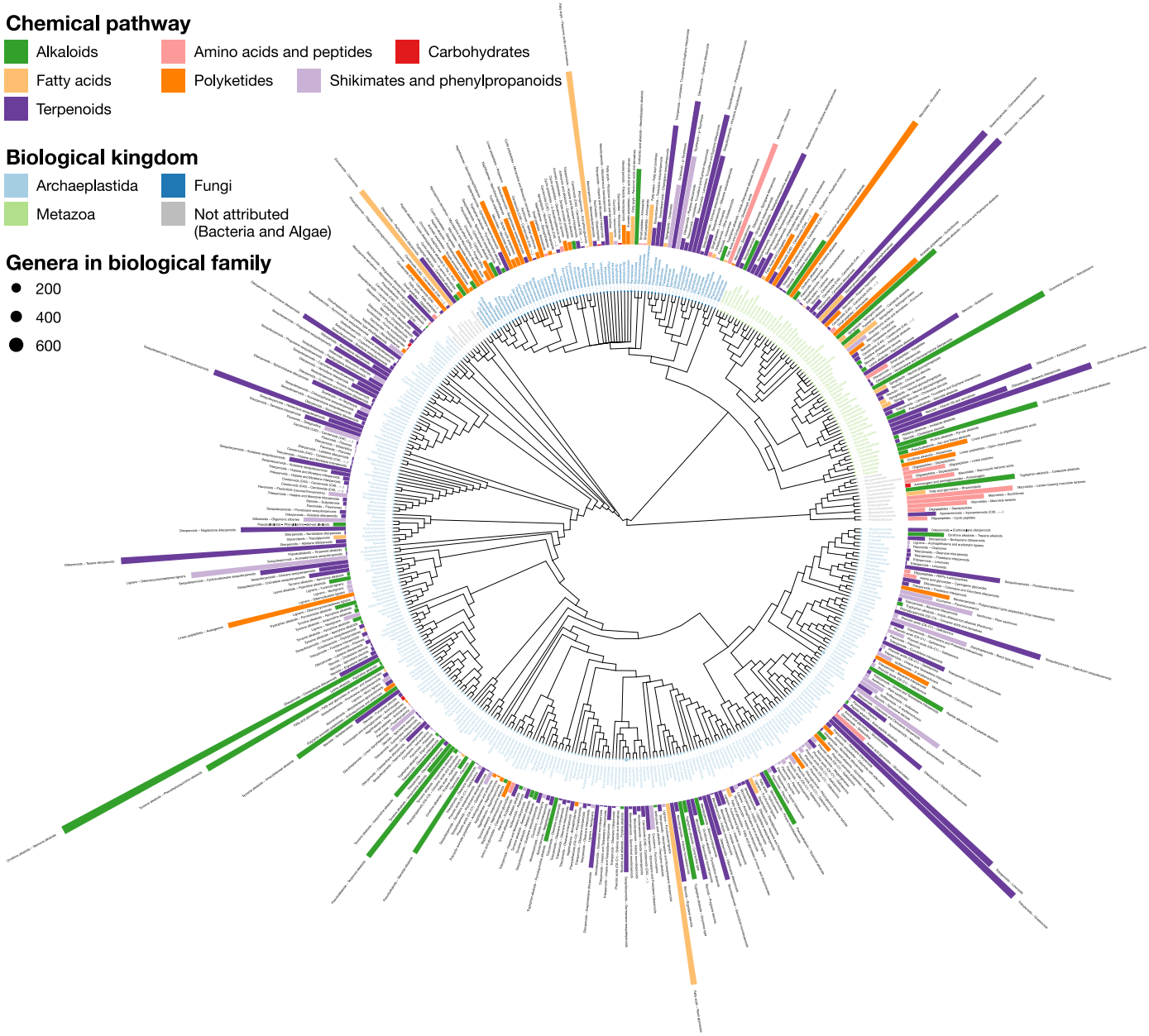


Figure 7: LOTUS provides new means of exploring and representing chemical and biological diversity. The tree generated from current LOTUS data builds on biological taxonomy and employs the kingdom as tips label color (only families containing 50+ chemical structures were considered). The outer bars correspond to the most specific chemical class found in the biological family. The height of the bar is proportional to a specificity score corresponding to the square of the number of structures reported in the chemical class within the given biological family over the product of the number of reported structures in the chemical class with the number of reported structures in the biological family. The bar color corresponds to the chemical pathway of the most specific chemical class in the NPClassifier classification system. The size of the leaf nodes corresponds to the number of genera reported in the family. The figure is vectorized and zoomable for detailed inspection and is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_chemically_interpreted_biological_tree.svg.

Figure 7 makes it possible to spot highly specific compound classes such as trinervitane terpenoids in the Termitidae, the rhizoxin macrolides in the Rhizopodaceae, or the quassinoids and limonoids typical, respectively, of Simaroubaceae and Meliaceae. Similarly, tendencies of more generic occurrence of NP can be observed. For example, within the fungal kingdom, Basidiomycotina appear to have a higher biosynthetic specificity toward terpenoids than other fungi, which mostly focus on polyketides production. When observed at a finer scale, down to the structure level, such chemotaxonomic representation can give valuable insights. For example, among all chemical structures, only two were found in all biological kingdoms, namely heptadecanoic acid (KEMQGTRYUADPNZ-UHFFFAOYSA-N) and *β*-carotene (OENHQHLEOONYIE-JLTXGRSLSA-N). Looking at the repartition of *β*-sitosterol (KZJWDPNRJALLNS-VJSFXXLFSA-N) within the overall biological tree, SI-5 plots its presence/absence *versus* those of its superior chemical classifications, namely the stigmastane, steroid and terpenoid derivatives, over the same tree used in Figure 7. The comparison of these five chemically-interpreted biological trees clearly highlights the increasing speciation of the *β*-sitosterol biosynthetic pathway in the Archaeplastida kingdom, while the superior classes are distributed across all kingdoms. Figure 7 is zoomable and vectorized for detailed inspection.

As illustrated, the possibility of data interrogation at multiple precision levels, from fully defined chemical structures to broader chemical classes, is of great interest, e.g., for taxonomic and evolution studies. This makes LOTUS a unique ressource for the advancement of chemotaxonomy, a discipline pioneered by Augustin Pyramus de Candolle and pursued by other notable researchers (Robert Hegnauer, Otto R. Gottlieb) (de Candolle, 1816; Gottlieb, 1982; Hegnauer, 1986a). Six decades after Hegnauer's publication of "Die Chemotaxonomie der Pflanzen" (Hegnauer, 1986b) much remains to be done for the advancement of this field of study and the LOTUS initiative aims to provide a solid basis for researchers willing to pursue these exciting explorations at the interface of chemistry, biology and evolution.

As shown recently in the context of spectral annotation (Dührkop et al., 2020), lowering the precision level of the annotation allows a broader coverage along with greater confidence. Genetic studies investigating the pathways involved and the organisms carrying the responsible biosynthetic genes would be of interest to confirm the previous observations. These forms of data interpretation exemplify the importance of reporting not only new structures, but also novel occurrences of known structures in organisms as comprehensive chemotaxonomic studies are pivotal for a better understanding of the metabolomes of living organisms.

The integration of multiple knowledge sources, e.g. genetics for NP producing gene clusters (Kautsar et al., 2019) combined to taxonomies and occurrences DB, also opens new opportunities to understand if an organism is responsible for the *biosynthesis* of a NP or merely *contains* it. This understanding is of utmost importance for the chemotaxonomic field and will help to understand to which extent microorganisms (endosymbionts) play a role in host development and its NP expression potential (SAIKKONEN, 2004).

Conclusion & Perspectives

Advancing Natural Products Knowledge

At its current development stage, data harmonized and curated throughout the LOTUS initiative remain imperfect and, by the very nature of research, at least partially biased (see Introduction). In the context of bioactive NP research, and due to global editorial practices, it should not be ignored that many publications tend to emphasize new compounds and/or those for which interesting bioactivity has been measured. Near-ubiquitous (primarily plant-based) compounds tend to be overrepresented in the NP literature, yet the implication of their wide distribution in nature and associated patterns of broad, non-specific activity are often underappreciated (Bisson et al., 2015). Ideally, all characterized compounds independent of structural novelty and/or bioactivity profile should be documented, and the expansion of verified structure-organism pairs is fundamental to the advancement of NP research.

The LOTUS initiative provides a framework for rigorous review and incorporation of new records and already presents a valuable overview of the distribution of NP occurrences studied to date. While current data presents a reasonable approximation of the chemistries of a few well-studied organisms such as *Arabidopsis thaliana*, they remain patchy for many other organisms represented in the dataset. Community participation is the most efficient means of achieving more comprehensive documentation of NP occurrences, and the comprehensive editing opportunities provided within LOTUS and through the associated Wikidata distribution platform open new opportunities for collaborative engagement. In addition to facilitating the introduction of new data, it also provides a forum for critical review of existing data, as well as harmonization and verification of existing NP datasets as they come online.

Fostering FAIRness and TRUSTworthiness

The LOTUS harmonized data and dissemination of referenced structure-organism pairs through Wikidata, enables novel forms of queries and transformational perspectives in NP research. As LOTUS follows the guidelines of FAIRness and TRUSTworthiness, all researchers across disciplines can benefit from this opportunity, whether the interest is in ecology and evolution, chemical ecology, drug discovery, biosynthesis pathway elucidation, chemotaxonomy, or other research fields that connect with NP.

The introduction of LOTUS even provides a new opportunity to advance the FAIR guiding principles for scientific data management and stewardship originally established in 2016 (Wilkinson et al., 2016). Researchers worldwide uniformly acknowledge the limitations caused by the intrinsic unavailability of essential (raw) data (Bisson et al., 2016). The lack of progress is, at least in part, due to elements in the dissemination channels of the classical print and static PDF publication formats that complicate or sometimes even discourage data sharing, e.g., due to page limitations and economically motivated mechanisms, including those involved in the focus on and calculation of journal impact factors. In particular raw data such as experimental readings, spectroscopic data, instrumental measurements, statistical, and other calculations are valued by all, but disseminated by only very few. The immense value of raw data and the desire to advance the public dissemination has recently been documented in detail for nuclear magnetic resonance (NMR) spectroscopic data by a large consortium of NP researchers (McAlpine et al., 2019). However, to generate the vital flow of contributed data, the effort associated with preparing and submitting content to open repositories as well as data reuse should be better acknowledged in academia, government, regulatory, and industrial environments (Cousijn et al., 2019, 2018; Pierce et al., 2019).

Opening New Perspectives for Spectral Data

The possibilities for expansion and future applications of the Wikidata-based LOTUS initiative are significant. For example, properly formatted spectral data, e.g., data obtained by MS or NMR, can be linked to the Wikidata entries for the respective chemical compounds. MassBank (Horai et al., 2010) and SPLASH (Wohlgemuth et al., 2010) identifiers are already reported in Wikidata, and this existing information can be used to report MassBank or SPLASH records for example for *Arabidopsis thaliana* compounds (https://w.wiki/3PJD). Such possibilities will help to bridge experimental data results obtained during the early stages of NP research with data that has been reported and formatted in different contexts. This opens exciting perspectives for structural dereplication, NP annotation, and metabolomic analysis. The authors have previously demonstrated that taxonomically-informed metabolite annotation is critical for the improvement of the NP annotation process (Rutz et al., 2019). Alternative approaches linking structural annotation to biological organisms have also shown substantial improvements (Hoffmann et al., 2021). The LOTUS initiative offers new opportunities for linking chemical objects to both their biological occurrences and spectral information and should significantly facilitate such applications.

Integrating Chemodiversity, Biodiversity, and Human Health

As shown in SI-5, observing the chemical and biological diversity at various granularities can offer new insights. Regarding the chemical objects involved, it will be important to document the taxonomies of chemical annotations for the Wikidata entries. However, this is a rather complex task, for which stability and coverage issues will have to be addressed first. Existing chemical taxonomies such as ChEBI, ClassyFire, or NPClassifier are evolving steadily, and it will be important to constantly update the tools used to make further annotations. Repositioning NP within their greater biosynthetic context is another major challenge - and active field of research. The fact that the LOTUS disseminates data through Wikidata will help facilitate its integration with biological pathway knowledge bases such as WikiPathways and contribute to this complex task (Martens et al., 2021; Slenter et al., 2018).

In the field of ecology, for example, molecular traits are gaining increased attention (Kessler and Kalske, 2018; Sedio, 2017). Conceptually, LOTUS can help associate classical plant traits (e.g., leaf surface area, photosynthetic capacities, etc.) with Wikidata biological organisms entries, and, thus, allow their integration and comparison with chemicals that are associated with the organisms. Likewise, the association of biogeography data documented in repositories such as GBIF could be further exploited in Wikidata to pursue the exciting but understudied topic of "chemodiversity hotspots" (Defossez et al., 2021).

Further NP-related information of great interest remains poorly formatted. One example of such a shortcoming relates to traditional medicine, including ethnomedicine and ethnobotany, which is the historical and empiric approach of mankind to discover and use bioactive products from Nature, primarily plants. The amount of knowledge generated in human history on the use of medicinal substances represents fascinating yet underutilized information. Notably, the body of literature on the pharmacology and toxicology of NP is compound-centric, increases steadily, and relatively scattered, but still highly relevant (not necessarily: sufficient) for exploring the role and potential utility of NP for human health. To this end, the LOTUS initiative represents a resource for new concepts by which such information could be valued and conserved in the digital era, as LOTUS provides a blueprint for appropriate formatting and sharing of such data (Allard et al., 2018; Geoffrey A. Cordell, 2017a, 2017b). This underscores the transformative value of the LOTUS initiative for the advancement of Traditional Medicine and its drug discovery potential in health systems worldwide.

Summary & Outlook

The various facets discussed above connect with ongoing and future developments that the tandem of the LOTUS initiative and its Wikidata integration can accommodate through a broader knowledge base. The information of the LOTUS initiative is already readily accessible by third party projects build on top of Wikidata such as the SLING project (https://github.com/ringgaard/sling, see entry for gliotoxin) or the Plant Humanities Lab project (https://lab.plant-humanities.org/, see entry for *Ilex guayusa*).

Behind the scenes, all underlying resources represent data in a multidimensional space and can be extracted as individual graphs that can be interconnected. The craft of appropriate federated queries allows users to navigate these graphs and fully exploit their potential (Kratochvíl et al., 2018; Waagmeester et al., 2020). The development of interfaces such as RDFFrames (Mohamed et al., 2020) will also facilitate the use of the wide arsenal of existing machine learning approaches to automate reasoning on these knowledge graphs.

Overall, the LOTUS initiative aims to make more and better data available. This project paves the way for the establishment of an open and expandable electronic NP resource. The design and efforts of the LOTUS initiative reflect our conviction that the integration of NP research results is long-needed and requires a truly open and FAIR knowledge base. We believe that the LOTUS initiative has the potential to fuel a virtuous cycle of research habits and, as a result, *contribute to a better understanding of Life and its chemistry*.

Methods

Data Curation

Gathering

Before their inclusion, the overall quality of the source was manually assessed to estimate the quality of referenced structure-organism pairs and the lack of ambiguities in the links between data and references. This led to the identification of thirty-six electronic NP resources as valuable LOTUS input. Data from the proprietary Dictionary of Natural Products (DNP v 29.2) was also used for comparison purposes only and is not publicly disseminated. FooDB was also curated but not publicly disseminated since its license did not allow sharing in Wikidata. SI-1 gives all necessary details regarding electronic NP resources access and characteristics.

Manual inspection of each electronic NP resource revealed that the structure, organism, and reference fields were widely variable in format and contents, thus requiring standardization to be comparable. The initial stage consisted of writing tailored scripts that are capable of harmonizing and categorizing knowledge from each source (Figure 1). This transformative process led to three categories: fields relevant to the chemical structure described, to the producing biological organism, and the reference describing the occurrence of the chemical structure in the producing biological organism. This process resulted in categorized columns for each source, providing an initial harmonized format for each table.

For all thirty-eight sources, if a single file or multiple files were accessible *via* a download option including FTP, data was gathered that way. For some sources, data was scraped (cf. SI-1). All scraping scripts can be found in the lotus-processor repository in the src/1_gathering directory (under each respective subdirectory). Data extraction scripts for the DNP are available and should allow users with a DNP license only to further exploit the data (src/1_gathering/db/dnp). The chemical structure fields, organism fields, and reference fields were manually categorized into three, two, and ten subcategories, respectively. For chemical structures, “InChI”, “SMILES”, and “chemical name” (not necessarily IUPAC). For organisms, “clean” and “dirty”, meaning lot text not referred to the canonical name was present or the organism was not described by its canonical name (e.g. “Compound isolated from the fresh leaves of *Citrus* spp.”). For the references, the original reference was kept in the “original” field. When the format allowed it, references were divided into: “authors”, “doi”, “external”, “isbn”, “journal”, “original”, “publishing details”, “pubmed”, “title”, “split”. The generic “external” field was used for all external cross-references to other websites or electronic NP resources (e.g. “also in knapsack”). The last subcategory, “split”, corresponds to a still non-atomic field after the removal of parts of the original reference. Other field titles are self-explanatory. The producing organism field was kept as a single field.

Harmonization

To perform the harmonization of all previously gathered sources, sixteen columns were chosen as described above. Upon electronic NP resources harmonization, resulting subcategories were divided and subject to further cleaning. The “chemical structure” fields were divided into files according to their subcategories (“InChI”, “names” and “SMILES”). A file containing all initial structures from all three subcategories was also generated. The same procedure was followed for organisms and references.

Cleaning

To obtain an unambiguously referenced structure-organism pair for Wikidata dissemination, the initial sixteen columns were translated and cleaned into three fields: the reported structure, the organism canonical name, and the reference. The structure was reported as InChI, together with its SMILES and InChIKey translation. The biological organism field was reported as three minimal necessary and sufficient fields, namely its canonical name and the taxonID and taxonomic DB corresponding to the latter. The reference was reported as four minimal fields, namely reference title, DOI, PMCID, and PMID, one being sufficient. For the forthcoming translation processes, automated solutions were used when available. However, for specific cases (common or vernacular names of the biological organisms, Traditional Chinese Medicine (TCM) names, and conversion between digital reference identifiers), no solution existed, thus requiring the use of tailored dictionaries. The initial entries (containing one or multiple producing organisms per structure, with one or multiple accepted names per organism) were cleaned into 2M+ referenced structure-organism pairs.

Chemical Structures

To retrieve as much information as possible from the original structure field(s) of each of the sources, the following procedure was followed. Allowed structural fields for the sources were divided into two types: structural (InChI, SMILES) or nominal (chemical name, not necessarily IUPAC). If multiple fields were present, structural identifiers were preferred over structure names. Among structural identifiers, when both identifiers led to different structures, InChI was preferred over SMILES. SMILES were translated to InChI using the RDKit (2021.03.1) implementation in Python 3.8 (src/2_curating/2_editing/structure/1_translating/smiles.py). They were first converted to ROMol objects which were then converted to InChI. When no structural identifier was available, the nominal identifier was translated to InChI first thanks to OPSIN (Lowe et al., 2011), a fast Java-based translation open-source solution. If no translation was obtained, chemical names were then submitted to the CTS (Wohlgemuth et al., 2010), once in lower case only, once with the first letter capitalized. If again no translation was obtained, candidates were then submitted to the Chemical Identifier Resolver *via* the *cts_convert* function from the webchem package (Szöcs et al., 2020). Before the translation process, some typical chemical structure-related greek characters (such as *α*, *ß*) were replaced by their textual equivalents (alpha, beta) to obtain better results. All pre-translation steps are included in the preparing_name function and are available in src/r/preparing_name.R.

The chemical sanitization step sought to standardize the representation of chemical structures coming from different sources. It consisted of three main stages (standardizing, fragment removal, and uncharging) achieved *via* the MolVS package. The initial standardizer function consists of six stages (RDKit Sanitization, RDKit Hs removal, Metals Disconnection, Normalization, Acids Reionization, and Stereochemistry recalculation) detailed in the molvs documentation. In a second step, the FragmentRemover functionality was applied using a list of SMARTS to detect and remove common counterions and crystallization reagents sometimes occurring in the input DB. Finally, the Uncharger function was employed to neutralize molecules when appropriate.

MarvinSuite was used for traditional and IUPAC names translation, Marvin 20.19, ChemAxon. When stereochemistry was not fully defined, (+) and (-) symbols were removed from names. All details are available in the following script: src/2_curating/2_editing/structure/4_enriching/naming.R. Chemical classification of all resulting structures was done using classyfireR (Djoumbou Feunang et al., 2016) and NPClassifier API.

After manual evaluation, structures remaining as dimers were discarded (all structures containing a “.” in their SMILES were removed).

From the 283,267 initial InChI, 242,068 (85%) sanitized structures were obtained, of which 185,929 (77%) had complete stereochemistry defined. 203,718 (72%) were uploaded to Wikidata. From the 248,185 initial SMILES, 207,658 (84%) sanitized structures were obtained, of which 98,685 (48%) had complete stereochemistry defined. 174,091 (70%) were uploaded to Wikidata. From the 49,675 initial chemical names, 27,932 (56%) sanitized structures were obtained, of which 17,460 (63%) had complete stereochemistry defined. 23,036 (46%) were uploaded to Wikidata. In total, 163,800 structures with fully defined stereochemistry were uploaded as “chemical compounds” (Q11173), and 106,669 structures without fully defined stereochemistry were uploaded as “group of stereoisomers” (Q59199015).

Biological Organisms

The cleaning process at the biological organism's level had three objectives: convert the original organism string to (a) taxon name(s), atomize fields containing multiple taxon names, and deduplicate synonyms. The original organism strings were treated with Global Names Finder (GNF) and Global Names Verifier (GNV), both tools coming from the Global Names Architecture (GNA) a system of web services that helps people to register, find, index, check and organize biological scientific names and interconnect on-line information about species. GNF allows scientific name recognition within raw text blocks and searches for found scientific names among public taxonomic DB. GNV takes names or lists of names and verifies them against various biodiversity data sources. Canonical names, their taxonID, and the taxonomic DB they were found in were retrieved. When a single entry led to multiple canonical names (accepted synonyms), all of them were kept. Because both GNF and GNV recognize scientific names and not common ones, common names were translated before a second resubmission.

Dictionaries

To perform the translations from common biological organism name to latin scientific name, specialized dictionaries included in DrDuke, FooDB, PhenolExplorer were aggregated together with the translation dictionary of GBIF Backbone Taxonomy. The script used for this was src/1_gathering/translation/common.R. When the canonical translation of a common name contained a specific epithet that was not initially present, the translation pair was discarded (for example, "Aloe" translated in "*Aloe vera*" was discarded). Common names corresponding to a generic name were also discarded (for example "Kiwi" corresponding to the synonym of an *Apteryx* spp. (https://www.gbif.org/species/4849989)). When multiple translations were given for a single common name, the following procedure was followed: the canonical name was split into species name, genus name, and possible subnames. For each common name, genus names and species names were counted. If both the species and genus names were consistent at more than 50%, they were considered consistent overall and, therefore, kept (for example, "Aberrant Bush Warbler" had "*Horornis flavolivaceus*" and "*Horornis flavolivaceus intricatus*" as translation; as both the generic ("*Horornis*") and the specific ("*flavolivaceus*") epithets were consistent at 100%, both ("*Horornis flavolivaceus*") were kept). When only the generic epithet had more than 50% consistency, it was kept (for example, "Angelshark" had "*Squatina australis*" and "*Squatina squatina*" as translation, so only "*Squatina*" was kept). Some unspecific common names were removed (see https://osf.io/gqhcn/) and only common names with more than three characters were kept. This resulted in 181,891 translation pairs further used for the conversion from common names to scientific names. For TCM names, translation dictionaries from TCMID, TMMC, and coming from the Chinese Medicine Board of Australia were aggregated. The script used for this was src/1_gathering/translation/tcm.R. Some unspecific common names were removed (see https://osf.io/zs7ky/). Careful attention was given to the Latin genitive translations and custom dictionaries were written (see https://osf.io/c3ja4/, https://osf.io/u75e9/). Organ names of the producing organism were removed to avoid wrong translation (see https://osf.io/94fa2/). This resulted in 7,070 translation pairs. Both common and TCM translation pairs were then ordered by decreasing string length, first translating the longer names to avoid part of them being translated incorrectly.

Translation

To ensure compatibility between obtained taxonID with Wikidata, the taxonomic DB 3 (ITIS), 4 (NCBI), 5 (Index Fungorum), 6 (GRIN Taxonomy for Plants), 8 (The Interim Register of Marine and Nonmarine Genera), 9 (World Register of Marine Species), 11 (GBIF Backbone Taxonomy), 12 (Encyclopedia of Life), 118 (AmphibiaWeb), 128 (ARKive), 132 (ZooBank), 147 (Database of Vascular Plants of Canada (VASCAN)), 148 (Phasmida Species File), 150 (USDA NRCS PLANTS Database), 155 (FishBase), 158 (EUNIS), 163 (IUCN Red List of Threatened Species), 164 (BioLib.cz), 165 (Tropicos - Missouri Botanical Garden), 167 (The International Plant Names Index), 169 (uBio NameBank), 174 (The Mammal Species of The World), 175 (BirdLife International), 179 (Open Tree of Life), 180 (iNaturalist) and 187 (The eBird/Clements Checklist of Birds of the World) were chosen. All other available taxonomic DB are listed at http://index.globalnames.org/datasource. To retrieve as much information as possible from the original organism field of each of the sources, the following procedure was followed: First, a scientific name recognition step, allowing us to retrieve canonical names was carried (src/2_curating/2_editing/organisms/subscripts/1_cleaningOriginal.R). Then, a subtraction step of the obtained canonical names from the original field was applied, to avoid unwanted translation of parts of canonical names. For example, Bromus mango contains "mango" as a specific epithet, which is also the common name for Mangifera indica. After this subtraction step, the remaining names were translated from vernacular (common) and TCM names to scientific names, with help of the dictionaries. For performance reasons, this cleaning step was written in Kotlin and used coroutines to allow efficient parallelization of that process (src/2_curating/2_editing/organisms/2_translating_organism_kotlin/). They were subsequently submitted again to scientific name recognition (src/2_curating/2_editing/organisms/3_cleaningTranslated.R).

After full resolution of canonical names, all obtained names were submitted to rotl (Michonneau et al., 2016) to obtain a unified taxonomy. From the 88,395 initial "clean" organism fields, 43,936 (50%) canonical names were obtained, of which 32,285 (37%) were uploaded to Wikidata. From the 300 initial "dirty" organism fields, 250 (83%) canonical names were obtained, of which 208 (69%) were uploaded to Wikidata.

References

The Rcrossref package (Chamberlain et al., 2020) interfacing with the Crossref API was used to translate references from their original subcategory ("original", "publishingDetails", "split", "title") to a DOI, the title of its corresponding article, the journal it was published in, its date of publication and the name of the first author. The first twenty candidates were kept and ranked according to the score returned by Crossref, which is a tf-idf score. For DOI and PMID, only a single candidate was kept. All parameters are available in src/functions/reference.R. All DOIs were also translated with this method, to eventually discard any DOI not leading to an object. PMIDs were translated, thanks to the entrez_summary function of the rentrez package (Winter, 2017). Scripts used for all subcategories of references are available in the directory src/2_curating/2_editing/reference/1_translating/. Once all translations were made, results coming from each subcategory were integrated, (src/2_curating/2_editing/reference/2_integrating.R) and the producing organism related to the reference was added for further treatment. Because the crossref score was not informative enough, at least one other metric was chosen to complement it. The first metric was related to the presence of the producing organism's generic name in the title of the returned article. If the title contained the generic name of the organism, a score of 1 was given, else 0. Regarding the subcategories "doi", "pubmed" and "title", for which the same subcategory was retrieved *via* crossref or rentrez, distances between the input's string and the candidates' one were calculated. Optimal string alignment (restricted Damerau-Levenshtein distance) was used as a method. Among "publishing details", "original" and "split" categories, three additional metrics were used: If the journal name was present in the original field, a score of 1 was given, else 0. If the name of the first author was present in the original field, a score of 1 was given, else 0. Those three scores were then summed together. All candidates were first ordered according to their crossref score, then by the complement score for related subcategories, then again according to their title-producing organism score, and finally according to their translation distance score. After this reranking step, only the first candidate was kept. Finally, the Pubmed PMCID dictionary (PMC-ids.csv.gz) was used to perform the translations between DOI, PMID, and PMCID (src/2_curating/2_editing/reference/3_cleaning.R).

From the 36,710 initial "original" references, 21,970 (60%) references with sufficient quality were obtained, of which 15,588 (71%) had the organism name in their title. 14,710 (40%) were uploaded to Wikidata. From the 21,953 initial "pubmed" references, 9,452 (43%) references with sufficient quality were obtained, of which 6,098 (65%) had the organism name in their title. 5,553 (25%) were uploaded to Wikidata. From the 37,371 initial "doi" references, 20,139 (54%) references with sufficient quality were obtained, of which 15,727 (78%) had the organism name in their title. 15,351 (41%) were uploaded to Wikidata. From the 29,600 initial "title" references, 17,417 (59%) references with sufficient quality were obtained, of which 12,675 (73%) had the organism name in their title. 10,725 (36%) were uploaded to Wikidata. From the 11,325 initial "split" references, 5,856 (52%) references with sufficient quality were obtained, of which 3,206 (55%) had the organism name in their title. 2,854 (25%) were uploaded to Wikidata. From the 3,314 initial "publishingDetails"

references, 119 (4%) references with sufficient quality were obtained, of which 59 (50%) had the organism name in their title. 58 (2%) were uploaded to Wikidata.

Realignment

In order to fetch back the referenced structure-organism pairs links in the original data, the cleaned structures, cleaned organisms, and cleaned references were re-aligned with the initial entries. This resulted in 6.2M+ referenced structure-organism pairs. Those pairs were not unique, with redundancies among electronic NP resources and different original categories leading to the same final pair (for example, entry reporting InChI=1/C21H20O12/c22-6-13-15(27)17(29)18(30)21(32-13)33-20-16(28)14-11(26)4-8(23)5-12(14)31-19(20)7-1-2-9(24)10(25)3-7/h1-5,13,15,17-18,21-27,29-30H,6H2/t13-,15+,17+,18-,21+/m1/s1 in *Crataegus oxyacantha* or InChI=1S/C21H20O12/c22-6-13-15(27)17(29)18(30)21(32-13)33-20-16(28)14-11(26)4-8(23)5-12(14)31-19(20)7-1-2-9(24)10(25)3-7/h1-5,13,15,17-18,21-27,29-30H,6H2/t13-,15+,17+,18-,21+/m1/s1 in *Crataegus stevenii* both led to OVSQVDMCBVZWGM-DTGCRPNFSA-N in *Crataegus monogyna*). After deduplication, 2M+ unique structure-organism pairs were obtained.

After the curation of all three objects, all of them were put together again. Therefore, the original aligned table containing the original pairs was joined with each curation result. Only entries containing a structure, an organism, and a reference after curation were kept. Each curated object was divided into minimal data (for Wikidata upload) and metadata. A dictionary containing original and curated object translations was written for each object to avoid those translations being made again during the next curation step (src/2_curating/3_integrating.R).

Validation

The pairs obtained after curation were of different quality. Globally, structure and organism translation was satisfactory whereas reference translation was not. Therefore, to assess the validity of the obtained results, a randomized set of 420 referenced structure-organism pairs was sampled in each reference subcategory and validated or rejected manually. Entries were sampled with at least 55 of each reference subcategory present (to get a representative idea of each subcategory) (src/3_analysing/1_sampling.R). An entry was only validated if: *i)* the structure (as any structural descriptor that could be linked to the final sanitized InChIKey) was described in the reference *ii)* the producing organism (as any organism descriptor that could be linked to the accepted canonical name) was described in the reference and *iii)* the reference was describing the occurrence of the chemical structure in the biological organism. Results obtained on the manually analyzed set were categorized according to the initial reference subcategory and are detailed in SI-2. To improve these results, further cleaning of the references was needed. This was done by accepting entries whose reference was coming from a DOI, a PMID, or from a title which restricted Damerau-Levenshtein distance between original and translated was lower than ten or if it was coming from one of the three main journals where occurrences are published (i.e., *Journal of Natural Products*, *Phytochemistry*, or *Journal of Agricultural and Food Chemistry*). For “split”, “publishingDetails” and “original” subcategories, the year of publication of the obtained reference, its journal, and the name of the first author were searched in the original entry and if at least two of them were present, the entry was kept. Entries were then further filtered to keep the ones where the reference title contained the first element of the detected canonical name. Except for COCONUT, exceptions to this filter were made for all DOI-based references. To validate those filtering criteria, an additional set of 100 structure-organism pairs were manually analyzed. F0.5 score was used as a metric. F0.5 score is a modified F1 score where precision has twice more weight than recall.

The F-score was calculated with ß = 0.5, as in Equation 3:

$$F_\beta = \left(1+\beta^2\right) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \quad (3)$$

Based on this first manually validated dataset, filtering criteria (src/r/filter.R) were established to maximize precision and recall. Another 100 entries were sampled, this time respecting the whole set ratios. After manual validation, 97% of true positives were reached on the second set. A summary of the validation results is given in SI-2. Once validated, the filtering criteria were established to the whole curated set to filter entries chosen for dissemination (src/3_analysing/2_validating.R).

Unit Testing

To provide robustness of the whole process and code, unit tests and partial data full-tests were written. They can run on the developer machine but also on the CI/CD system (GitLab) upon each commit to the codebase.

Those tests assess that the functions are providing results coherent with what is expected especially for edge cases detected during the development. The Kotlin code has tests based on JUnit and code quality control checks based on Ktlint, Detekt and Ben Mane’s version plugin.

Data Dissemination

Wikidata

All the data produced for this work has been made available on Wikidata under a Creative Commons 0 license according to Wikidata:Licensing. This license is a “No-right-reserved” license that allows most reuses.

Lotus.NaturalProducts.Net (LNPN)

The web interface is implemented following the same protocol as described in the COCONUT publication (Sorokina et al., 2021) i.e. the data are stored in a MongoDB repository, the backend runs with Kotlin and Java, using the Spring framework, and the frontend is written in React.js, and completely Dockerized. In addition to the diverse search functions available through this web interface, an API is also implemented, allowing programmatic LNPN querying. The complete API usage is described on the “Documentation” page of the website. LNPN is part of the NaturalProducts.net portal, an initiative aimed at gathering diverse open NP resources in one place.

Data Interaction

Data Retrieval

Bulk retrieval of a frozen (2021-05-23) version of LOTUS data is also available at https://osf.io/eydjs/.

lotus-wikidata-exporter allows the download of all chemical compounds with a “found in taxon” property. That way, it does not only get the data produced by this work, but any that would have existed beforehand or that would have been added directly on Wikidata by our users. It makes a copy of all the entities (compounds, taxa, references) into a local triplestore that can be queried with SPARQL as is or converted to a TSV file for inclusion in other projects. It is currently adapted to export directly into the SSOT thus allowing direct reuse by the processing/curation pipeline.

Data Addition

Wikidata

Data is loaded by the Kotlin importer available in the lotus-wikidata-importer repository under a GPL V3 license and imported into Wikidata. The importer processes the curated outputs grouping references, organisms, and compounds together. It then checks if they already exist in Wikidata (using SPARQL or a direct connection to Wikidata depending on the kind of data). It then uses *update* or *insert*, also called *upsert*, the entities as needed. The script currently takes the tabular file of the referenced structure-organism pairs resulting from the LOTUS curation process as input. It is currently being adapted to use directly the SSOT and avoid an unnecessary conversion step. To import references, it first double checks for the presence of duplicated DOIs and utilizes the Crossref REST API to retrieve metadata associated with the DOI, the support for other citation sources such as Europe PMC is in progress. The structure-related fields are only subject to limited processing: basic formatting of the molecular formula by subscripting of the numbers. Due to limitations in Wikidata, the molecule names are dropped if they are longer than 250 characters and likewise the InChI strings cannot be stored if they are longer than 1500 characters.

Uploaded taxonomical DB identifiers are currently restricted to ITIS, GBIF, NCBI Taxon, Index Fungorum, IRMNG, WORMS, VASCAN, and iNaturalist. The taxa levels are currently limited to family, subfamily, tribe, subtribe, genus, species, variety. The importer checks for the existence of each item based on their InChIKey and upserts the compound with the *found in taxon* statement and the associated organisms and references.

LNPN

From the onset, LNPN has been importing data directly from the frozen tabular data of the LOTUS dataset (https://osf.io/hgjdb/). In future versions, LNPN will directly feed on the SSOT.

Data Edition

The bot framework lotus-wikidata-importer was adapted such that, in addition to batch upload capabilities, it can also edit erroneously created entries on Wikidata. As massive edits have a large potential to disrupt otherwise good data, progressive deployment of this script is used, starting by editing progressively 1, 10, then 100 entries that are manually checked. Upon validation of 100 entries, the full script is run and check its behavior checked at regular intervals. An example of a corrected entry is as follows: https://www.wikidata.org/w/index.php?title=Q105349871&type=revision&diff=1365519277&oldid=1356145998

Curation interface

A web-based (Kotlin, Spring Boot for the back-end, and TypeScript with Vue for the front-end) curation interface is currently in construction. It will allow mass-editing of entries and navigate quick navigation in the SSOT for the curation of new and existing entries. This new interface is intended to become open to the public to foster the curation of entries by further means, driven by the users. In line with the overall LOTUS approach, any modification made in this curation interface will be mirrored after validation on Wikidata and LNPN.

Code Availability

General Repository

All programs written for this work can be found in the following group: https://gitlab.com/lotus7.

Processing

The source data curation system is available at https://gitlab.com/lotus7/lotus-processor. This program takes the source data as input and outputs curated data, ready for dissemination. The first step involves checking if the source data has already been processed. If not, all three elements (biological organism, chemical structures, and references) are submitted to various steps of translation and curation, before validation for dissemination.

Wikidata

Import

The Wikidata importer is available at https://gitlab.com/lotus7/lotus-wikidata-importer. This program takes the processed data resulting from the lotusProcessor subprocess as input and uploads it to Wikidata. It performs a SPARQL query to check which objects already exist. If needed, it creates the missing objects. It then updates the content of each object. Finally, it updates the chemical compound page with a “found in taxon” statement complemented with a “stated in” reference.

Export

The Wikidata exporter is available at https://gitlab.com/lotus7/lotus-wikidata-exporter. This program takes the structured data in Wikidata corresponding to chemical compounds found in taxa with a reference associated as input and exports it in both RDF and tabular formats for further use. Two subsequent options are (a) that the end-user can directly use the exported data.; or (b) that the exported data, which can be new or modified since the last iteration, is used as new source data in lotusProcessor.

LNPN

The LNPN website and processing system is available at https://github.com/mSorok/LOTUSweb. This system takes the processed data resulting from the lotusProcessor as input and uploads it on https://lotus.naturalproducts.net. The repository is not part of the main GitLab group as it benefits from already established pipelines developed by CS and MS. The website allows searches from different points of view, complemented with taxonomies for both the chemical and biological sides. Many chemical molecular properties and molecular descriptors that otherwise are unavailable in Wikidata are also provided.

Code Freezing

All repository hyperlinks in the manuscript point to the preprint branches by default. The links contain all programs and code before submission (2021-02-23) and will eventually be updated to a publication branch using modifications resulting from the peer-reviewing process. As the code evolves, readers are invited to refer to the main branch of each repository for the most up-to-date code. A frozen version (2021-02-23) of all programs and code is also available in the LOTUS OSF repository (https://osf.io/pmgux/).

Programs and packages

R

The R versions used for the project were 4.0.2 up to 4.1, and R-packages used were, in alphabetical order: ChemmineR (3.42.1) (Cao et al., 2008), chorddiag (0.1.2) (Flor, 2020), ClassyfireR (0.3.6) (Djoumbou Feunang et al., 2016), data.table (1.13.6) (Dowle and Srinivasan, 2020), DBI (1.1.1) (R Special Interest Group on Databases (R-SIG-DB) et al., 2021), gdata (2.18.0) (Warnes et al., 2017), ggalluvial (0.12.3) (Brunson, 2020), ggfittext (0.9.1) (Wilkins, 2020), ggnewscale (0.4.5) (Campitelli, 2021), ggraph (2.0.4) (Pedersen, 2020), ggstar (1.0.1) (Xu, 2021), ggtree (2.4.1) (Yu et al., 2016), ggtreeExtra (1.0.1) (Xu et al., 2021), Hmisc (4.4-2) (Jr et al., 2020), jsonlite (1.7.2) (Ooms, 2014), pbmcapply (1.5.0) (Kuang et al., 2019), plotly (4.9.3) (Sievert, 2020), rcrossref(1.1.0) (Chamberlain et al., 2020), readxl (1.3.1) (Wickham and Bryan, 2019), rentrez (1.2.3) (Winter, 2017), rotl (3.0.11) (Michonneau et al., 2016), rvest (0.3.6) (Wickham, 2020), splitstackshape (1.4.8) (Mahto, 2019), RSQLite (2.2.3) (Müller et al., 2021), stringdist (0.9.6.3) (Loo, 2014), stringi (1.5.3) (Gagolewski, 2020), tidyverse (1.3.0) (Wickham et al., 2019), treeio (1.14.3) (Wang et al., 2020), UpSetR (1.4.0) (Gehlenborg, 2019), vroom (1.3.2) (Hester and Wickham, 2020), webchem (1.1.1) (Szöcs et al., 2020), XML (3.99-05) (Lang, 2020), xml2 (1.3.2) (Wickham et al., 2020)

Python

The Python version used was 3.8.6, and the Python packages utilized were, in alphabetical order: faerun (0.3.2) (Probst and Reymond, 2018a), map4 (1.0) (Capecchi et al., 2020), matplotlib (3.1.3) (Hunter, 2007), Molvs (0.1.1), pandas (1.1.4) (Reback et al., 2020), rdkit (2021.03.1) ("RDKit: Open-source cheminformatics," 2021), scipy (1.5.0) (Virtanen et al., 2020), tmap (1.0.4) (Probst and Reymond, 2020).

Kotlin

Kotlin packages used were as follows: Common: Kotlin 1.4.21 up to 1.4.30, Univocity 2.9.0, OpenJDK 15, Kotlin serialization 1.0.1, konnector 0.1.27, Log4J 2.14.0 Wikidata Importer Bot:, WikidataTK 0.11.1, CDK 2.3 (Willighagen et al., 2017), RDF4J 3.6.0, Ktor 1.5.0, KotlinXCli 0.3.1, Wikidata data processing: Shadow 5.0.0 Quality control and testing: Ktlint 9.4.1, Kotlinter 3.3.0, Detekt 1.15.0, Ben Mane's version plugin 0.36.0, Junit 5.7.0

Additional executable files

GNFinder v.0.12.1, GNVerifier v.0.3.1, OPSIN v.2.5.0 (Lowe et al., 2011)

Data Availability

A snapshot of the obtained data at the time of submission is available at the following OSF repository (data): https://osf.io/pmgux/. The https://lotus.nprod.net website is intended to gather news and features related to the LOTUS initiative in the future.

Acknowledgments

JLW and PMA are thankful to the Swiss National Science Foundation for supporting part of this project through the SNF Sinergia grant CRSII5_189921. JB and AR are really thankful to JetBrains for the Free educational license of IntelliJ and the excellent support received on Youtrack. JB, JGG, and GFP gratefully acknowledge the support of this work by grant U41 AT008706 and supplemental funding to P50 AT000155 from NCCIH and ODS of the NIH. MS and CS are supported by the German Research Foundation within the framework ChemBioSys (Project-ID 239748522, SFB 1127). The work on the Wikidata IDSM/Sachem endpoint was supported by an ELIXIR CZ research infrastructure project grant (MEYS Grant No: LM2018131) including access to computing and storage facilities. The authors would like to thank Dmitry Mozzherin for his work done for the Global Names Architecture and related improvements. The authors would also like to thank Layla Michán for starting to add pigment information on Wikidata. EW and DM acknowledge the Scholia grant from the Alfred P. Sloan Foundation under grant number G-2019-11458. The authors would also like to thank contributors of all electronic NP resources used in this work and the NP community at large.

Competing interests

The authors declare no competing interest.

Author contributions

	Conceptualization	Data curation	Formal analysis	Funding acquisition	Investigation	Methodology	Project administration	Resources	Software	Supervision	Validation	Visualization	Writing - original draft	Writing - review and editing	Additional - LNPN website	Additional - NAPRALERT	Additional - Sachem, IDSM	Additional - Wikidata	**Total**
AG												1		1					**2**
AR	3	3	3		3	3	3		3		3	3	3	3				2	**35**
CS				1				1						1	1				**4**
DM														1				2	**3**
EW														1				2	**3**
GFP				1				2						2		1			**6**
JB	3	2	3		2	3	2	1	3	2	2			2		1		3	**29**
JGa														2			2		**4**
JGr														1		1			**2**
J-LW				1				2						2					**5**
JV				1				1									2		**4**
MS									2					2	3				**7**
P-MA	3	2	3	1	2	3	2	1	2	3	2		3	3				1	**31**
RP																		1	**1**
RS														1				1	**2**
Total	**9**	**7**	**9**	**5**	**7**	**9**	**7**	**8**	**10**	**5**	**7**	**4**	**6**	**22**	**4**	**3**	**4**	**12**	

Supporting Information

SI 1 Data Sources List

Table SI-1: Data Sources List

database	type	initial retrieved unique observations	cleaned referenced structure-organism pairs	pairs validated for wikidata export	website	article	retrieval	license	contact	varia
afrotryp	open	312	135	28	-	article (Ibezim et al., 2017)	download	license	Fidele Ntie-Kang or Ngozi Justina Nwodo	-
alkamid	open	4,434	2,582	2,076	website	article (Boonen et al., 2012)	script	license	Bart De Spiegeleer	-
biofacquim	open	531	683	534	website (old version)	article_old article_new (Pilón-Jiménez et al., 2019)	download	license	José Medina-Franco	-
biophytmol	open	546	628	353	website	article (Sharma et al., 2014)	script	license	Anshu Bhardwaj	website often down
carotenoiddb	open	2,922	1,199	639	website	article (Yabuzaki, 2017)	script	license	yzjunko@gmail.com	-
coconut	open	583,623	345,328	34,429	website	article (Sorokina and Steinbeck, 2020b)	download	license	Maria Sorokina	zenodo
cyanometdb	open	1,930	1,844	1,774	-	article (Jones et al., 2021)	download	license	elisabeth.janssen@eawag.ch	-
datawarrior	open	589	1,062	102	website	article (Sander et al., 2015)	download	icense	thomas.sander@idorsia.com	no real link to the dataset inside it
dianatdb	open	290	404	27	website	article (Madariaga-Mazón et al., 2021)	download	license	amadariaga@iquimica.unam.mx or kmtzm@unam.mx	-
dnp	commercial	210,832	258,328	-	website	-	script	-	support@taylorfrancis.com	commercial
drduke	open	90,675	9,072	4,266	website	-	download	license	agref@usda.gov	-
foodb	restricted	82,415	361	-	website	-	download	license	jreid3@ualberta.ca (Jennifer)	-
inflamnat	open	665	656	282	-	article (Zhang et al., 2018)	download	license	xiaoweilie@ynu.edu.cn	-
knapsack	open	116,284	143,062	62,727	website	article (Afendi et al., 2012; Shinbo et al., 2006)	script	license	skanaya@gtc.naist.jp	-
metabolights	open	32,928	32,484	4,919	website	article (Haug et al., 2019)	download	license	-	-
mibig	open	1,310	1,142	548	website	article (Kautsar et al., 2019)	download	license	Tilmann Weber or Marnix Medema	-
mitishamba	open	1,073	1,159	368	website	"article" (Derese, Solomon et al., 31st August to 3rd September 2015)	script	license	-	-
nanpdb	open	5,752	6,447	5,303	website	article (Ntie-Kang et al., 2017)	script	license	ntiekfidele@gmail.com stefan.guenther@pharmazie.uni-freiburg.de	-
napralert	commercial	681,401	380,860	261,545	website	article (Graham and Farnsworth, 2010)	-	license	napralert@uic.edu	-
npass	open	290,539	34,403	23,060	website	article (Zeng et al., 2018)	download	license	phacyz@nus.edu.sg jiangyy@sz.tsinghua.edu.cn jiaochen@163.com	-
npatlas	open	29,006	49,968	44,466	website	article (van Santen et al., 2019)	download	license	rlining t@sfu.ca	-
npcare	open	7,763	4,650	2,525	website	article (Choi et al., 2017)	download	license	choihwanho@gmail.com	-

database	type	initial retrieved unique observations	cleaned referenced structure-organism pairs	pairs validated for wikidata export	website	article	retrieval	license	contact	varia
npedia	open	82	99	23	website	article (Tomiki et al., 2006)	script	license	hisyo@riken.jp npd@riken.jp	-
nubbe	open	2,189	2,614	2,389	website	article (Pilon et al., 2017)	-	license	Vanderlan S. Bolzani	-
pamdb	open	3,061	3,198	64	website	article (Huang et al., 2018)	download	license	awilks@rx.umaryland.edu aoglesby@rx.umaryland.edu mkane@rx.umaryland.edu	-
phenolexplorer	open	8,968	12,027	6,536	website	article (Rothwell et al., 2013)	download	license	scalberta@iarc.fr	-
phytohub	open	2,363	1,451	50	website	"article" (Giacomoni et al., 2017)	script	license	claudine.manach@inra.fr	-
procardb	open	6,606	9,933	70	website	article (Nupur et al., 2016)	script	license	Anil Kumar Pinnaka Ashwani Kumar	-
respect	open	2,759	607	263	website	article (Sawada et al., 2012)	download	license	ksaito@psc.riken.jp	-
sancdb	open	861	987	774	website	article (Hatherley et al., 2015)	script	license	Özlem Tastan Bishop	-
streptomedb	open	71,638	38,343	19,021	website	article (Klementz et al., 2016)	download	license	stefan.guenther@pharmazie.uni-freiburg.de	-
swmd	open	1,075	1,616	1,377	website	article (Davis and Vasanthi, 2011)	script	license	Dicky.John@gmail.com	-
tmdb	open	2,116	841	17	website	article (Yue et al., 2014)	script	license	Xiao-Chun Wan Guan-Hu Bao	currently down
tmmc	open	15,033	5,771	2,662	website	article (Kim et al., 2015)	download	license	Jeong-Ju Lee	-
tppt	open	27,182	28,412	941	website	article (Günthardt et al., 2018)	download	license	thomas.bucheli@agroscope.admin.ch	-
unpd	open	340,319	552,307	352,814	website	article (Gu et al., 2013)	-	-	lirongc@pku.edu.cn xiaojxu@pku.edu.cn	-
wakankensaku	open	367	49	41	website	-	script	-	-	-
wikidata	open	nextStep	nextStep	nextStep	website	-	download	-	-	-

SI 2 Summary of the Validation Statistics

Table SI-2: Summary of the Validation Statistics

	First validation dataset (n =420)								Second validation dataset (n = 100)	
Reference Type	**True positives**	**False positives**	**False negatives**	**True negatives**	**Relative abundance**	**Precision**	**Recall**	**$F_{0.5}$ score**	**True positives**	**False negatives**
Original	80	6	7	11	0.31	0.93	0.92	0.92	38	1
Pubmed	37	1	5	6	0.30	0.97	0.88	0.92	5	1
DOI	115	6	0	6	0.19	0.95	1.00	0.97	43	1
Title	38	2	0	16	0.12	0.95	1.00	0.97	7	0
Split	8	0	15	27	0.08	1.00	0.35	0.52	4	0
Publishing details	1	0	1	32	0.01	1.00	0.50	0.67	0	0
Total	279	15	28	98	1.00	-	-	-	**97**	**3**
Corrected total	-	-	-	-	-	**0.96**	**0.89**	**0.91**	-	-

SI 3 Wikidata SPARQL Queries

Query 1 - *Arabidopsis thaliana*

This query answers to the following question:

> What are the compounds found in Mouse-ear cress (*Arabidopsis thaliana*)?

Link: https://w.wiki/3HMX

```
# What are the compounds found in Mouse-ear cress (Arabidopsis thaliana)?
SELECT DISTINCT ?structure ?structureLabel ?structure_inchikey WHERE {
  ?structure wdt:P235 ?structure_inchikey; # get the inchikey
    p:P703 ?statement.                     # statement found in taxon
  ?statement ps:P703 wd:Q158695;           # Wikidata identifier of your taxa of interest (here Arabidopsis
        thaliana).
                                           # You can remove the Qxxxxxx and hit Ctrl+space, type the first letters
        and it should autocomplete
    prov:wasDerivedFrom ?ref.              # get results with a reference
  ?ref pr:P248 ?art.                       # get the reference ID
  ?art wdt:P356 ?art_doi.                  # get the reference DOI
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Query 2 - *β*-sitosterol

This query answers to the following question:

> Which organisms are known to contain *β*-sitosterol?

Link: https://w.wiki/3HLy

```
# Which organisms are known to contain β-sitosterol?
SELECT DISTINCT ?taxon ?taxonName WHERE {
  {
    wd:Q121802 p:P703 ?stmt.          # limits the search to β-sitosterol
    ?stmt ps:P703 ?taxon.             # found in taxon
    {
      ?stmt prov:wasDerivedFrom ?ref.
      ?ref pr:P248 ?art.              # stated in
      ?art wdt:P356 ?art_doi.         # DOI of the reference
    }
  }
  ?taxon wdt:P225 ?taxonName.         # scientific name of the taxon
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Query 3 - *β*-sitosterol stereoisomers

This query answers to the following question:

> Which organisms are known to contain stereoisomers of *β*-sitosterol?

Link: https://w.wiki/3Jgs

```
# Which organisms are known to contain stereoisomers of β-sitosterol?
SELECT ?compound ?compoundLabel ?InChIKey ?taxonname
WITH {
  SELECT ?compound ?InChIKey WHERE {
    wd:Q121802 wdt:P235 ?queryKey .          # β-sitosterol
    ?compound wdt:P235 ?InChIKey .           # get the inchikey of β-sitosterol
    FILTER (regex(str(?InChIKey), concat("^", substr($queryKey,1,14), "-"))) # results containing the first 14
        character of the inchikey of β-sitosterol
    FILTER ( ?InChIKey != ?queryKey )        # but not containing the inchikey of β-sitosterol
  }
} AS %compounds
WHERE {
  INCLUDE %compounds
  ?compound wdt:P703/wdt:P225 ?taxonname .  # found in taxon / taxon scientific name
  ?compound rdfs:label ?compoundLabel.      # compound name
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER BY ASC(?InChIKey)                     # sort by inchikey
```

Query 4 - Pigments

This query answers to the following question:

> Which pigments are found in which taxa, according to which reference?

Link: https://w.wiki/3H3o

```
# Which pigments are found in which taxa, according to which reference?
# special thanks goes to @biocolores (Layla Michán) for updating this information!
SELECT DISTINCT ?compound ?compoundLabel ?taxon ?taxonname ?DOI
WITH {
  SELECT ?compound WHERE {
    ?compound wdt:P31*/wdt:P279* wd:Q161179.  # get pigments
  }
} AS %compounds
WITH {
  SELECT ?compound ?P703statement WHERE {
    INCLUDE %compounds
            ?compound p:P703 ?P703statement.  # check for "found in taxon" statements
  }
} AS %P703statement
WITH {
  SELECT ?compound ?taxon ?DOI WHERE {
    INCLUDE %P703statement
            ?P703statement ps:P703 ?taxon ;   # get the respective taxa
            prov:wasDerivedFrom / pr:P248 [   # get the reference supporting that statement
              wdt:P356 ?DOI                   # get the DOI for the reference
            ] .
  }
} AS %taxa
WHERE {
  {
    INCLUDE %taxa

            ?taxon wdt:P225 ?taxonname .      # get the taxon name
  }
  ?compound rdfs:label ?compoundLabel .       # get compound labels
  FILTER (LANG(?compoundLabel) = "en") .      # filter for English
}
ORDER BY ASC(?compoundLabel)
LIMIT 10000
```

Query 5 - Sister taxon compounds

This query answers to the following question:

> What are examples of organisms where compounds were found in an organism sharing the same parent taxon, but not the organism itself?

Link: https://w.wiki/3HM6

```
# What are examples of organisms where compounds were found in an organism sharing the same parent taxon, but not
        the organism itself?
SELECT DISTINCT ?compound ?compoundLabel ?taxonname_with_compound ?taxonname_without_compound ?parent_taxon WITH{
  SELECT DISTINCT ?compound ?taxon_with_compound ?parent_taxon
  WHERE {
    ?compound wdt:P235 ?inchikey.
    SERVICE bd:sample { ?compound wdt:P703 ?taxon_with_compound . bd:serviceParam bd:sample.limit 1000 }
    ?taxon_with_compound wdt:P171 ?parent_taxon .
  }
} AS %taxon_with_compound
WITH
{
  SELECT DISTINCT ?taxon_without_compound ?parent_taxon ?compound
  WHERE {
    INCLUDE %taxon_with_compound
    ?taxon_without_compound wdt:P171 ?parent_taxon .
    FILTER (?taxon_with_compound != ?taxon_without_compound)
  }
} AS %taxon2
WHERE {
  INCLUDE %taxon_with_compound
  INCLUDE %taxon2
  FILTER NOT EXISTS {  ?compound wdt:P703 ?taxon_without_compound .}
  ?taxon_with_compound wdt:P225 ?taxonname_with_compound .
  ?taxon_without_compound wdt:P225 ?taxonname_without_compound .
  ?compound rdfs:label ?compoundLabel.
  FILTER(LANG(?compoundLabel) = "en").
}
```

Query 6 - *Zephyranthes* sister taxon compounds

This query answers to the following question:

> Which *Zephyranthes* species lack compounds known from at least two species in the genus?

Link: https://w.wiki/3HJf

```
# Which Zephyranthes species lack compounds known from at least two species in the genus?
PREFIX target: <http://www.wikidata.org/entity/Q191364> # Zephyranthes
SELECT DISTINCT ?compound ?compoundLabel ?taxon_with_compound ?another_taxon_with_compound ?taxon_without_compound
WITH
{
  SELECT DISTINCT ?compound ?taxon_YES_1 ?taxon_YES_2
  WHERE {
    ?compound wdt:P703 ?taxon_YES_1 .
    ?compound wdt:P703 ?taxon_YES_2 .
    ?taxon_YES_1 wdt:P171 target: .
    ?taxon_YES_2 wdt:P171 target: .
    FILTER (?taxon_YES_2 != ?taxon_YES_1)
  }
} AS %taxa_with_compound
WITH
{
  SELECT DISTINCT ?taxon_NO ?compound
  WHERE {
    INCLUDE %taxa_with_compound
    ?taxon_NO wdt:P171 target: .
    FILTER (?taxon_YES_1 != ?taxon_NO)
  }
} AS %taxon_without_compond
WHERE {
  INCLUDE %taxa_with_compound
  INCLUDE %taxon_without_compond
  FILTER NOT EXISTS { ?compound wdt:P703 ?taxon_NO .}
  VALUES ?classes {
    wd:Q11173
    wd:Q59199015
  }
  ?taxon_YES_1 wdt:P225 ?taxon_with_compound .
  ?taxon_YES_2 wdt:P225 ?another_taxon_with_compound .
  ?taxon_NO wdt:P225 ?taxon_without_compound .
  ?compound wdt:P31*/wdt:P279* ?classes .
  ?compound rdfs:label ?compoundLabel.
  FILTER(LANG(?compoundLabel) = "en").
}
```

Query 7 - Antibiotic-like compounds

This query answers to the following question:

> How many compounds are structurally similar to compounds labeled as antibiotics? Results are grouped by the parent taxon of the organism they were

Link: https://w.wiki/3HMA

```
# How many compounds are structurally similar to compounds labeled as antibiotics?
# Results are grouped by the parent taxon of the organism they were found in.
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#> # prefixes needed for structural similarity search
PREFIX idsm: <https://idsm.elixir-czech.cz/sparql/endpoint/>
SELECT ?parent_taxon ?parent_taxon_name (COUNT(DISTINCT ?compound) AS ?count) WHERE {
  SERVICE idsm:wikidata {
    SERVICE <https://query.wikidata.org/bigdata/namespace/wdq/sparql> {
      ?antibiotic ((wdt:P279*)/wdt:P2868/wdt:P486) "D000900"; # get antibiotics
        wdt:P233 ?smiles.                                     # get SMILES
    }
    ?compound sachem:similarCompoundSearch _:b40.
    _:b40 sachem:query ?smiles;
      sachem:cutoff "0.9"^^xsd:double.                        # similarity cut-off at 0.9 similarity
  }
  hint:Prior hint:runFirst "true"^^xsd:boolean.
  ?compound wdt:P703 ?taxon.                                  # found in taxon
  ?taxon wdt:P171 ?parent_taxon.                              # get the parent taxon
  OPTIONAL { ?parent_taxon wdt:P225 ?parent_taxon_name. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP BY ?parent_taxon ?parent_taxon_name                     # count per parent taxon
ORDER BY DESC (?count)
```

Query 8 - Triples

This query answers to the following question:

> Which are the available referenced structure-organism pairs? (example limited to 1000 results)

Link: https://w.wiki/3JpE

```
SELECT DISTINCT ?structure ?structure_inchikey ?taxon ?taxon_name ?reference ?reference_doi WHERE {
  ?structure wdt:P235 ?structure_inchikey;       # get the inchikey
    p:P703[                                      # statement found in taxon
     ps:P703 ?taxon;                             # get the taxon
     (prov:wasDerivedFrom/pr:P248) ?reference ]. # get the reference
  ?taxon wdt:P225 ?taxon_name.                   # get the taxon scientific name
  ?reference wdt:P356 ?reference_doi.            # get the reference DOI
}
LIMIT 1000
```

Query 9 - Indolic scaffold

This query answers to the following question:

> Which organisms contain indolic scaffolds? Count occurrences, group and order the results by the parent taxon.

Link: https://w.wiki/3HMD

```
# Which organisms contain indolic scaffold? Group and order the results by the parent taxon.
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#> # prefixes needed for structural similarity search
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX idsm: <https://idsm.elixir-czech.cz/sparql/endpoint/>

SELECT ?parent_taxon ?parent_taxon_name (COUNT(DISTINCT ?compound) AS ?count)  WHERE {
  SERVICE idsm:wikidata {
    ?compound sachem:substructureSearch
              [ sachem:query "NCCC1=CNC2=C1C=CC=C2" ] # indolic scaffold
  }
  hint:Prior hint:runFirst true.                      # hint to evaluate the idsm service first
  ?compound p:P703 ?statement;                        # found in taxon
            wdt:P235 ?inchikey.                       # get the inchikey
  ?statement ps:P703 ?taxon.                          # get the taxon
  ?taxon wdt:P171 ?parent_taxon.                      # get the parent taxon
  ?parent_taxon wdt:P225 ?parent_taxon_name.          # get the parent taxon name
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP BY ?parent_taxon ?parent_taxon_name
ORDER BY DESC (?count)
```

Query 10 - Senior NP chemists

This query answers to the following question:

> How many structure-organism pairs have been referenced by certain authors? (Here, two senior natural products chemists and co-authors of this paper

Link: https://w.wiki/3HML

```
# How many structure-organism pairs have been referenced by certain authors?
# Here, two senior natural products chemists are compared to the late Ferdinand Bohlmann
#defaultView:BarChart
SELECT ?authors_namesLabel (COUNT(DISTINCT(?compound)) AS ?count) WHERE {
  ?compound p:P703 ?stmt.           # statement found in taxon
    ?stmt prov:wasDerivedFrom ?ref. # get the reference
    ?ref pr:P248 ?art.              # stated in
  VALUES ?authors_names {
    wd:Q56084663                    # JLW
    wd:Q40259636                    # GFP
    wd:Q1405133                     # A german chemist of the 20th century ... Ferdinand Bohlmann
  }
  ?art wdt:P50 ?authors_names.      # limit to references containing the author names
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP BY ?authors_namesLabel
ORDER BY DESC (?count)
```

SI 4 Wikidata Entry Creation Tutorial

Tutorial for manual creation
available at https://osf.io/7dk8h/ and https://oolonek.github.io/dendron/notes/235ba226-b0da-4c23-bbb7-c46c4a65d2f1.html

Manual addition of a referenced structure-organism pair to Wikidata

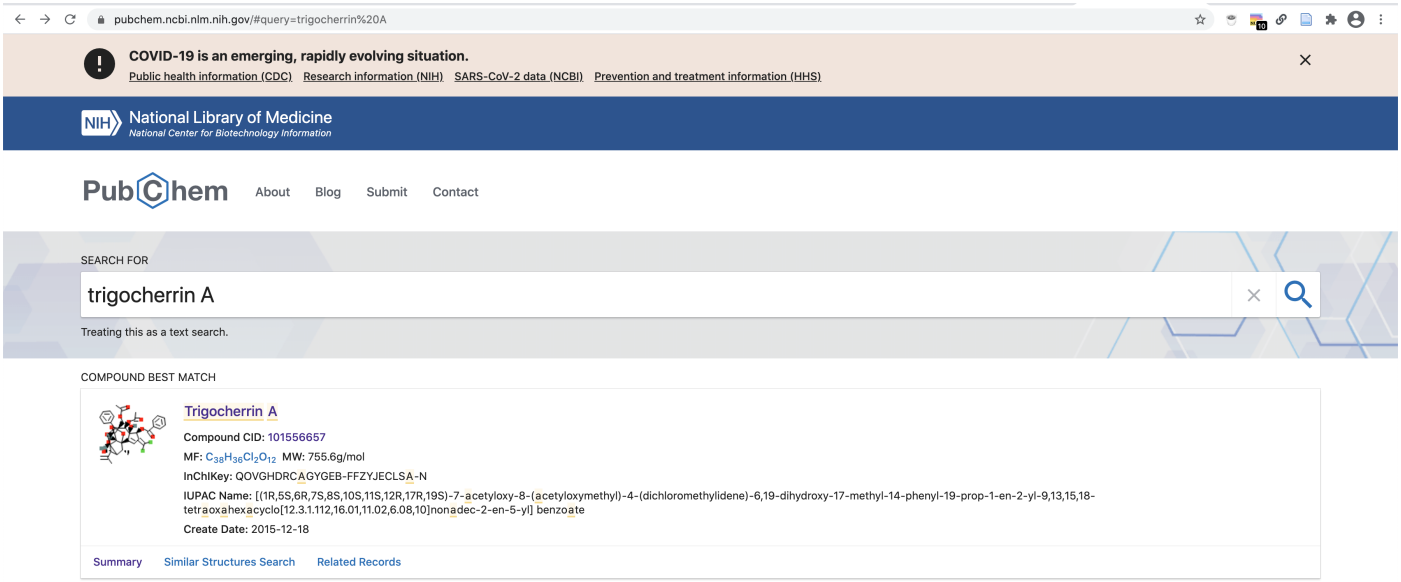
Select a referenced structure-organism pair

Throughout this demonstration, we are going to use the following example: > trigocherrin A is found in *Trigonostemon cherrieri*, as stated in Trigocherrin A, the first natural chlorinated daphnane diterpene orthoester from Trigonostemon cherrieri.

Fetch the information for the referenced structure-organism pair

Structure

Search PubChem for your compound, here trigocherrin A.This leads to https://pubchem.ncbi.nlm.nih.gov/compound/101556657.



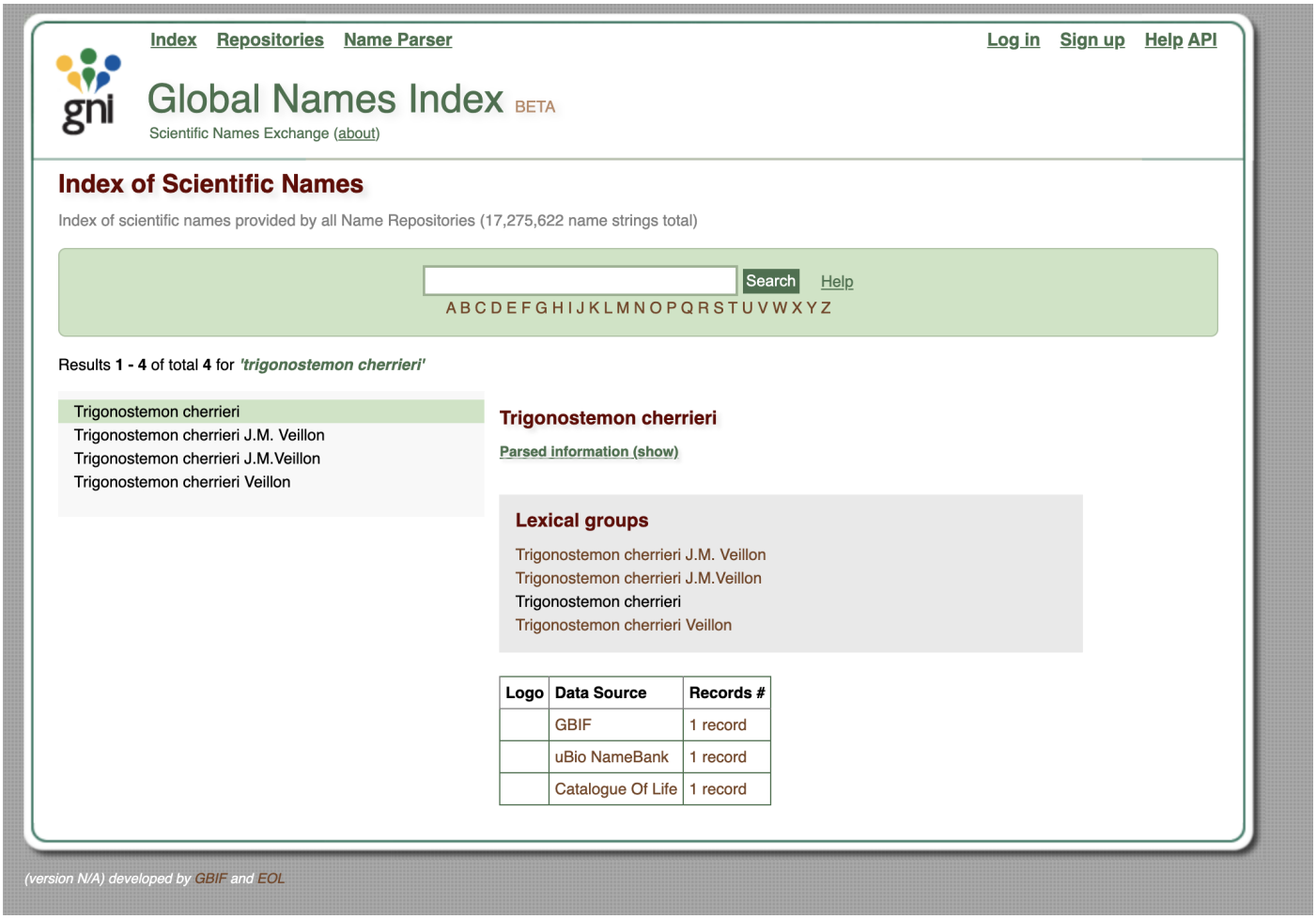
tutorial-image-01

From there, you can fetch the compound's name, InChIKey and InChI as well as its Canonical and Isomeric SMILES. Here we keep, respectively:

```
* trigocherrin A
* QOVGHDRCAGYGEB-FFZYJECLSA-N
* InChI=1S/C38H36Cl2O12/c1-18(2)35(44)27-19(3)37-25-16-24(31(39)40)28(48-32(43)22-12-8-6-9-13-22)36(25,45)33(47-21(5)42)34(17-46-20(4)41)29(49-34)26(37)30(35)51-38(50-27,52-37)23-14-10-7-11-15-23/h6-16,19,26-30,33,44-45H,1,17H2,2-5H3/t19-,26+,27?,28+,29+,30-,33-,34+,35+,36-,37+,38?/m1/s1
* CC1C2C(C3C4C1(C5=CC(=C(Cl)Cl)C(C5(C(C6(C4O6)COC(=O)C)OC(=O)C)O)OC(=O)C7=CC=CC=C7)OC(O2)(O3)C8=CC=CC=C8)(C(=C)C)O
* C[C@@H]1C2[C@]([C@H]3[C@H]4[C@]1(C5=CC(=C(Cl)Cl)[C@@H]([C@]5([C@@H]([C@@]6([C@H]4O6)COC(=O)C)OC(=O)C)O)OC(=O)C7=CC=CC=C7)OC(O3)(O2)C8=CC=CC=C8)(C(=C)C)O
```

Organism

You can check if your organism name is correctly spelled using the Global Names resolver service: http://gni.globalnames.org/name_strings?search_term=trigonostemon+cherrieri&commit=Search.



tutorial-image-02

Alternatively, you can use gnfinder in your command line interface to check for the spelling of your organism string.

```
echo "Trigonostemion cherrieri" | gnfinder find -c -l eng

{
  "metadata": {
    "date": "2021-02-27T18:44:41.640982+01:00",
    "gnfinderVersion": "v0.11.1",
    "withBayes": true,
    "tokensAround": 0,
    "language": "eng",
    "detectLanguage": false,
    "totalWords": 2,
    "totalCandidates": 1,
    "totalNames": 1
  },
  "names": [
    {
      "cardinality": 2,
      "verbatim": "Trigonostemion cherrieri",
      "name": "Trigonostemion cherrieri",
      "odds": 77581.46698350731,
      "start": 0,
      "end": 24,
      "annotationNomenType": "NO_ANNOT",
      "annotation": "",
      "verification": {
        "bestResult": {
          "dataSourceId": 1,
          "dataSourceTitle": "Catalogue of Life",
          "taxonId": "1575885",
          "matchedName": "Trigonostemon cherrieri Veillon",
          "matchedCardinality": 2,
          "matchedCanonicalSimple": "Trigonostemon cherrieri",
          "matchedCanonicalFull": "Trigonostemon cherrieri",
          "classificationPath":
"Plantae|Tracheophyta|Magnoliopsida|Malpighiales|Euphorbiaceae|Trigonostemon|Trigonostemon cherrieri",
          "classificationRank": "kingdom|phylum|class|order|family|genus|species",
          "classificationIds": "3939764|3942634|3942724|3942777|3942795|4210752|1575885",
          "editDistance": 1,
          "stemEditDistance": 1,
          "matchType": "FuzzyCanonicalMatch"
        },
        "dataSourcesNum": 13,
        "dataSourceQuality": "HasCuratedSources",
        "retries": 1
      }
    }
  ]
}
```

For misspellings like *Trigonstemion cherrieri*, gnfinder can help resolve them, in this case to *Trigonostemon cherrieri*.

Reference

Make sure that you have the correct Digital Object Identifier (DOI) for it. For "Trigocherrin A, the first natural chlorinated daphnane diterpene orthoester from Trigonostemon cherrieri", this is **10.1021/ol2030907**. Note that DOIs are uppercase-normalized in Wikidata.

Check for the presence of your compound in Wikidata

Using the compound's InChIKey (i.e. `QOVGHDRCAGYGEB-FFZYJECLSA-N` for trigocherrin A), run a SPARQL query to check if your compound is present in Wikidata or not:

```
SELECT ?item ?itemLabel WHERE {
  VALUES ?classes {
    wd:Q11173 # chemical compound
    wd:Q59199015 # group of stereoisomers
    wd:Q79529 # chemical substance
    wd:Q17339814 # group of chemical substances
    wd:Q47154513 # structural class of chemical compounds
  }
  ?item wdt:P31 ?classes. # instance of
  ?item wdt:P235 'QOVGHDRCAGYGEB-FFZYJECLSA-N'
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Try this query. You can adapt it by replacing the InChIKey with the one for your compound.

Alternatively you can use the following Scholia link (replace by your compounds InChIKey) https://scholia.toolforge.org/inchikey/QOVGHDRCAGYGEB-FFZYJECLSA-N

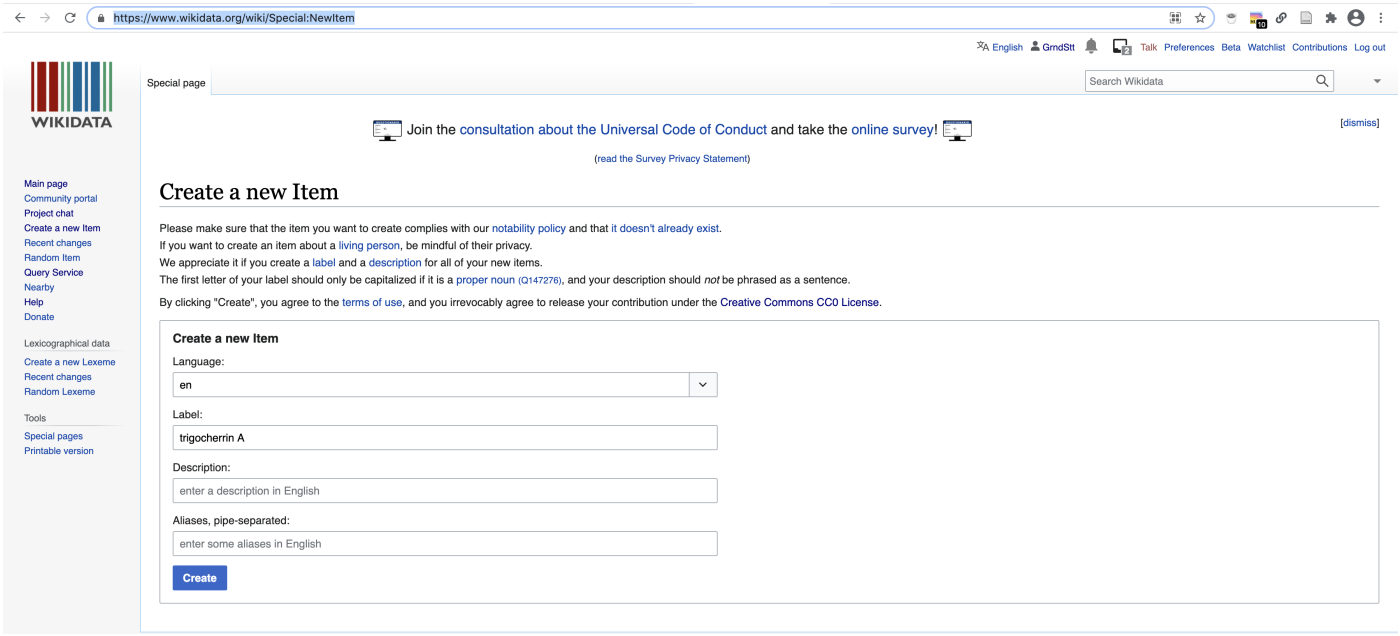
If your compound is already present on Wikidata, you can directly skip to the Add the biological source information section below.

Add your data manually to Wikidata

First, if you do not have a Wikidata account already, it is advisable that you create one via https://www.wikidata.org/wiki/Special:CreateAccount. While an account is not strictly required for manual edits, having one will be useful if you want to contribute more than once, and it helps in getting your contributions recognized. Note that Wikidata accounts are integrated with accounts across the Wikimedia ecosystem, so if you already have an account on, say, any Wikipedia or on Wikispecies, then you can use the same credentials on Wikidata.

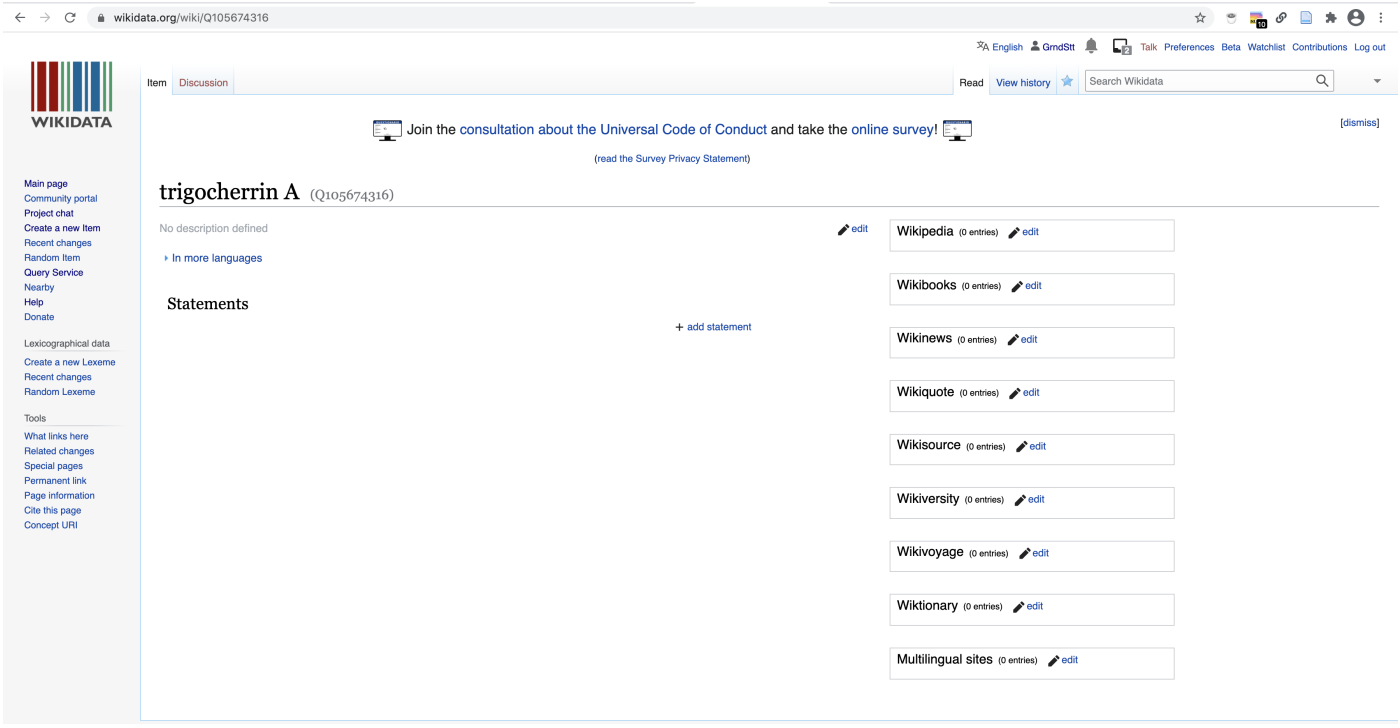
If you are unfamiliar with how Wikidata works, you can start by reading the Wikidata introduction page https://www.wikidata.org/wiki/Wikidata:Introduction and have a look at the Wikidata Tours page https://www.wikidata.org/wiki/Wikidata:Tours.

Now that you are all set up, you can go to Wikidata's page for creating new items, https://www.wikidata.org/wiki/Special:NewItem:



tutorial-image-03

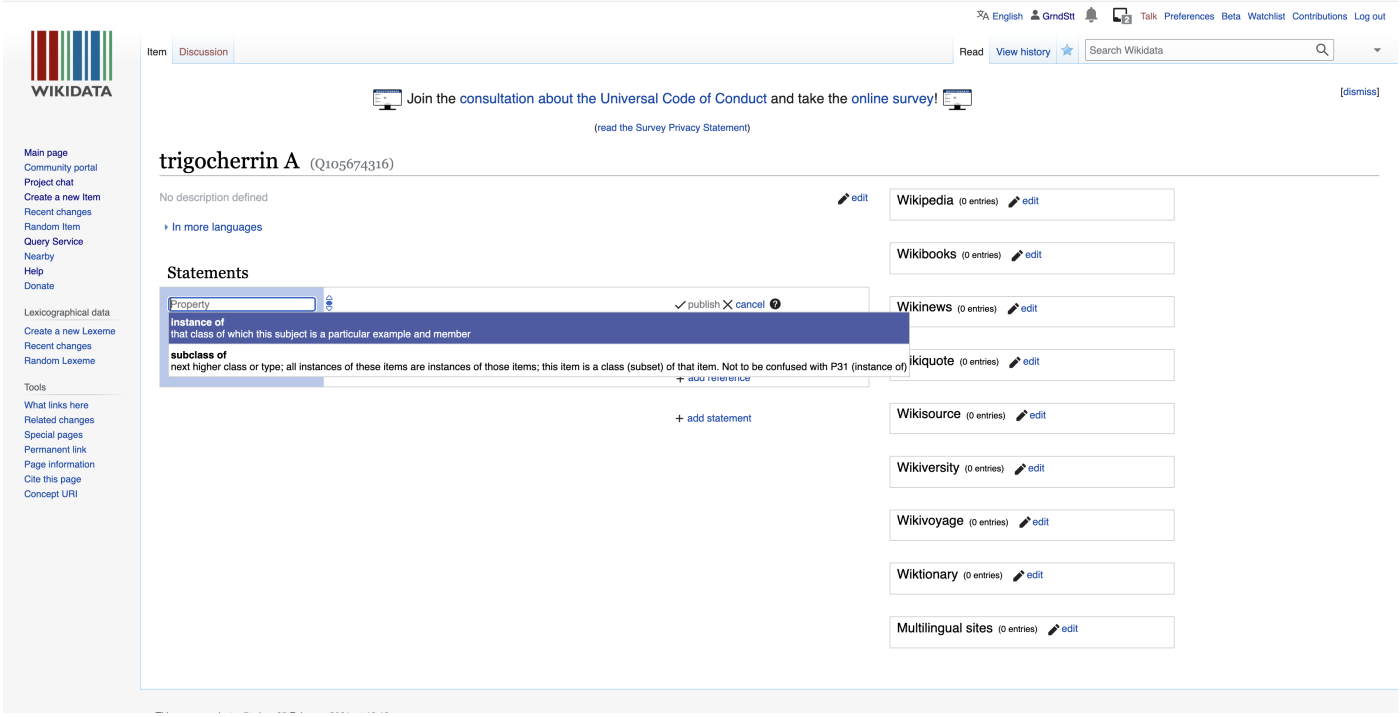
An empty page with a new Wikidata identifier is created



tutorial-image-04

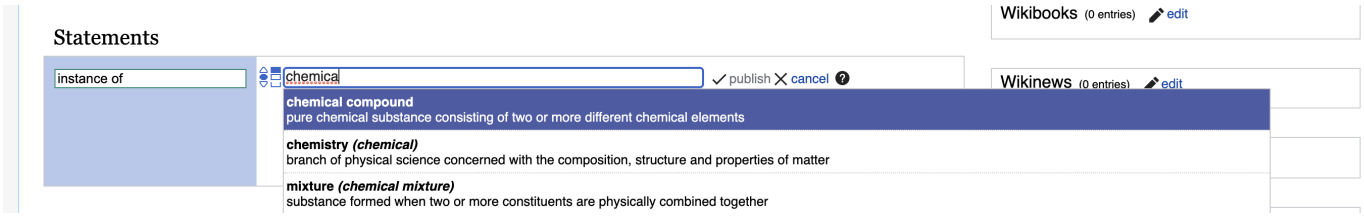
Add the chemical compound information

Create a new statement for `is an instance of`



tutorial-image-05

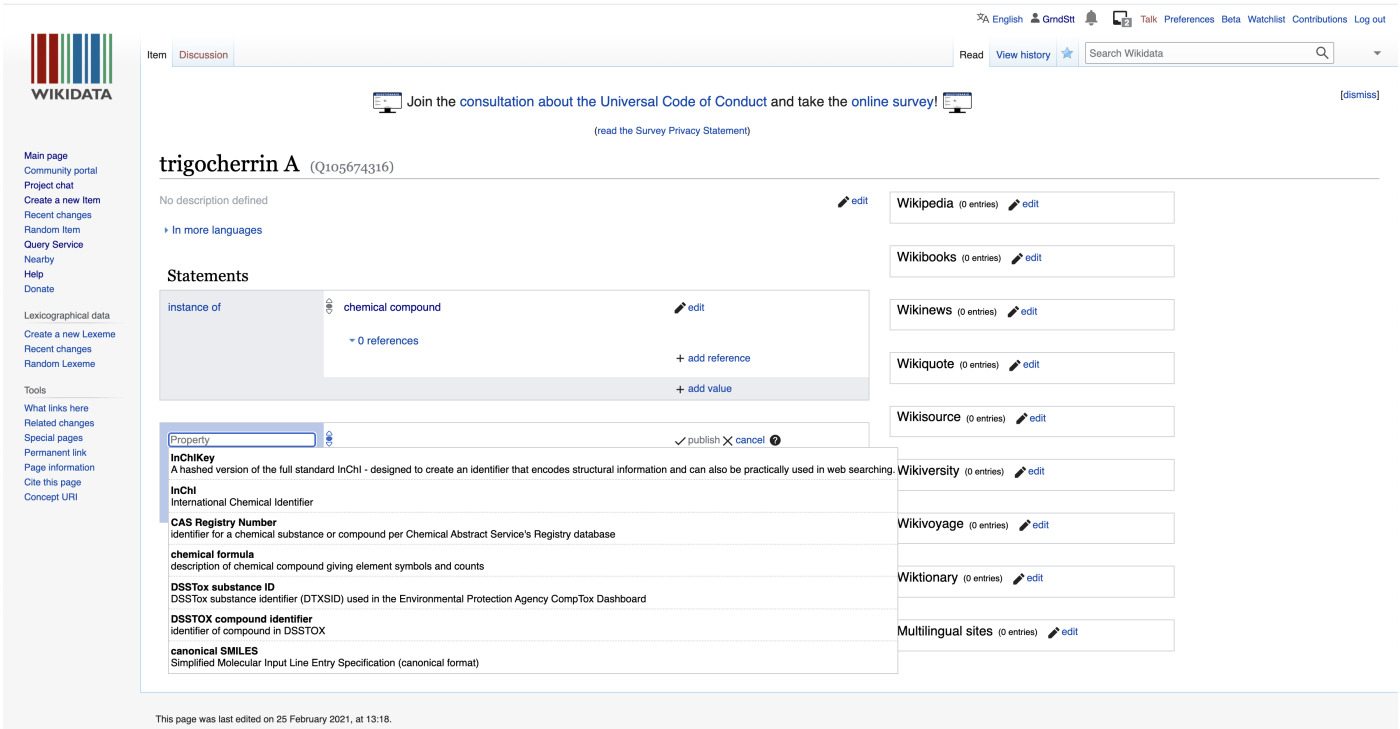
and select chemical compound (i.e. Q11173):



tutorial-image-06

Click `publish` to save your changes and make them public.

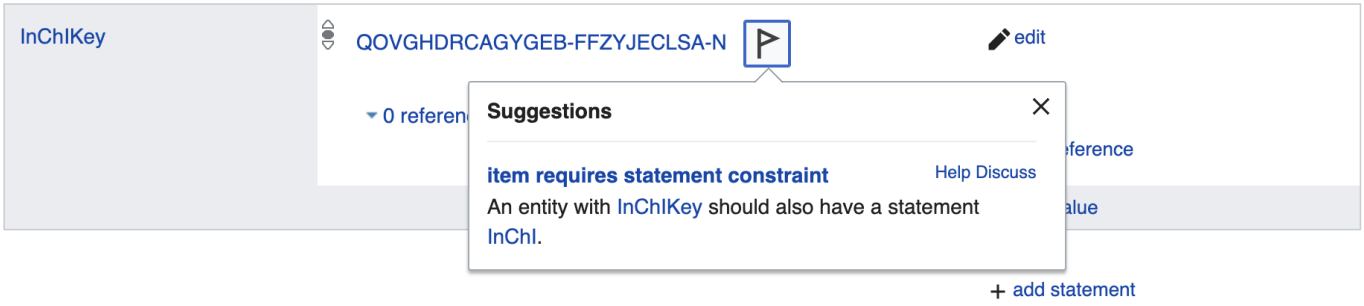
Since you created a new item about an instance of a chemical compound, the user interface will automatically propose to you a set of additional statements commonly found on items about chemical compounds.



tutorial-image-07

You can then go on and fill these in.

Here, we start with the InChIKey. Note the little flag which will automatically tell you if you have some problems with the recently created statements.



tutorial-image-08

Here, Wikidata tells us that if we add an InChiKey, we will need to also add an InChI. Logical, but good to have a reminder!

Let's go ahead and add the InChI string.

Likewise, the addition of an isomeric SMILES string will require us to add a Canonical SMILES.

Note that you might have to copy and paste the SMILES string from PubChem to a plain text editor and then back to Wikidata because of some formatting issues when copy pasting directly from PubChem.

```
C[C@@H]1C2[C@]([C@H]3[C@H]4[C@]1(C5=CC(=C(Cl)Cl)[C@@H]([C@]5([C@@H]
([C@@]6([C@H]4O6)COC(=O)C)OC(=O)C)O)OC(=O)C7=CC=CC=C7)OC(O3)(O2)C8=CC=CC=C8)(C(=C)C)O

CC1C2C(C3C4C1(C5=CC(=C(Cl)Cl)C(C5(C(C6(C4O6)COC(=O)C)OC(=O)C)O)OC(=O)C7=CC=CC=C7)OC(O2)(O3)C8=CC=CC=C8)(C(=C)C)O
```

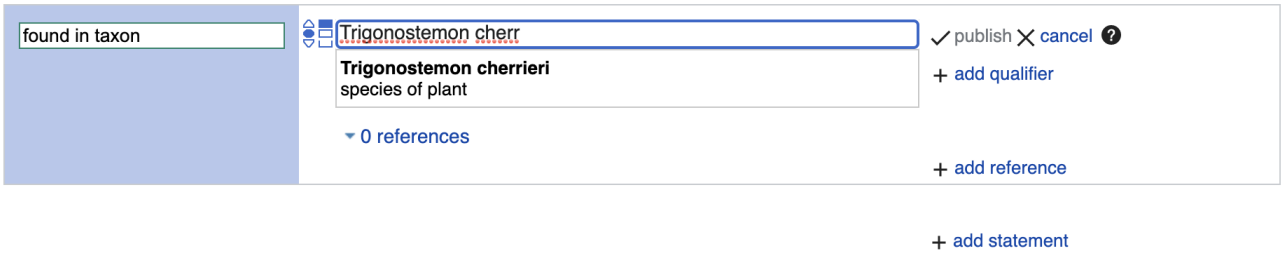
Add the biological source information

Now let's add the `found in taxon` property (P703).

Just click on `Add a new statement` and type in the first letters of the property you want to add:

tutorial-image-09

Again, type in the first letters of the taxon, and if the organism is present, it will autocomplete. Here is how this looks like for *Trigonostemon cherrieri*:



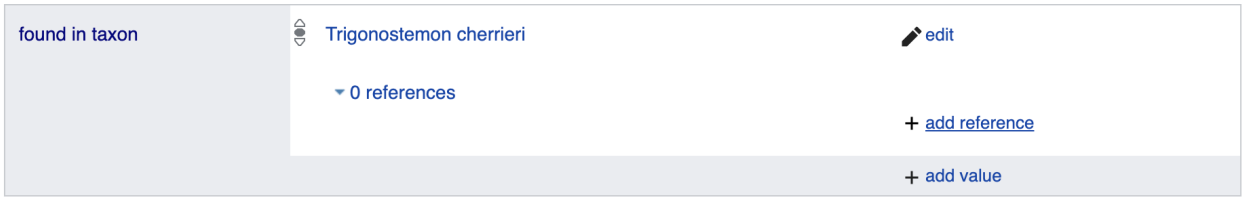
tutorial-image-10

Click `publish` to save your changes and make them public.

If your target taxon is not yet present on Wikidata and you are sure you have a valid taxon name that is spelled correctly, then you can go to https://www.wikidata.org/wiki/Special:NewItem, as described in the Add your data manually to Wikidata section. For items about taxa, the `instance of` statement should have a value `taxon` (i.e. Q16521). As for chemical compounds, the user interface will then suggest to you further statements to add. For taxa, these include taxon name, parent taxon and taxon rank.

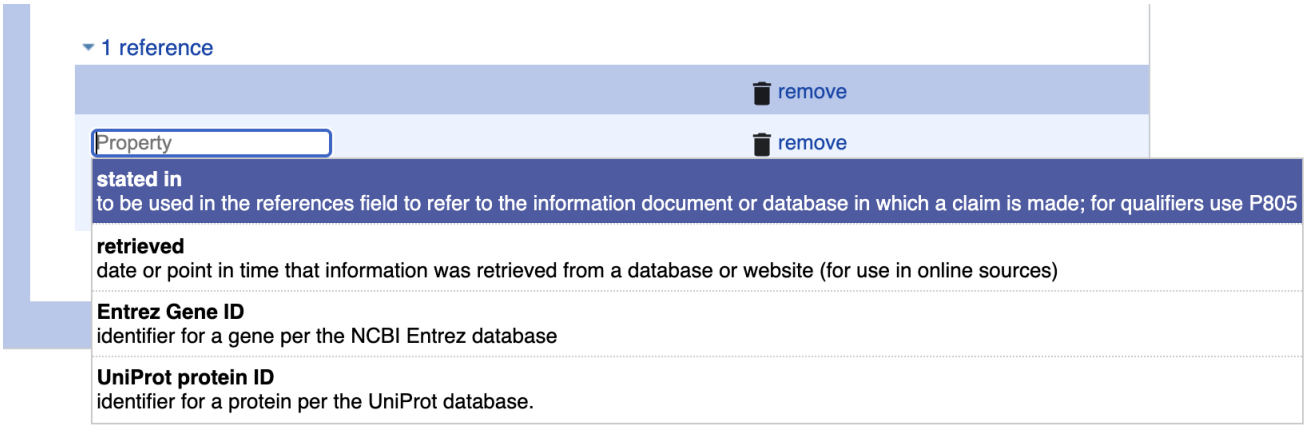
Add the reference documenting the structure-organism pair

Finally, since we report referenced structure-organisms pairs, we need to add the reference for this newly created `compound found in taxon` relationship. This happens on the item about the compound, just below the `found in taxon` statement. Click on the `0 references` link and then on `add reference`:



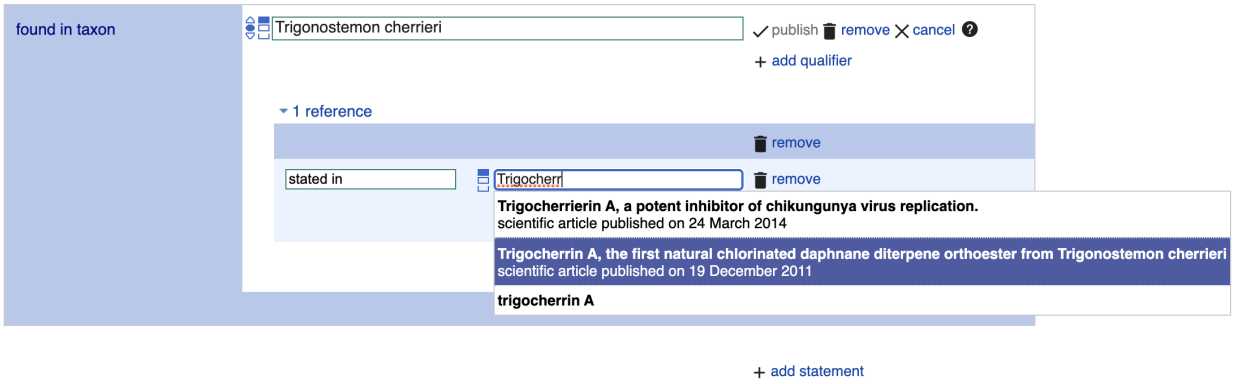
tutorial-image-11

Here, we use the `stated in` property (P248):



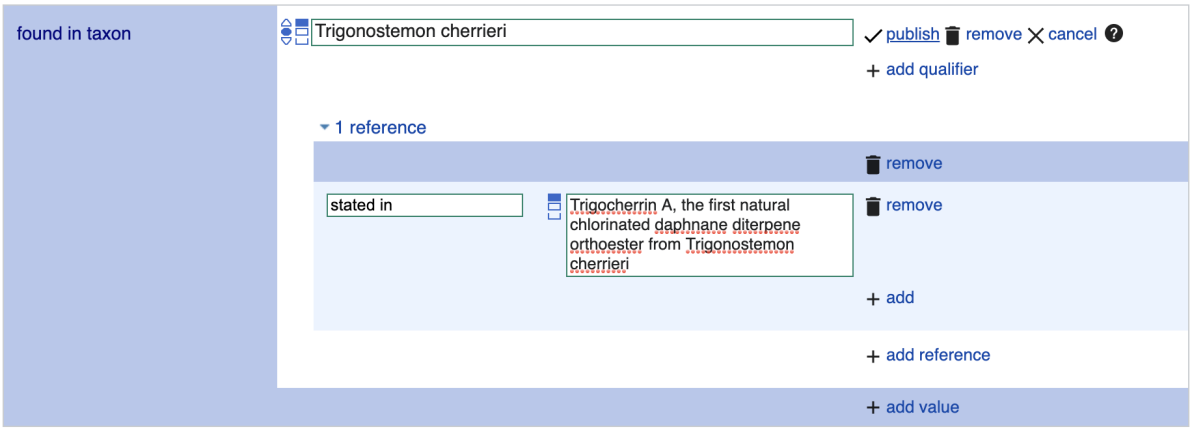
tutorial-image-12

Now, type in the first letters or word of the scientific publication documenting the natural product occurence, autocompletion happens again. Note that multiple publications might have the same title, and that there could be minor differences in punctuation or special characters between the information you and Wikidata have about the same reference. If you are not sure whether your target reference is already in Wikidata, you can use its DOI to check, as outlined in the Check whether your target reference is already on Wikidata section.



tutorial-image-13

Click `publish` to save your changes and make them public.



tutorial-image-14

Check whether your target reference is already on Wikidata

If you are not sure whether your target reference is already in Wikidata, you can use its DOI to check. For our DOI `10.1021/ol2030907`, the URL https://scholia.toolforge.org/doi/***10.1021/ol2030907*** will lead you to a Scholia page about this publication: https://scholia.toolforge.org/work/Q83059010.

Scholia visualizes information from Wikidata, so if it has an entry for your target reference, then so does Wikidata, and both of them will use the same identifier (in this case Q83059010). If you prefer to resolve your DOI to Wikidata directly, you can do so by using the uppercase-normalized DOI in the following URL pattern: https://hub.toolforge.org/P356:**10.1021/OL2030907**, which will lead you to the respective Wikidata page, in this case Q83059010.

If you think that no Wikidata entry exists for your target reference, you can use the DOI in the URL pattern https://sourcemd.toolforge.org/index_old.php?id=**10.1021/ol2030907**&doit=Check+source, which will trigger a check with both Crossref and Wikidata, and if no Wikidata entry can be found, the metadata from Crossref will be fetched and presented to you for creating the respective Wikidata item semi-automatically. Using such semi-automated workflows does require and account that is a minimum number of days old and has made a minimum number of edits on Wikidata.

If you are interested the annotation of article with topics in Scholia here is a tutorial https://laurendupuis.github.io/Scholia_tutorial/

Voilà !

You have added your first referenced structure-organism relationship to Wikidata and made a valuable contribution to the community. You can add further statements, e.g. `molecular formula`, or `SPLASH code` for linking to spectral data.

The Wikidata entry https://www.wikidata.org/wiki/Q105674316 has been started using these instructions.

You can run a SPARQL query and check that everything went smoothly by modifying the InChiKey line in the following SPARQL query:

```
SELECT ?item ?itemLabel ?taxonLabel ?artLabel WHERE {
  VALUES ?classes {
    wd:Q11173 # chemical compound
    wd:Q59199015 # group of stereoisomers
    wd:Q79529 # chemical substance
    wd:Q17339814 # group of chemical substances
    wd:Q47154513 # structural class of chemical compounds
  }
  ?item wdt:P31 ?classes. # instance of
  ?item wdt:P235 'QOVGHDRCAGYGEB-FFZYJECLSA-N' # InChiKey
  {
    ?item p:P1582 ?stmt. # natural product of taxon
    ?stmt ps:P1582 ?taxon. # natural product of taxon
    OPTIONAL {
      ?stmt prov:wasDerivedFrom ?ref.
      ?ref pr:P248 ?art. # stated in
    }
  }
  UNION
  {
    ?item p:P703 ?stmt. # found in taxon
    ?stmt ps:P703 ?taxon. # found in taxon
    OPTIONAL {
      ?stmt prov:wasDerivedFrom ?ref.
      ?ref pr:P248 ?art. # stated in
    }
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

SI 5 Complement to Figure 7

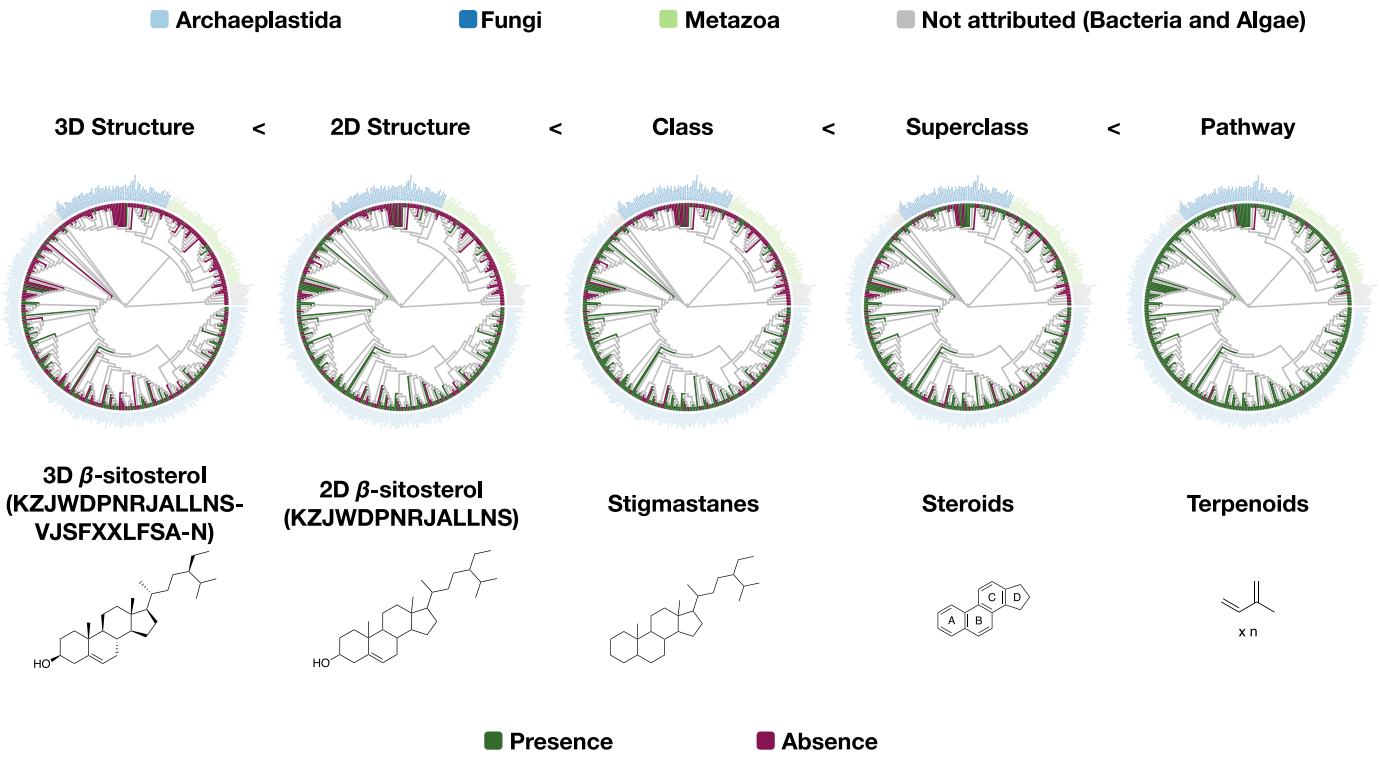


Figure SI-5: Complement to Figure 7: distribution of *β*-sitosterol and related chemical parents among families with at least 50 reported compounds present in LOTUS. Script used for the generation of each tree in the figure is the same (src/4_visualizing/plot_magicTree.R) as for Figure 7 as both figures are related. The figure is available under CC0 license at https://commons.wikimedia.org/wiki/File:Lotus_initiative_1_chemically_interpreted_biological_tree_supplement.svg.

References

Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S. 2012. KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research. *Plant and Cell Physiology* **53**:e1–e1. doi:10.1093/pcp/pcr165

Agosti D, Johnson NF. 2002. Taxonomists need better access to published data. *Nature* **417**:222–222. doi:10.1038/417222b

Allard P-M, Bisson J, Azzollini A, Pauli GF, Cordell GA, Wolfender J-L. 2018. Pharmacognosy in the digital era: shifting to contextualized metabolomics. *Current Opinion in Biotechnology* **54**:57–64. doi:10.1016/j.copbio.2018.02.010

All natural. 2007.. *Nature Chemical Biology* **3**:351–351. doi:10.1038/nchembio0707-351

Balietti S, Mäs M, Helbing D. 2015. On Disciplinary Fragmentation and Scientific Progress. *PLOS ONE* **10**:e0118747. doi:10.1371/journal.pone.0118747

Bisson J, McAlpine JB, Friesen JB, Chen S-N, Graham J, Pauli GF. 2015. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *Journal of Medicinal Chemistry* **59**:1671–1690. doi:10.1021/acs.jmedchem.5b01009

Bisson J, Simmler C, Chen S-N, Friesen JB, Lankin DC, McAlpine JB, Pauli GF. 2016. Dissemination of original NMR data enhances reproducibility and integrity in chemical research. *Natural Product Reports* **33**:1028–1033. doi:10.1039/c6np00022c

Boonen J, Bronselaer A, Nielandt J, Veryser L, De Tré G, De Spiegeleer B. 2012. Alkamid database: Chemistry, occurrence and functionality of plant N-alkylamides. *Journal of Ethnopharmacology* **142**:563–590. doi:10.1016/j.jep.2012.05.038

Brunson J. 2020. ggalluvial: Layered Grammar for Alluvial Plots. *Journal of Open Source Software* **5**:2017. doi:10.21105/joss.02017

Campitelli E. 2021. ggnewscale: Multiple fill and colour scales in “ggplot2” (manual).

Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. 2008. ChemmineR: a compound mining framework for R. *Bioinformatics* **24**:1733–1734. doi:10.1093/bioinformatics/btn307

Capecchi A, Probst D, Reymond J-L. 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* **12**:43. doi:10.1186/s13321-020-00445-4

Chamberlain S, Zhu H, Jahn N, Boettiger C, Ram K. 2020. rcrossref: Client for Various “CrossRef” “APIs”.

Choi H, Cho SY, Pak HJ, Kim Y, Choi J-y, Lee YJ, Gong BH, Kang YS, Han T, Choi G, Cho Y, Lee S, Ryoo D, Park H. 2017. NPCARE: database of natural products and fractional extracts for cancer regulation. *Journal of Cheminformatics* **9**:2. doi:10.1186/s13321-016-0188-5

Cordell GA. 2017a. Cognate and cognitive ecopharmacognosy — in an anthropogenic era. *Phytochemistry Letters* **20**:540–549. doi:10.1016/j.phytol.2016.10.009

Cordell GA. 2017b. Sixty Challenges – A 2030 Perspective on Natural Products and Medicines Security. *Natural Product Communications* **12**:1934578X1701200. doi:10.1177/1934578x1701200849

Cousijn H, Feeney P, Lowenberg D, Presani E, Simons N. 2019. Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal* **18**:9. doi:10.5334/dsj-2019-009

Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Lemberger T, Murphy F, Polischuk P, Taylor S, Martone M, Clark T. 2018. A data citation roadmap for scientific publishers. *Scientific Data* **5**:180259. doi:10.1038/sdata.2018.259

Davis GDJ, Vasanthi AHR. 2011. Seaweed metabolite database (SWMD): A database of natural compounds from marine algae. *Bioinformation* **5**:361–364. doi:10.6026/97320630005361

de Candolle AP de. 1816. Essai sur les propriétés médicales des plantes, comparées avec leurs formes extérieures et leur classification naturelle. Paris: Chez Crochard, Libraire, doi:10.5962/bhl.title.112422

Defossez E, Pitteloud C, Descombes P, Glauser G, Allard P-M, Walker TWN, Fernandez-Conradi P, Wolfender J-L, Pellissier L, Rasmann S. 2021. Spatial and evolutionary predictability of phytochemical diversity. *Proceedings of the National Academy of Sciences* **118**:e2013344118. doi:10.1073/pnas.2013344118

Derese, Solomon, Ndakala, Albert, Rogo, Michael, Maynim, Cholastica, Oyim, James. 31st August to 3rd September 2015. Mitishamba database: a web based in silico database of natural products from Kenya plants.

Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E, Greiner R, Wishart DS. 2016. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8**:61. doi:10.1186/s13321-016-0174-y

Dowle M, Srinivasan A. 2020. data.table: Extension of $data.\ ame$.

Dührkop K, Nothias LF, Fleischauer M, Ludwig M, Hoffmann MA, Rousu J, Dorrestein PC, Böcker S. 2020. Classes for the masses: Systematic classification of unknowns using fragmentation spectra. *Cold Spring Harbor Laboratory*. doi:10.1101/2020.04.17.046672

Flor M. 2020. chorddiag: Interactive Chord Diagrams.

Gagolewski M. 2020. R package stringi: Character string processing facilities.

GBIF.org. 2020.. *GBIF Home Page*. https://www.gbif.org

Gehlenborg N. 2019. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets.

Giacomoni F, Bento Da Silva AL, Bronze M, Gladine C, Hollman P, Kopec R, Low Yanwen D, Micheau P, Nunes Dos Santos MC, Pavot B, Schmidt G, Morand C, Urpi Sarda M, Vazquez Manjarrez N, Verny M-A, Wiczkowski W, Knox C, Manach C. 2017. PhytoHub, an online platform to gather expert knowledge on polyphenols and other dietary phytochemicals.

Gottlieb OR. 1982. Micromolecular Evolution, Systematics and Ecology. Springer Science and Business Media LLC. doi:10.1007/978-3-642-68641-2

Graham JG, Farnsworth NR. 2010. The NAPRALERT Database as an Aid for Discovery of Novel Bioactive Compounds. Elsevier BV. pp. 81–94. doi:10.1016/b978-008045382-8.00060-5

Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. 2013. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS ONE* **8**:e62839. doi:10.1371/journal.pone.0062839

Günthardt BF, Hollender J, Hungerbühler K, Scheringer M, Bucheli TD. 2018. Comprehensive Toxic Plants–Phytotoxins Database and Its Application in Assessing Aquatic Micropollution Potential. *Journal of Agricultural and Food Chemistry* **66**:7577–7588. doi:10.1021/acs.jafc.8b01639

Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, Tastan Bishop Ö. 2015. SANCDB: a South African natural compound database. *Journal of Cheminformatics* **7**:29. doi:10.1186/s13321-015-0080-8

Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C. 2019. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Research* gkz1019. doi:10.1093/nar/gkz1019

Hegnauer R. 1986a. Phytochemistry and plant taxonomy — an essay on the chemotaxonomy of higher plants. *Phytochemistry* **25**:1519–1535. doi:10.1016/s0031-9422(00)81204-2

Hegnauer R. 1986b. Chemotaxonomie der Pflanzen. Springer Science and Business Media LLC. doi:10.1007/978-3-0348-9314-5

Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. 2013. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5**:7. doi:10.1186/1758-2946-5-7

Helmy M, Crits-Christoph A, Bader GD. 2016. Ten Simple Rules for Developing Public Biological Databases. *PLOS Computational Biology* **12**:e1005128. doi:10.1371/journal.pcbi.1005128

Hester J, Wickham H. 2020. vroom: Read and write rectangular text data quickly (manual).

Hoffmann MA, Nothias L-F, Ludwig M, Fleischauer M, Gentry EC, Witting M, Dorrestein PC, Dührkop K, Böcker S. 2021. Assigning confidence to structural annotations from mass spectra with COSMIC. *Cold Spring Harbor Laboratory*. doi:10.1101/2021.03.18.435634

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **45**:703–714. doi:10.1002/jms.1777

Huang W, Brewer LK, Jones JW, Nguyen AT, Marcu A, Wishart DS, Oglesby-Sherrouse AG, Kane MA, Wilks A. 2018. PAMDB: a comprehensive Pseudomonas aeruginosa metabolome database. *Nucleic Acids Research* **46**:D575–D580. doi:10.1093/nar/gkx1061

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**:90–95. doi:10.1109/mcse.2007.55

Ibezim A, Debnath B, Ntie-Kang F, Mbah CJ, Nwodo NJ. 2017. Binding of anti-Trypanosoma natural products from African flora against selected drug targets: a docking study. *Medicinal Chemistry Research* **26**:562–579. doi:10.1007/s00044-016-1764-y

Jarmusch AK, Wang M, Aceves CM, Advani RS, Aguirre S, Aksenov AA, Aleti G, Aron AT, Bauermeister A, Bolleddu S, Bouslimani A, Caraballo Rodriguez AM, Chaar R, Coras R, Elijah EO, Ernst M, Gauglitz JM, Gentry EC, Husband M, Jarmusch SA, Jones KL, Kamenik Z, Le Gouellec A, Lu A, McCall L-I, McPhail KL, Meehan MJ, Melnik AV, Menezes RC, Montoya Giraldo YA, Nguyen NH, Nothias LF, Nothias-Esposito M, Panitchpakdi M, Petras D, Quinn RA, Sikora N, van der Hooft JJJ, Vargas F, Vrbanac A, Weldon KC, Knight R, Bandeira N, Dorrestein PC. 2020. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nature Methods* **17**:901–904. doi:10.1038/s41592-020-0916-7

Jones MR, Pinto E, Torres MA, Dörr F, Mazur-Marzec H, Szubert K, Tartaglione L, Dell'Aversano C, Miles CO, Beach DG, McCarron P, Sivonen K, Fewer DP, Jokela J, Janssen EM-L. 2021. CyanoMetDB, a comprehensive public database of secondary metabolites from cyanobacteria. *Water Research* **196**:117017. doi:10.1016/j.watres.2021.117017

Jr FEH, Dupont with contributions from C, others many. 2020. Hmisc: Harrell Miscellaneous.

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2019. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research* gkz882. doi:10.1093/nar/gkz882

Kessler A, Kalske A. 2018. Plant Secondary Metabolite Diversity and Species Interactions. *Annual Review of Ecology, Evolution, and Systematics* **49**:115–138. doi:10.1146/annurev-ecolsys-110617-062406

kim H, Wang M, Leber C, Nothias L-F, Reher R, Kang KB, van der Hooft JJJ, Dorrestein P, Gerwick W, Cottrell G. 2020. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *American Chemical Society (ACS)*. doi:10.26434/chemrxiv.12885494.v1

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**:D1102–D1109. doi:10.1093/nar/gky1033

Kim S-K, Nam S, Jang H, Kim A, Lee J-J. 2015. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complementary and Alternative Medicine* **15**:218. doi:10.1186/s12906-015-0758-5

Klementz D, Döring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, Erber A, Santillana I, Thomas OS, Bechthold A, Günther S. 2016. StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes. *Nucleic Acids Research* **44**:D509–D514. doi:10.1093/nar/gkv1319

Kratochvíl M, Vondrášek J, Galgonek J. 2019. Interoperable chemical structure search service. *Journal of Cheminformatics* **11**:45. doi:10.1186/s13321-019-0367-2

Kratochvíl M, Vondrášek J, Galgonek J. 2018. Sachem: a chemical cartridge for high-performance substructure search. *Journal of Cheminformatics* **10**:27. doi:10.1186/s13321-018-0282-y

Kuang K, Kong Q, Napolitano F. 2019. pbmcapply: Tracking the Progress of Mc*pply with Progress Bar.

Lang DT. 2020. XML: Tools for Parsing and Generating XML Within R and S-Plus.

Lee CJ, Sugimoto CR, Zhang G, Cronin B. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology* **64**:2–17. doi:10.1002/asi.22784

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, Westbrook J. 2020. The TRUST Principles for digital repositories. *Scientific Data* **7**:144. doi:10.1038/s41597-020-0486-7

Loo M. 2014. The stringdist Package for Approximate String Matching. *The R Journal* **6**:111. doi:10.32614/rj-2014-011

Lowe DM, Corbett PT, Murray-Rust P, Glen RC. 2011. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling* **51**:739–753. doi:10.1021/ci100384d

Madariaga-Mazón A, Naveja JJ, Medina-Franco JL, Noriega-Colima KO, Martinez-Mayorga K. 2021. DiaNat-DB: a molecular database of antidiabetic compounds from medicinal plants. *RSC Advances* **11**:5172–5178. doi:10.1039/d0ra10453a

Mahto A. 2019. splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values.

Martens M, Ammar A, Riutta A, Waagmeester A, Slenter D, Hanspers K, A. Miller R, Digles D, Lopes E, Ehrhart F, Dupuis LJ, Winckers LA, Coort S, Willighagen EL, Evelo CT, Pico AR, Kutmon M. 2021. WikiPathways: connecting communities. *Nucleic Acids Research* **49**:D613–D621. doi:10.1093/nar/gkaa1024

McAlpine JB, Chen S-N, Kutateladze A, MacMillan JB, Appendino G, Barison A, Beniddir MA, Biavatti MW, Bluml S, Boufridi A, Butler MS, Capon RJ, Choi YH, Coppage D, Crews P, Crimmins MT, Csete M, Dewapriya P, Egan JM, Garson MJ, Genta-Jouve G, Gerwick WH, Gross H, Harper MK, Hermanto P, Hook JM, Hunter L, Jeannerat D, Ji N-Y, Johnson TA, Kingston DGI, Koshino H, Lee H-W, Lewin G, Li J, Linington RG, Liu M, McPhail KL, Molinski TF, Moore BS, Nam J-W, Neupane RP, Niemitz M, Nuzillard J-M, Oberlies NH, Ocampos FMM, Pan G, Quinn RJ, Reddy DS, Renault J-H, Rivera-Chávez J, Robien W, Saunders CM, Schmidt TJ, Seger C, Shen B, Steinbeck C, Stuppner H, Sturm S, Taglialatela-Scafati O, Tantillo DJ, Verpoorte R, Wang B-G, Williams CM, Williams PG, Wist J, Yue J-M, Zhang C, Xu Z, Simmler C, Lankin DC, Bisson J, Pauli GF. 2019. The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. *Natural Product Reports* **36**:35–107. doi:10.1039/c7np00064b

Michonneau F, Brown JW, Winter DJ. 2016. rotl: an R package to interact with the Open Tree of Life data. *Methods in Ecology and Evolution* **7**:1476–1481. doi:10.1111/2041-210x.12593

Mohamed A, Abuoda G, Ghanem A, Kaoudi Z, Aboulnaga A. 2020. RDFFrames: Knowledge Graph Access for Machine Learning Tools.

Mongia M, Mohimani H. 2021. Repository scale classification and decomposition of tandem mass spectral data. *Scientific Reports* **11**:8314. doi:10.1038/s41598-021-87796-6

Müller K, Wickham H, James DA, Falcon S. 2021. RSQLite: "SQLite" interface for r (manual).

Nielsen FÅ, Mietchen D, Willighagen E. 2017. Scholia And Scientometrics With Wikidata. *Zenodo*. doi:10.5281/zenodo.1036595

Noteborn HPJM, Lommen A, van der Jagt RC, Weseman JM. 2000. Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops. *Journal of Biotechnology* **77**:103–114. doi:10.1016/s0168-1656(99)00210-2

Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, A. Moumbock AF, Malange YI, Njume LE, Yong JN, Sippl W, Günther S. 2017. NANPDB: A Resource for Natural Products from Northern African Sources. *Journal of Natural Products* **80**:2067–2076. doi:10.1021/acs.jnatprod.7b00283

Nupur LNU, Vats A, Dhanda SK, Raghava GPS, Pinnaka AK, Kumar A. 2016. ProCarDB: a database of bacterial carotenoids. *BMC Microbiology* **16**:96. doi:10.1186/s12866-016-0715-6

Ooms J. 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects.

Pedersen TL. 2020. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.

Pierce HH, Dev A, Statham E, Bierer BE. 2019. Credit data generators for data reuse. *Nature* **570**:30–32. doi:10.1038/d41586-019-01715-4

Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS. 2017. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Scientific Reports* **7**:7215. doi:10.1038/s41598-017-07451-x

Pilón-Jiménez B, Saldívar-González F, Díaz-Eufracio B, Medina-Franco J. 2019. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules* **9**:31. doi:10.3390/biom9010031

Probst D, Reymond J-L. 2020. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **12**:12. doi:10.1186/s13321-020-0416-x

Probst D, Reymond J-L. 2018a. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **34**:1433–1435. doi:10.1093/bioinformatics/btx760

Probst D, Reymond J-L. 2018b. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *Journal of Chemical Information and Modeling* **58**:1–7. doi:10.1021/acs.jcim.7b00425

Rasberry L, Willighagen E, Nielsen F, Mietchen D. 2019. Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata. *Research Ideas and Outcomes* **5**:e35820. doi:10.3897/rio.5.e35820

RDKit: Open-source cheminformatics. 2021.

Reback J, McKinney W, Jbrockmendel, Bossche JVD, Augspurger T, Cloud P, Gfyoung, Sinhrks, Hawkins S, Roeschke M, Klein A, Terji Petersen, Tratner J, She C, Ayd W, Naveh S, Garcia M, Schendel J, Hayden A, Saxton D, Jancauskas V, McMaster A, Battiston P, Skipper Seabold, Chris-B1, H-Vetinari, Kaiqi Dong, Hoyer S, Overmeire W, Gorelli M. 2020. pandas-dev/pandas: Pandas 1.1.4. Zenodo. doi:10.5281/zenodo.4161697

Rees J, Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* **5**:e12581. doi:10.3897/bdj.5.e12581

Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remon A, M'Hiri N, Garcia-Lobato P, Manach C, Knox C, Eisner R, Wishart DS, Scalbert A. 2013. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database* **2013**:bat070–bat070. doi:10.1093/database/bat070

R Special Interest Group on Databases (R-SIG-DB), Wickham H, Müller K. 2021. DBI: R database interface (manual).

Rutz A, Dounoue-Kubo M, Ollivier S, Bisson J, Bagheri M, Saesong T, Ebrahimi SN, Ingkaninan K, Wolfender J-L, Allard P-M. 2019. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Frontiers in Plant Science* **10**:1329. doi:10.3389/fpls.2019.01329

SAIKKONEN K. 2004. Evolution of endophyte?plant symbioses. *Trends in Plant Science* **9**:275–280. doi:10.1016/j.tplants.2004.04.005

Sander T, Freyss J, von Korff M, Rufener C. 2015. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling* **55**:460–473. doi:10.1021/ci500588j

Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, Akiyama K, Sakurai T, Matsuda F, Aoki T, Hirai MY, Saito K. 2012. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry* **82**:38–45. doi:10.1016/j.phytochem.2012.07.007

Sedio BE. 2017. Recent breakthroughs in metabolomics promise to reveal the cryptic chemical traits that mediate plant community composition, character evolution and lineage diversification. *New Phytologist* **214**:952–958. doi:10.1111/nph.14438

Sharma A, Dutta P, Sharma M, Rajput NK, Dodiya B, Georrge JJ, Kholia T, Bhardwaj A, OSDD Consortium. 2014. BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. *Journal of Cheminformatics* **6**:46. doi:10.1186/s13321-014-0046-2

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D, Kanaya S. 2006. KNApSAcK: A Comprehensive Species-Metabolite Relationship Database. Springer Science and Business Media LLC. pp. 165–181. doi:10.1007/3-540-29782-0_13

Sievert C. 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC.

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46**:D661–D667. doi:10.1093/nar/gkx1064

Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. 2021. COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics* **13**:2. doi:10.1186/s13321-020-00478-9

Sorokina M, Steinbeck C. 2020a. COCONUT: the COlleCtion of Open NatUral producTs. doi:10.5281/zenodo.3778405

Sorokina M, Steinbeck C. 2020b. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **12**:20. doi:10.1186/s13321-020-00424-9

Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. 2020. **webchem** : An *R* Package to Retrieve Chemical Information from the Web. *Journal of Statistical Software* **93**. doi:10.18637/jss.v093.i13

Tomiki T, Saito T, Ueki M, Konno H, Asaoka T, Suzuki R, Uramoto M, Kakeya H, Osada H. 2006. RIKEN natural products encyclopedia (RIKEN NPEdia), a chemical database of RIKEN natural products depository (RIKEN NPDepo). *Proceedings of the Symposium on Chemoinformatics* **2006**:JL6–JL6. doi:10.11545/ciqs.2006.0.JL6.0

Tsugawa H. 2018. Advances in computational metabolomics and databases deepen the understanding of metabolisms. *Current Opinion in Biotechnology* **54**:10–17. doi:10.1016/j.copbio.2018.01.008

U.S. Department of Agriculture, Agricultural Research Service. Dr. Duke's Phytochemical and Ethnobotanical Databases. 1992–2016.. *Home Page*. https://phytochem.nal.usda.gov/

van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Neto FC, Castaño-Espriu L, Chang C, Clark TN, Cleary Little JL, Delgadillo DA, Dorrestein PC, Duncan KR, Egan JM, Galey MM, Haeckl FPJ, Hua A, Hughes AH, Iskakova D, Khadilkar A, Lee J-H, Lee S, LeGrow N, Liu DY, Macho JM, McCaughey CS, Medema MH, Neupane RP, O'Donnell TJ, Paula JS, Sanchez LM, Shaikh AF, Soldatou S, Terlouw BR, Tran TA, Valentine M, van der Hooft JJJ, Vo DA, Wang M, Wilson D, Zink KE, Linington RG. 2019. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science* **5**:1824–1833. doi:10.1021/acscentsci.9b00806

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**:261–272. doi:10.1038/s41592-019-0686-2

Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. 2020. Wikidata as a knowledge graph for the life sciences. *eLife* **9**:e52614. doi:10.7554/elife.52614

WAKANKENSAKU - Main Page. 2020. https://wakankensaku.inm.u-toyama.ac.jp/wiki/Main_Page

Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. 2020. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution* **37**:599–603. doi:10.1093/molbev/msz240

Warnes GR, Bolker B, Gorjanc G, Grothendieck G, Korosec A, Lumley T, MacQueen D, Magnusson A, Rogers J, others. 2017. gdata: Various r programming tools for data manipulation (manual).

Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **28**:31–36. doi:10.1021/ci00057a005

Wickham H. 2020. rvest: Easily Harvest (Scrape) Web Pages.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* **4**:1686. doi:10.21105/joss.01686

Wickham H, Bryan J. 2019. readxl: Read Excel Files.

Wickham H, Hester J, Ooms J. 2020. xml2: Parse XML (manual).

Wilkins D. 2020. ggfittext: Fit Text Inside a Box in "ggplot2".

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**:160018. doi:10.1038/sdata.2016.18

Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C. 2017. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics* **9**:33. doi:10.1186/s13321-017-0220-4

Winter D. 2017. rentrez: An R package for the NCBI eUtils API. *The R Journal* **9**:520. doi:10.32614/rj-2017-058

Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. 2010. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **26**:2647–2648. doi:10.1093/bioinformatics/btq476

Xu S. 2021. ggstar: Star Layer for "ggplot2" (manual).

Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, Wu T, Hu E, Yu G. 2021. ggtreeExtra: Compact visualization of richly annotated phylogenetic data. *Research Square Platform LLC*. doi:10.21203/rs.3.rs-155672/v2

Yabuzaki J. 2017. Carotenoids Database: structures, chemical fingerprints and distribution among organisms. *Database* **2017**. doi:10.1093/database/bax004

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2016.: an package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**:28–36. doi:10.1111/2041-210x.12628

Yue Y, Chu G-X, Liu X-S, Tang X, Wang W, Liu G-J, Yang T, Ling T-J, Wang X-G, Zhang Z-Z, Xia T, Wan X-C, Bao G-H. 2014. TMDB: A literature-curated database for small molecular compounds found from tea. *BMC Plant Biology* **14**:243. doi:10.1186/s12870-014-0243-1

Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, Wang Y, Tan Y, Gao D, Wang B, Chen Z, Chen W, Jiang YY, Chen YZ. 2018. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research* **46**:D1217–D1222. doi:10.1093/nar/gkx1026

Zhang R, Lin J, Zou Y, Zhang X-J, Xiao W-L. 2018. Chemical Space and Biological Target Network of Anti-Inflammatory Natural Products. *Journal of Chemical Information and Modeling* **59**:66–73. doi:10.1021/acs.jcim.8b00560