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

This manuscript (permalink) was automatically generated from lotusnprod/lotus-manuscript@21649de on May 3, 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

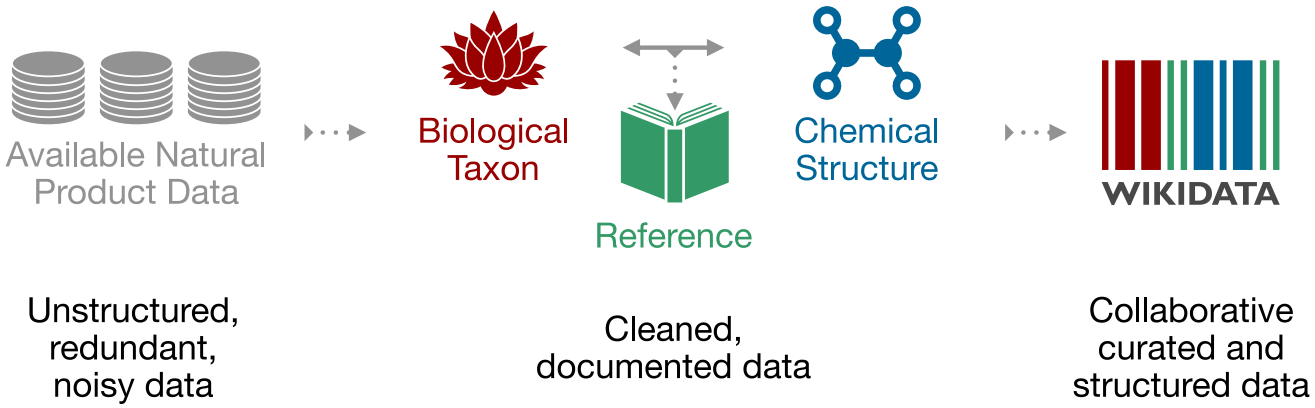
- **Jonathan Bisson** ✉
  0000-0003-1640-9989 · bjonnh · Bjonnh
  Center for Natural Product Technologies and WHO Collaborating Centre for Traditional Medicine (WHO CC/TRM), Pharmacognosy Institute; College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612,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

✉ — correspondence preferred via GitLab Issues. Otherwise, address correspondence to jean-luc.wolfender@unige.ch, bjo@uic.edu, and pierre-marie.allard@unige.ch.

Abstract

As bio- and chemoinformatics are reshaping natural products (NP) research, existing databases impose intrinsic and practical limits regarding access, chemical or taxonomic scope, field standardization, and interoperability. Further limitations result from essential but missing links to the primary literature, which contain the experimental information that documents structures, organisms and their relationships. Sharing such consolidated knowledge via an open platform has strong transformative potential for NP research. The LOTUS initiative has now completed the first steps towards the harmonization, curation, validation, and open dissemination of 500,000+ 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, chemical and taxonomic scope, and interoperability, thereby overcoming many of the limitations of existing electronic resources. Furthermore, embedding LOTUS data into the vast Wikidata knowledge graph facilitates new biological and chemical insights. This opens new possibilities for community curation of data and evolving publication models. All code developed within the LOTUS initiative for data gathering, curation, and dissemination are publicly available on https://gitlab.com/lotus7 and https://github.com/mSorok/LOTUSweb. Collectively, LOTUS evolves NP knowledge management to function in a world of openly shared electronic resources.



Introduction

Evolution of Electronic Natural Product Resources

Natural products (NP) research is a transdisciplinary field with scopes ranging from fundamental structural aspects of naturally-occuring molecular entities to their effects on living organisms, and even the study of chemically-mediated interactions within entire ecosystems. Technological and methodological developments are constantly reshaping NP research, a field steeped in history, with a rich legacy of experimental practices (Allard et al., 2018). In particular, contemporary bioinformatic approaches enable the (re-)interpretation and (re-)annotation of datasets documenting molecular aspects of biodiversity (Olivon et al., 2017). To efficiently annotate previously reported NP, or to identify and report new entities, these tools rely on properly maintained NP database (DB) (Tsugawa, 2018). Despite the ambiguous definition of "natural" ("All natural," 2007), assuming that it is reasonable to consider a NP is a chemical entity found in a living organism, a future-oriented electronic NP resource should contain, at minimum, a list of chemical entities, organisms, and references to the work(s) establishing the links between them. If large, well-structured, and freely accessible DB, composed only of 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) exists, their scarce interlinkage limits their applications. Currently, no open, cross-kingdom and comprehensive electronic NP resource links NP and their producing organisms, along with information about the experimental works describing those links. If referenced structure-organism pairs are critical evidence for pharmacognosy and related NP research, they remain poorly accessible (Cordell, 2017a).

Pioneering efforts to address such issues 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 over 50,000 structures and over 100,000 structure-organism pairs. However, the organism field is not standardized and downloads are complicated. The NAPRALERT dataset (Graham and Farnsworth, 2010), compiled over five decades, gathers annotated data derived from over 200,000 primary NP literature sources and contains 200,000 distinct compound names and structural elements, along with over 500,000 records of distinct structure-organism pairs. In total, it represents over 900,000 total records of pairs due to equivalent structure-organism pairs reported in different citations. NAPRALERT however uses an access model with limited free searches. Finally, the NPAtlas (Santen et al., 2019) is a 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, such as the plant kingdom.

The majority of available electronic NP resources provide entries without referencing their origin, thereby breaking the precious link required for tracing information back to the original data and assessing its quality. Even valuable efforts for compiling NP data made by commercial products, such as the Dictionary of Natural Products (DNP), are missing proper documentation of this critical information, thereby precluding further computational use or exhaustive review. Building on experience with the recently published COlleCtion of Open NatUral producTs (COCONUT) (Sorokina et al., 2021), the LOTUS project addresses these shortcomings. At its current stage of development, LOTUS features >500,000 distributed, referenced structure-organism pairs. After extensive data curation and harmonization, each documented structure-organism pair became standardized at the chemical, biological, and reference level. Collectively, LOTUS offers a high quality, computer-interpretable knowledge base.

Accommodating Principles of FAIRness and TRUSTworthiness

In awareness of the multi-facetted pitfalls associated with implementing, using, and maintaining classical scientific DBs (Helmy et al., 2016), and to enhance current and future sharing, the LOTUS project selected the Wikidata platform for disseminating its resources. Since its creation, Wikidata has focused on cross-disciplinary and multilingual support. It 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. They are machine-interpretable and can be enriched with qualifiers and references. They correspond to ~90 million entries, which can be grouped into classes as diverse as countries, songs, disasters, or chemical compounds. They 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, this article introduces the development and implementation of a workflow for NP occurrences curation and dissemination using both FAIR and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles (Lin et al., 2020). The presented data upload and retrieval procedures ensure optimal accessibility by the research community, allowing any researcher to contribute and reuse the data with a clear and open license (Creative Commons 0).

Despite many advantages, Wikidata hosting of the LOTUS project has some drawbacks. While the SPARQL query language offers a powerful way to query available data, it can also appear intimidating at first for an inexperienced user. Furthermore, some typical queries of molecular electronic NP resources such as structural or spectral search 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. LNPN is periodically updated with the latest LOTUS data. The advantage of this dual hosting is that it yields 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 project and its multiple data interaction options provide the basis for transparent and sustainable ways to access, share, and create knowledge on NP occurrence. More broadly, the LOTUS project will foster cross-fertilization of the fields of chemistry and biology and associated disciplines.

Structure of this Article

The first section presents an overview of the LOTUS project at its current stage of development, keeping in mind that it is a constant work-in-progress. As part of this overview, the central curation and dissemination elements of the LOTUS project are explained in detail. The second section addresses the interaction between LOTUS and its end-users, including data retrieval, addition, and editing. The third section is dedicated to the interpretation of LOTUS data and illustrates the dimensions and qualities of the current LOTUS dataset from the chemical and biological perspectives.

Results & Discussion

Outline of the LOTUS Blueprint

Building on the standards established by each related Wikidata projects (chemistry (Wikidata:WikiProject Chemistry), taxonomy (Wikidata:WikiProject Taxonomy), and source metadata (Wikidata:WikiProject Source MetaData)), a NP chemistry oriented subproject was created (Wikidata:WikiProject Chemistry/Natural products). Its central data is constituted of 3 minimal sufficient objects:

- A chemical compound 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 this DB.
- A reference object describing the structure-organism pair, with the associated article title and a Digital Object Identifier (DOI), a PubMed ID (PMID), or a PubMed Central ID (PMCID).

As data formats are largely inhomogeneous among existing electronic NP resources, fields related to chemical structure, biological organism, and literature reference are variable and essentially not standardized. Therefore, LOTUS implements multiple stages of harmonization, cleaning, validation, and dissemination (Figure 1, stages 1 to 4). LOTUS employs the Single Source of Truth approach (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. The SSOT approach consists of a PostgreSQL DB that structures links and data schemes such that every data element is in a single place. To accommodate the addition of data directly from new data sources or at the Wikidata level), LOTUS' processing pipeline is tailored to efficiently include and diffuse novel or curated data (Figure 1, stage 4). This iterative workflow relies both on data addition and retrieval actions 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.

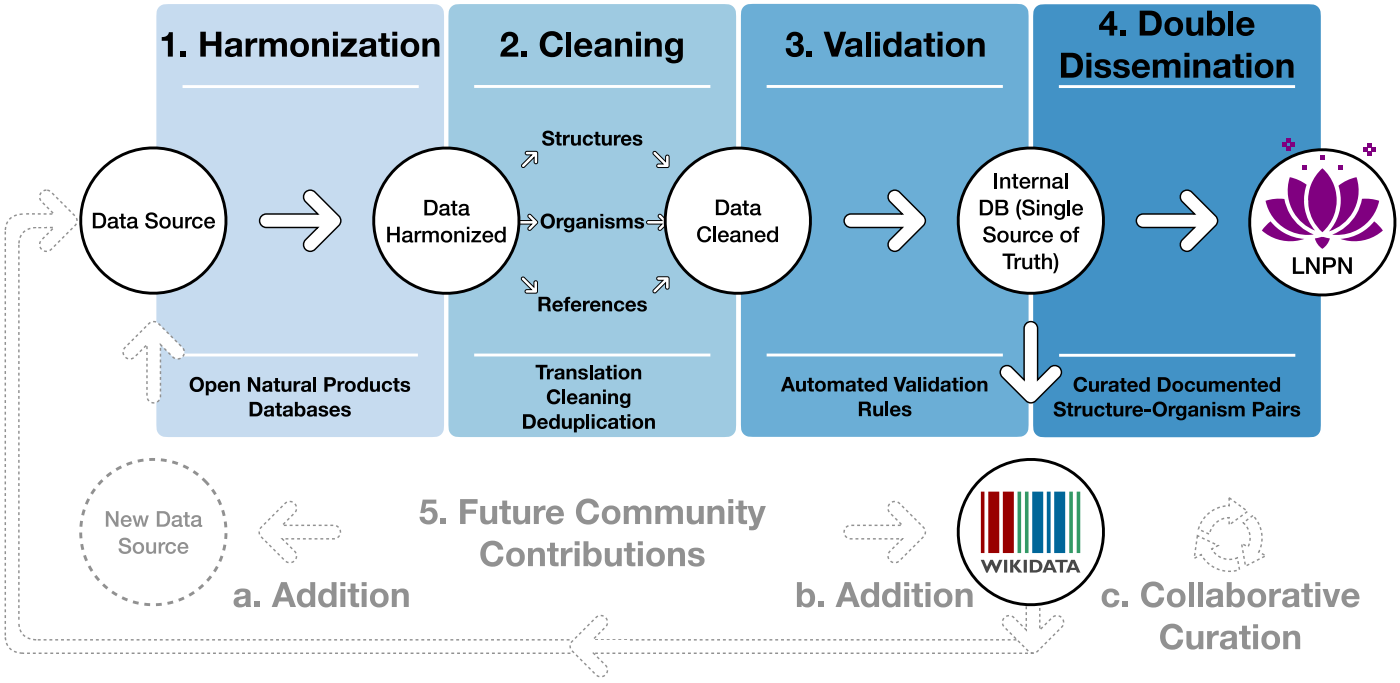


Figure 1: Overview of the workflow in the LOTUS project The process consists of four stages: (1) Harmonization, (2) Cleaning, (3) Validation, and (4) Dissemination. The process was designed to incorporate (5) future contributions either in the form of new data addition (a and b) or via curation of existing data (c). This an iterative process that allows the community to participate in the global effort to document NP occurrences (the "virtuous circle" of NP).

All stages of the process are described in the https://gitlab.com/lotus7 project and at https://github.com/mSorok/LOTUSweb. At the time of submission, over 742,000 LOTUS entries contained a curated chemical structure, biological organism, and reference and are available on both Wikidata and LNPN. As the LOTUS data volume is expected to increase over time, a frozen (as of 2021-02-23) tabular version of this dataset with its associated metadata is made available at https://osf.io/hgjdb/.

Data Harmonization

Multiple data sources were processed and are described hereafter. All publicly accessible electronic NP resources included in COCONUT that contain referenced structure-organism pairs were processed. They were complemented with COCONUT's own structure-organism documented pairs (Sorokina and Steinbeck, 2020a) and the following additional electronic NP resources: Dr. Duke ("U.S. Department of Agriculture, Agricultural Research Service. Dr. Duke's Phytochemical and Ethnobotanical Databases" 1992), Cyanometdb (Jones et al., 2020), Datawarrior (López-López et al., 2019), a subset of NAPRALERT (Graham and Farnsworth, 2010), Wakankensaku ("Main Page - WAKANKENSAKU," n.d.), and DiaNat-DB (Madariaga-Mazón, 2021). The owners of the electronic NP resources not explicitly licensed as open were individually contacted for permission to access and reuse data. The complete list of data sources and related information is available as SI-1. All necessary scripts for gathering and harmonization can be found in the lotusProcessor repository in the src/1_gathering directory. All subsequent iterations with additional data sources (either the updated versions of the same data sources or new ones), will start by comparing the new data sources with previously gathered ones at the SSOT level to curate data only once.

Data Cleaning & Vaidation

As described in Figure 1, the data curation process was divided into four stages: harmonization, cleaning, and validation. Cleaning of each of the three central objects (chemical, biological, and reference) of the referenced pairs was performed. Given the data size with over 2.5M initial entries, manual validation was unfeasible. Curating the references turned out to be a particularly problematic point of the process. Whereas organisms are typically reported by at least their vernacular or scientific denomination and structures by their SMILES, InChI, InChIKey, or image (case not covered in this work), references suffer from largely insufficient reporting standards. Better reporting, together with new tools, such as Scholia (Nielsen et al., 2017; Rasberry et al., 2019), relying on Wikidata, Fatcat, or Semantic Scholar should lead towards improved references-related information retrieval in the future. Despite the poor standardization of the initial reference field, this latter is the only way to establish the validity of the structure-organism pair. Therefore, in addition to the entries curated during data processing, 420 referenced structure-organism pairs were analyzed manually to establish rules for automatic filtering of the entries. This filter was then applied to all entries. To confirm the efficacy of the filtering process, a subset of 100 diverse automatically curated and validated entries was manually checked, and showed 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/hgjdb/.

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

	Structure	Organism	Reference

	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

Table 1 shows an example of a referenced structure-organism pair before and after curation, which resolved the structure to an InChIKey, the organism to a valid taxonomic name, and the reference to a DOI. Challenging examples encountered during the development of the curation process were compiled in an edge case table (tests/tests.tsv) to allow for automatic unit testing. These tests allow a continuous revalidation of any change made to the code, ensuring that no corrected error can reappear.

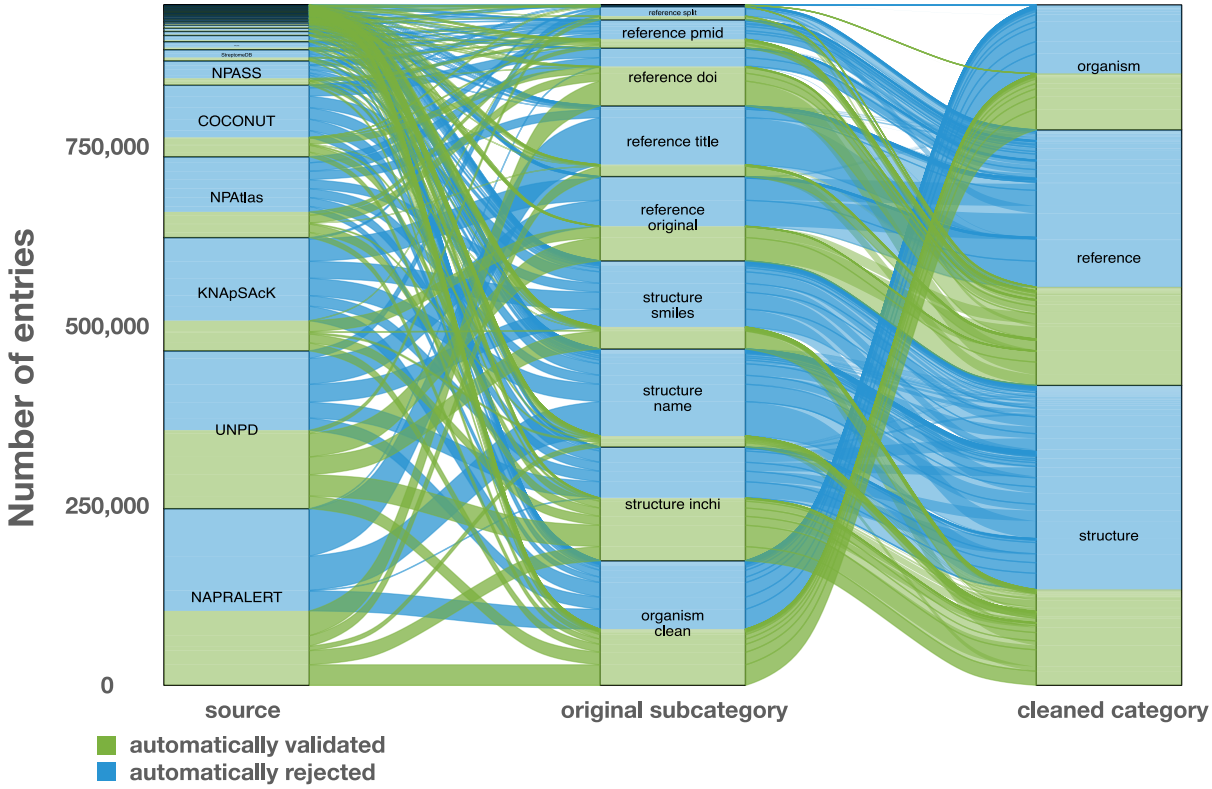


Figure 2: Alluvial plot of the data transformation flow within the LOTUS project during the automated curation and validation processes. The graph demonstrates how curation and validation of each informational object eventually yields curated information. The figure also reflects the relative proportions of the stream of data by showing the contributions from the various sources ("source" block), the composition of the harmonized subcategories ("original subcategory" block), and the validated data after curation ("cleaned category" block). Automatically validated entries are represented in green and rejected ones in blue.

The alluvial plot in Figure 2 illustrates the individual contribution of each source and original subcategory that led to the cleaned categories of structure, organism, and reference. For example, the graph highlights the important contribution of the DOI category of references contained in NAPRALERT towards the current set of validated references. 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 documented structure-organism pairs that formed the processed part of the SSOT. Out of a total of more than 2M pairs, the manual and automatic validation retained 740,000+ pairs (~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

Researchers should benefit from all results of published scientific studies immediately upon publication. This is considered as the foundation of scientific investigation and a prerequisite for effectively directing new research efforts based on prior available information. To achieve this, research results have to be made publicly available and reusable. As computers are now the main instrument of any scientist, all research data including those in publications should be disseminated in computer-readable format and following the FAIR principles. LOTUS employs Wikidata as the repository for the referenced structure-organism pairs as this enables the documented research data to be integrated with a pre-existing vast body of chemical and biological knowledge.

The dynamic nature of the Wikidata platform fosters the continuous curation of deposited data through its user communities. Independence from institutional funding represents another major advantage of Wikidata. Wikidata knowledge base and the option to use elaborated SPARQL queries allow the exploration of the dataset from a sheer unlimited number of angles. The openness of Wikidata also offers unique 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 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 and can discourage end-users. Furthermore, traditional ways to query NP electronic NP resources such as structural or spectral searches are currently not within the scope of Wikidata.

Using the pre-existing COCONUT template, LNPN hosting allows the user to perform structural searches in a more classical way, e.g., by drawing a molecule. It addresses the shortcomings of the current lack of structural searches in Wikidata for the community. Future versions of LOTUS and COCONUT are envisioned to be augmented by predicted MS spectra and are expected to be hosted at the naturalproducts.net portal to allow mass, fragment, 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 always evolving field, all the taxon identifiers from all accepted taxonomic DB for a given taxon name were kept. This implies that for a given name, multiple taxonomies from different taxonomic DB are allowed. 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 this development. OTL also benefits from regular expert curation and new data. As the taxonomic identifier property for this DB did not exist in Wikidata, its creation was requested and obtained (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.

Data Interaction from the User Point-of-view

With the data being available in multiple formats, the possibilities to interact with the LOTUS data are numerous. Some basic and more advanced examples on how to retrieve, add and edit LOTUS data are provided hereafter.

Data Retrieval

LOTUS data can be queried and retrieved either on Wikidata directly or on LNPN, both of these options presenting unique advantages. Wikidata offers modularity at the cost of potentially complex access to the data. LNPN offers a Graphical User Interface (GUI) with chemical structure drawing possibility, easy structural or biological filtering, and advanced chemical descriptors, but with a more rigid structure. A frozen (2021-02-23) version of LOTUS data is available at https://osf.io/zupqj/.

Hereafter, finer approaches to directly interrogate the up-to-date LOTUS data both in Wikidata and LNPN are detailed.

Wikidata

The simplest way to search for NP occurrence information in Wikidata is by typing directly the name of a chemical structure in the "Search Wikidata" field. For example by typing "erysodine" the user lands on the Wikidata page of this compound (Q27265641). Scrolling down to the "found in taxon" statement gives a view of the biological organisms reported to contain this chemical compound. Under each taxon name, clicking on the reference link will then display the scientific publication documenting the occurrence.

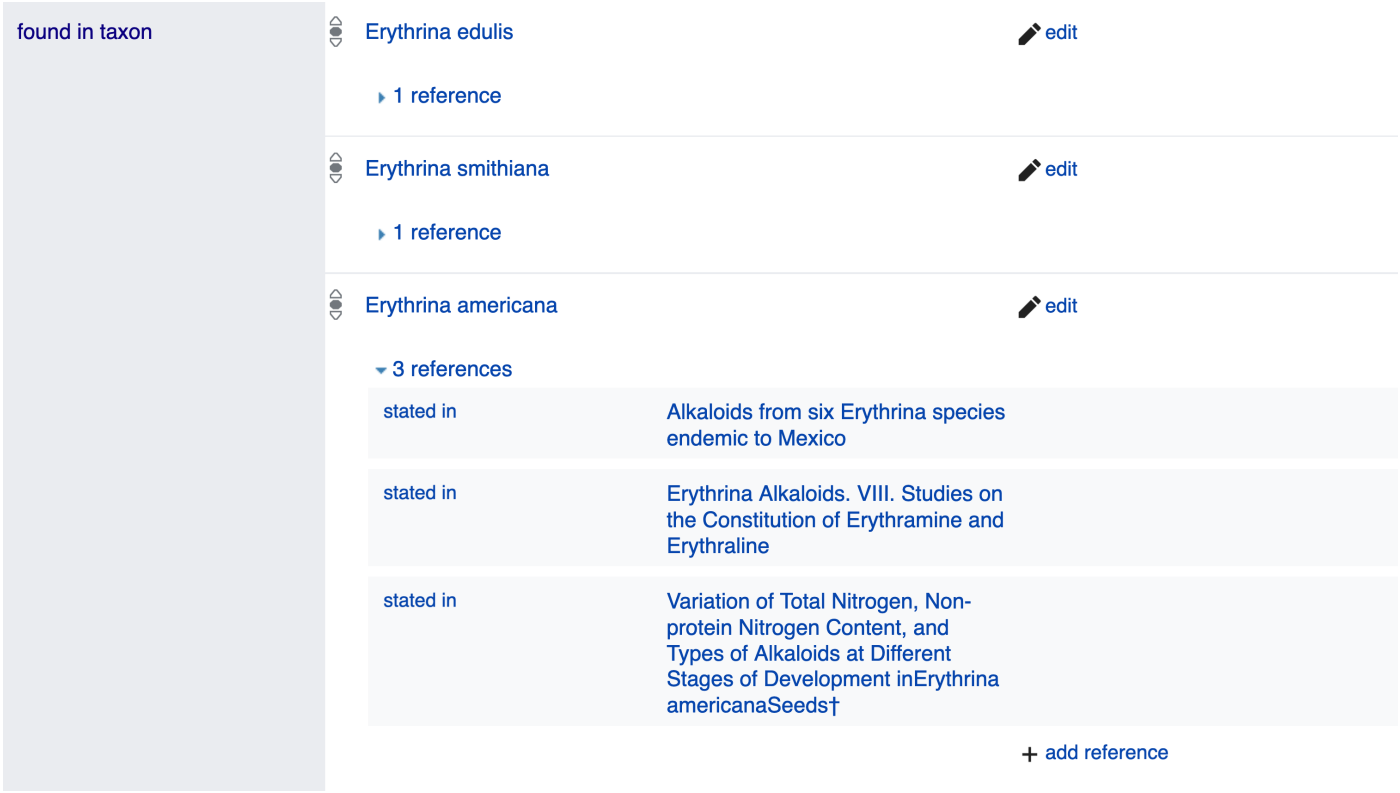


Figure 3: Illustration of the "found in taxon" statement section on the Wikidata page of erysodine Q27265641 showing a selection of containing taxa and the references documenting these occurrences.

For more elaborated queries, the most efficient way is to write SPARQL queries using the Wikidata Query Service or a direct connection to the SPARQL endpoint. Below are some examples, from simple to more elaborated queries demonstrating what can be done using this approach. The full text queries are also available with some explanations in SI-3.

Table 2: Example of a given referenced structure-organism pair before and after curation

Question	Wikidata SPARQL query
What are the compounds present in Mouse-ear cress (*Arabidopsis thaliana*)?	https://w.wiki/32y8
Which organisms are known to contain compounds sharing the planar structure of β-sitosterol?	https://w.wiki/334q
Which organisms are known to contain stereoisomers of β-sitosterol?	https://w.wiki/334s
Which pigments are found in which taxa, according to which reference?	https://w.wiki/38Rt
What are examples of organisms where compounds were reported to be produced by an organism sharing the same parent taxon, but not the organism itself?	https://w.wiki/3359

Question	Wikidata SPARQL query
Which *Zephyranthes* species lack compounds known from at least two sister species?	https://w.wiki/335x
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/32Qb
Which compounds are found in a biological organism, according to which references?	https://w.wiki/335C
Which organisms contain indolic scaffold? Results are grouped by the parent taxon and ordered.	https://w.wiki/32KZ
How many structure-organism pairs have been referenced by these authors? (Here, two senior natural products chemists and co-authors of this paper are compared to the late Ferdinand Bohlmann).	https://w.wiki/32$m

The queries presented in Table 2 are only selected examples and many other ways to interrogate LOTUS data can be formulated. A combination of these 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 compounds are found in a biological organism, according to which references?" - retrieves all chemical compounds (Q11173) or groups of stereoisomers (Q59199015) with found in taxon (P703) statements supported by a bibliographic reference (Q10358455): https://w.wiki/335C. Data can then be exported in various formats, such as classical tabular formats, json or html tables. At the time of publication, it returned 798,853 entries. A frozen result of the query is available at https://osf.io/xgyhm/. 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 reported structures in a given organism (e.g., structures found in Citrus aurantium (Q61127949) https://w.wiki/sFp). Alternatively, all organisms containing a given chemical structure can be queried (e.g., here all organisms in which β-sitosterol (Q121802) was reported https://w.wiki/dFz). For programmatic access, the WikidataLotusExporter repository also allows retrieval in RDF format and as tsv tables. As previously mentioned, some typical queries of molecular electronic resources such as structure or similarity search are not directly available in Wikidata. The SPARQL language 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, their high-performance chemical DB cartridge for PostgreSQL for fingerprint-guided substructure and similarity search (Kratochvíl, 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 this project. An example of a query containing substructure search is given in Table 2 (https://w.wiki/32KZ). The endpoint indexes isomeric (P2017) and canonical (P233) SMILES code available in Wikidata. To ensure that data are kept up-to-date, SMILES codes are downloaded from Wikidata automatically daily. The endpoint allows users to run federated queries and thus proceed to structure-oriented searches on the LOTUS data hosted at Wikidata. For example, the following SPARQL query, https://w.wiki/32KZ, will return a list of all organisms producing compounds with an indolic scaffold. The list 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, simple queries can be achieved by typing in the molecule name (e.g. protopine) or pasting a SMILES, InChI, InChIKey string, or a Wikidata identifier. All compounds found in a given organism can be found by typing the organism name at the species or any higher taxa level (e.g. Tabernanthe iboga). The compound search by chemical class is also enabled

in the same way. 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). From LNPN the bulk data can be retrieved as an SDF or SMILES file, 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 programmatically via the API, which is also described in the documentation.

Data Addition

One strong advantage of LOTUS consists in the possibility given to users to contribute to the NP occurrences documentation effort by adding new data or editing existing data. As all of the LOTUS data is stored in the SSOT, it is also used to avoid reprocessing of previously treated elements. However, at the moment, the SSOT is not open for direct write access to the public to maintain its coherence and allow us to evolve the scheme. To add or modify data in LOTUS, the users can employ the following approaches.

Sources

The LOTUS process will regularly re-import both the current sources and new ones. New and edited information from these electronic NP resources will be checked against the SSOT and if not present or different, it will follow the curation pipeline and will be further stored into the SSOT. Any researcher can, thus, contribute to these external electronic NP resources as means of providing new data for LOTUS, keeping in mind the delay between data addition and subsequent inclusion into LOTUS.

Wikidata

The currently favored approach to add new data to LOTUS is to edit Wikidata directly. This data will then be automatically imported into the SSOT DB. 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 to have a Wikidata account and to follow the general object editing guidelines. Regarding the addition of NP-centered objects (i.e., documented structure-organisms pairs) refer to the WikiProject Chemistry/Natural products group page. A tutorial for the manual creation and upload to Wikidata of a documented structure-organism pair is available in SI-3. 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

At the end of the previously described initial curation process, more than 500,000 referenced structure-organism pairs were validated for uploading on Wikidata. To automate the Wikidata upload 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 batch upload of properly curated and referenced structure-organism pairs to Wikidata. ScriptsPrograms for data addition to Wikidata can be found in the repository

WikidataLotusImporter. The following Xtools page offers an overview of the latest activity performed by our NPimporterBot, using those programs.

Data Edition

Even if correct at a given time point, scientific advances can invalidate or correct previously discovered 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. They often do thousands of editions without any need for human input. This helps to reduce automatically the amount of incorrect data that would otherwise require manual editing. However, manual curation by human experts remains the highest standard. Valuing this, interested users are invited to follow the manual curation tutorial in SI-4. The Scholia platform currently does not offer a user-friendly edition interface for scientific references. The adaptation of such a framework to edit the documented structure-pairs in the LOTUS project could facilitate future expert curation, including manual efforts that cannot be replaced by automated scripts or AI efforts.

Data Interpretation

To illustrate the nature and dimensions of the LOTUS dataset some selected examples of data interpretation are shown. First, 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). Then individual electronic NP resources participation to LOTUS data is resumed using the upset plot depiction, which allows the visualization of dataset intersection (Figure 5). In these two previous examples, the cases of β-sitosterol for the chemical structure and of Arabidopsis thaliana for the biological organism, are explored to provide well-documented entries to the reader. Finally, a biologically-interpreted chemical tree and a chemically-interpreted biological tree are presented (Figure 6 and Figure 7). They illustrate the overall chemical and biological coverage of LOTUS by linking family-specific classes of chemical structures to their taxonomic position. Figure 4, Figure 6, and Figure 7 were generated using the frozen data (as of 2021-02-23) available for download at the following link: https://osf.io/hgjdb/. Figure 5 required a dataset containing information from the commercial DNP and the complete data used for its generation is therefore unfortunately not available for public distribution. All scripts used for the generation of the figures (including SI-5) are available in the lotusProcessor repository in the src/4_visualizing directory for reproducibility purposes.

Organisms per Structure and Structure per Organisms Distribution

As depicted in Figure 4, three organisms on average are reported per chemical structure and eleven structures per organism. Half of the structures are reported in only one organism and five structures or fewer are reported in half of the organisms. Metabolomics studies suggest that these numbers are clearly underrated (Noteborn et al., 2000; Wang et al., 2019) and these numbers suggest that a better reporting of the metabolites during a phytochemical investigation could greatly improve coverage. This incomplete coverage may partially be explained by the fact that, usually, only newly described or bioactive structures are accepted for publication in classical NP research journals.

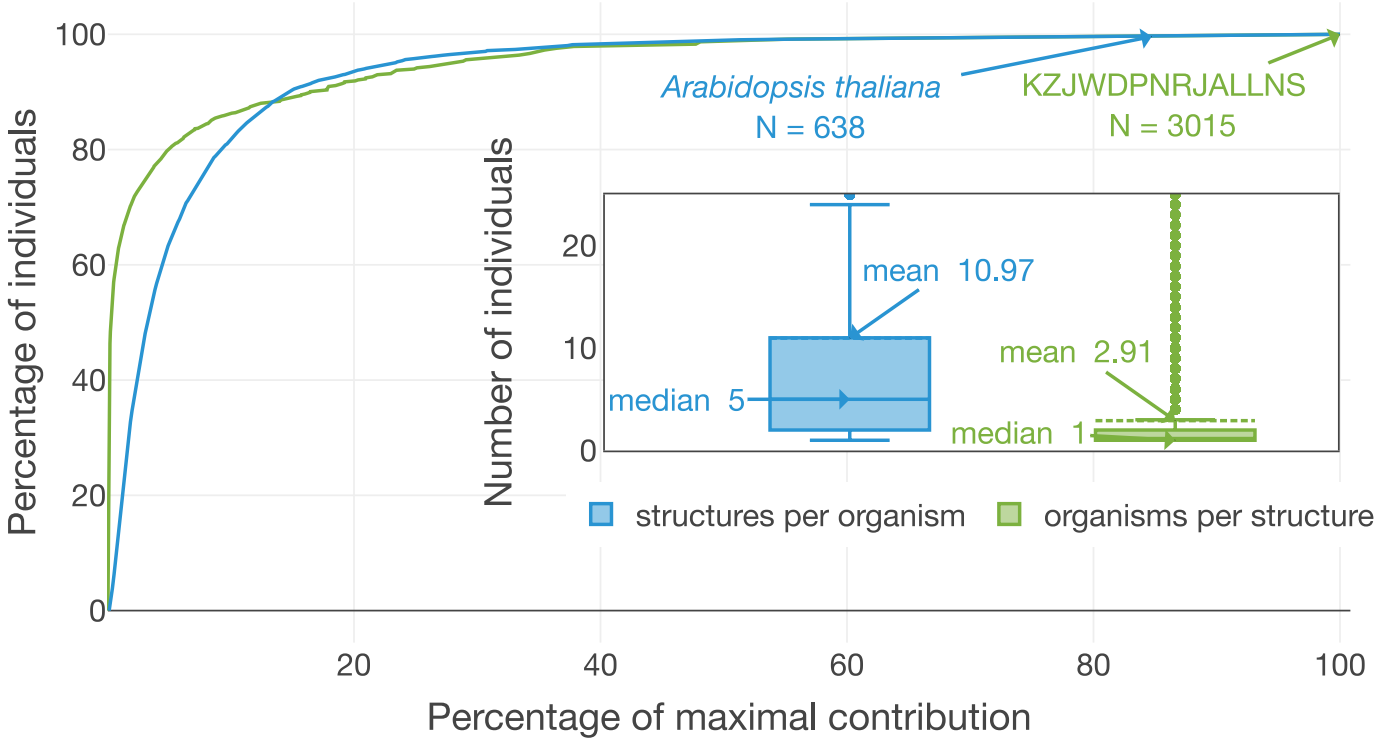


Figure 4: Distribution of "structures found in organisms" and"organisms containing structures". Number of organisms linked to the 2D structure of β-sitosterol (KZJWDPNRJALLNS) and the chemical diversity of Arabidopsis thaliana are highlighted as two notable examples.

Individual Electronic NP Resources Contribution to LOTUS

The added value of assembling multiple NP electronic NP resources in LOTUS is illustrated in Figure 5, showing the individual electronic NP resources contribution to the ensemble of chemical structures found in Arabidopsis thaliana ("Mouse-ear cress"; Q147096) (A) and to the ensemble of taxa containing the two-dimensional structure corresponding to β-sitosterol (Q121802) and (Q63409374) (B), a compound of ubiquitous occurrence in higher plants.

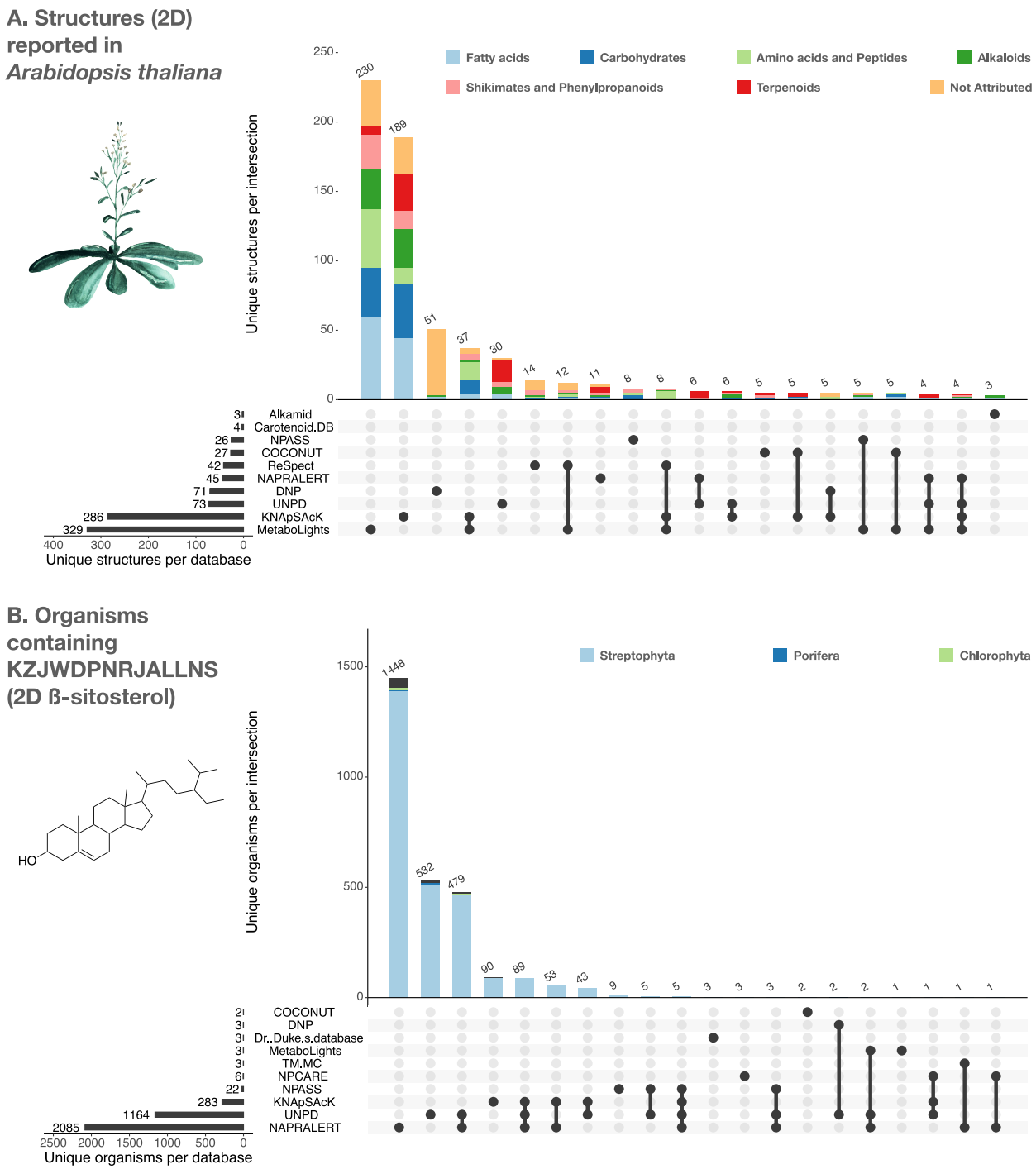


Figure 5: Upset plots of the individual electronic NP resource contribution to 2D structures found in Arabidopsis thaliana (A) and organisms containing the 2D structure of β-sitosterol (KZJWDPNRJALNS) (B). Upset plots are evolved Venn diagrams, allowing to represent intersections between multiple sets. 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 two sets. The vertical bars indicate the number of entries in the intersection. For example, 479 organisms containing the structure of β-sitosterol are present in both UNPD and NAPRALERT, which in turn, respectively report 1164 and 2085 organisms containing the structure of β-sitosterol.

Figure 5 A. shows that the chemical pathways distribution (according to NPClassifier (Kim et al., n.d.) across electronic NP resources is not conserved. Note that being specially tailored for NP, NPClassifier was preferred over ClassyFire (Feunang et al., 2016) but both chemical taxonomies are available as metadata in the frozen LOTUS export (https://osf.io/hgjdb/) 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. This upset plot indicates the poor overlap of preexisting electronic NP resources and the added value of an aggregated dataset. This is also illustrated in Figure 5 B., where the number of organisms for which the 2D structure of β-

sitosterol (KZJWDPNRJALLNS) has been reported for each intersection is shown. NAPRALERT has by far the highest number of entries (2085 in total), while other electronic NP resource complement this well (UNPD, for example, has 532 organisms where β-sitosterol is reported that are not overlapping with the ones reported in NAPRALERT). Interestingly, 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.

Biologically-interpreted Chemical Tree

LOTUS chemical diversity can be visualized via a TMAP (Figure 6), a visualization method allowing the structural organization of very large chemical datasets as a two-dimensional tree [Probst et al 2020 TMAP]. Using Faerun, an interactive HTML file is generated to display metadata and molecule structures (using the Java-Script library SmilesDrawer) [Probst et al 2018 FUN + Probst et al 2018 SmilesDrawer]. It is to be noticed that planar structures for all compounds (with MAP4 encoding) were used for the TMAP generation since LOTUS contains a mix of planar and 3D structures [Capecchi et al 2020]. As a result, the same planar structure may be present multiple times in the TMAP if this structure was reported with different stereochemistry and/or as a planar structure. Since structures are organized in the tree according to their molecular fingerprint, a good clustering of compounds of the same NPClassifier chemical class can be observed (Figure 6 A.). For clarity reasons, the eight most represented chemical classes of LOTUS plus the Quassinoids and Carotenoids (C40, β-β) classes are mapped, with an example of a quassinoid (yellow star) and a carotenoid (green star) with their corresponding location in the TMAP.

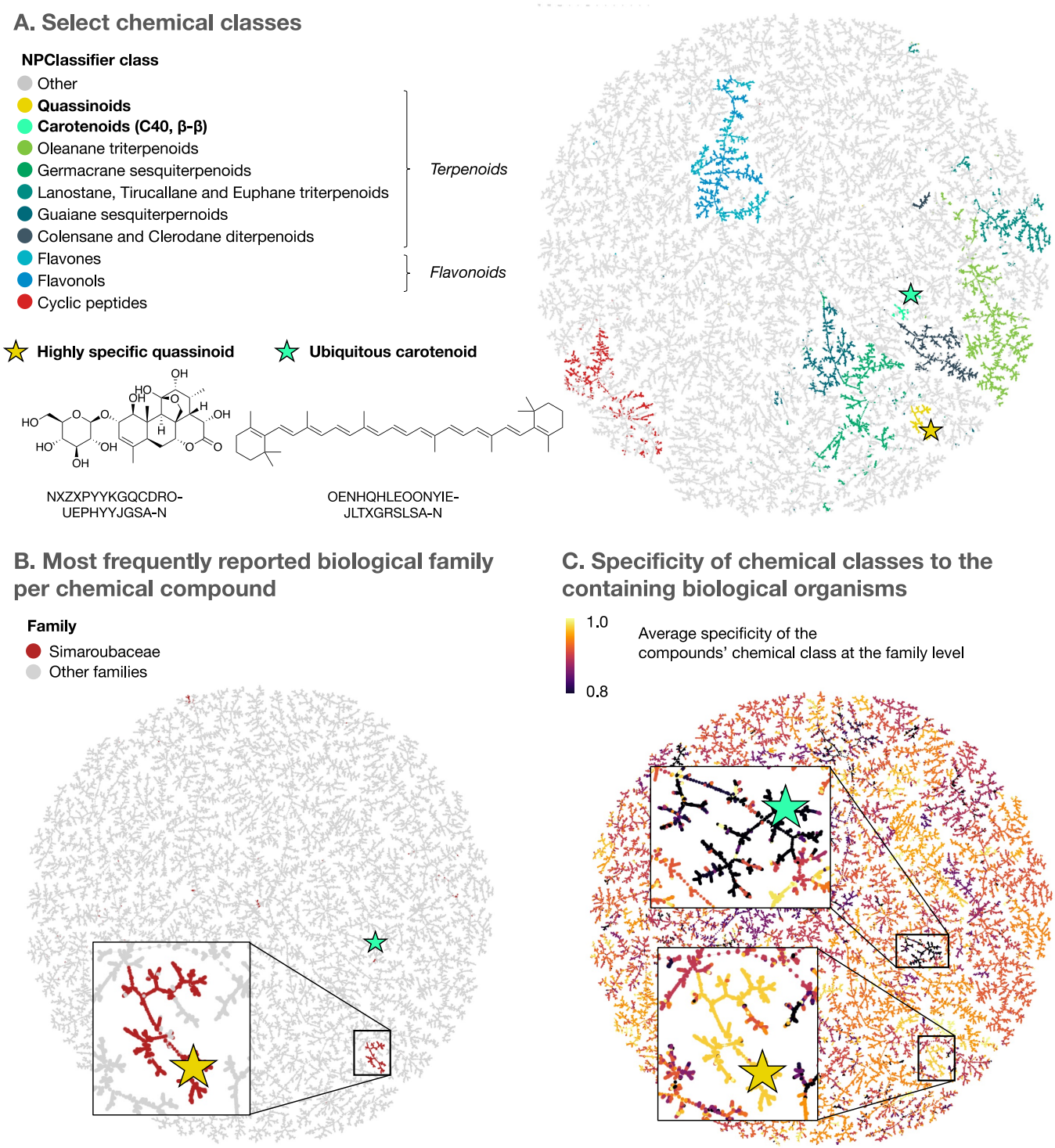


Figure 6: TMAP visualizations of the chemical diversity present in LOTUS where each dot corresponds to a chemical structure. Dots are colored according to the NPClassifier chemical class they belong to (A), the most frequently reported biological family (Simaroubaceae or other) (B), and to the specificity of chemical classes to the containing biological organisms (C). A quassinoid (yellow star) and a carotenoid (C40, β-β) (green star) are mapped as an example in all visualizations. The biological specificity at a given taxonomic level for a given structure is defined as the number of occurrences of the most frequently reported taxon divided by the number of biosources reported. For clarity in the plotting, the average specificity of the structure's NPClassifier class is displayed and values below 0.8 have been set to 0.8. An average chemical class biological specificity of 1 at the family level means that compounds in that NPClassifier class were only reported in one family. An interactive HTML visualization of the LOTUS TMAP is available at https://osf.io/kqa8b/.

To explore relationships between chemistry and biology, it is possible to map biological information such as the most reported biological family (Figure 6 B.) and the chemical class biological specificity (Figure 6 C.) on the TMAP. These visualizations allow to highlight chemical classes specific to a given taxon, such as the Quassinoids in the Simaroubaceae family example. It is striking to see how well in that case the chemical class (Quassinoids) and the most reported biological family (Simaroubaceae)

overlap, highlighting a very specific chemical class (specificity of 0.97 at the biological family level). On the other hand, it is also possible to identify classes that are widely spread among living organisms, such as the Carotenoids (C40, β-β) class (specificity of 0.79 at the biological family level).

Chemically-interpreted Biological Tree

A summary of the biological and chemical diversity covered by LOTUS is illustrated in Figure 7. To limit biases due to underreporting while keeping a reasonable display size, only families with at least 50 reported compounds were kept for this illustration. 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), calculated as the square of the number of structures reported in the chemical class within the given family, over the product of the number of reported structures in the chemical class and the number of reported structures in the biological family.

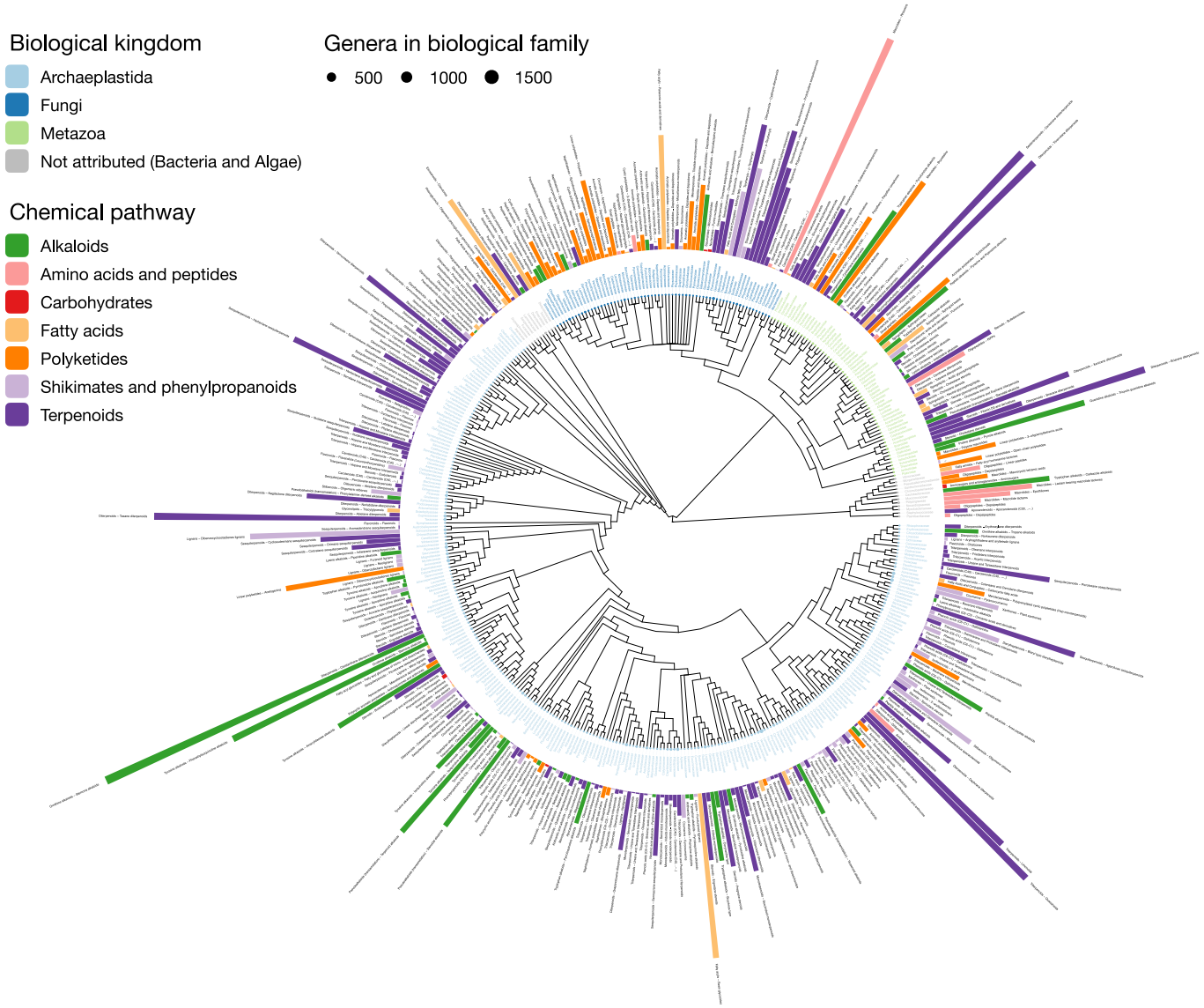


Figure 7: Chemical and biological diversity present in LOTUS. The tree corresponds to the biological taxonomy, with the kingdom as label color. The size of the leaves node corresponds to the number of genera reported in the family. The outer bars correspond to the most specific chemical class found in 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 family, over the product of the number of reported structures with the number of reported structures in the biological family. The bar's color corresponds to the chemical pathway (NPClassifier classification system) of the most specific chemical class.

In Figure 7, it is possible to spot highly specific compound classes such as trinervitane terpenoids in the Termitidae, rhizoxin macrolides in the Rhizopodaceae, or typical quassinoids and limonoids from the Simaroubaceae and Meliaceae, respectively. More generic tendencies can also 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 the β-sitosterol (KZJWDPNRJALLNS-VJSFXXLFSA-N) within the overall biological tree, SI-5 plots its presence/absence vs 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 were distributed across all kingdoms. As illustrated, the possibility of data interrogation at multiple precision levels (from fully defined 3D structures to broader chemical classes) is of great interest, e.g., for taxonomic and evolution studies.

As shown recently in the context of spectral annotation (Dührkop et al., 2020), lowering the precision level of the annotation allows a broader coverage together with greater confidence. Genetic studies investigating the pathways involved and the organisms carrying the responsible biosynthetic genes responsible 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.

Conclusion & Perspectives

At its current development stage, and despite its diligent assembly, data compiled in LOTUS remains imperfect and, by its nature, at least partially biased. As discussed above, in the context of bioactive NP research and due to global editorial practices, publications tend to emphasize new compounds for which interesting bioactivity has been measured. In contrast, ubiquitous compounds are poorly documented. For the time being, until more rigorous data becomes available, this provides a partial yet relatively comprehensive view on the actual metabolome of all studied organisms, including the more thoroughly studied model organisms. The comprehensive editing opportunities provided within LOTUS and by the associated Wikidata distribution platform open new opportunities for collaborative community engagement, which the authors believe is the most efficient means of correcting any remaining bias.

Moreover, the dissemination of referenced structure-organism pairs through Wikidata, together with the harmonized data format enables query of an increasingly globalized NP research knowledge from a radically novel perspective. All researchers involved in NP research and special aspects of metabolism can benefit from this opportunity, whether the interest is in ecology and evolution, chemical ecology, drug discovery, biosynthesis pathway elucidation, chemotaxonomy, or similar research fields.

The introduction of LOTUS also provides a new opportunity to advance the FAIR guiding principles for scientific data management and stewardship established in 2016 (Wilkinson et al., 2016). Researchers widely acknowledge the availability of data that has been unavailable, and at least in part even been discouraged, e.g., due to page limitations and mechanisms involved in the calculation of journal impact factors) by the classical print and static PDF publication formats. 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 NMR spectroscopic data by a large consortium of NP researchers (McAlpine et al., 2019). While first steps in this direction are currently being taken (for example: Harvard Dataverse repositories for individual research groups [Pauli group at UIC: dataverse.harvard.edu/dataverse/gfpuic] and projects [dataverse.harvard.edu/dataverse/cenapt]; an NIH-funded raw NMR data repository [www.np-mrd.org/]), one missing element in this multifaceted development is the adequate incentivization of community efforts; to generate the vital flow of contributed data, the effort associated with preparing and submitting content to open repositories should be better acknowledged in academia, government, regulatory, and industrial environments. Moreover, data reuse should also be acknowledged (Cousijn et al., 2018, 2019; Pierce et al., 2019).

The possibilities for expansion and future applications of Wikidata-hosted LOTUS data are significant. For example, properly formatted spectral data, e.g. data obtained by mass spectrometry or nuclear magnetic resonance, can be linked to the Wikidata entries for the respective chemical compounds. MassBank (Horai et al., 2010) and SPLASH (Wohlgemuth et al., 2016) identifiers are already reported in Wikidata, and this existing information can be used to report MassBank records for single organisms such as Arabidopsis thaliana compounds (https://w.wiki/335H). 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 previously in different content. This will open exciting new perspectives in the area of structural dereplication and NP annotation. The authors have previously demonstrated that taxonomically-informed metabolite annotation can critically improve the NP annotation process (Rutz et al., 2019). Alternative approaches linking structural annotation to biological organisms have also shown tremendous improvements (Hoffmann et al., 2021). The availability of LOTUS as an open repository that links chemical objects to both their spectral information and biological occurrences will facilitate such applications significantly.

As shown in SI-5, observing the chemical and biological diversity at various granularity levels can offer new insights. Regarding the chemical objects involved, it will be important to implement chemical taxonomy 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. The fact that LOTUS is disseminated on Wikidata should facilitate its integration within biological pathway DB such as WikiPathways and contribute to this complex task (Martens et al., 2020; Slenter et al., 2018). In the field of ecology, 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 the 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 “chemodiverse hotspots” (Defossez et al., 2021).

Other NP-related information is of great interest but unfortunately remains poorly formatted. For example, traditional medicine (including ethnomedicine and ethnobotany) 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. To this end, LOTUS represents a resource for new concepts by which such information could be valued and conserved in the digital era, as it provides a blueprint for appropriate formatting and sharing of such data (Allard et al., 2018; Cordell, 2017b, 2017a).

The various facets discussed above represent future developments that the combination of LOTUS and the Wikidata knowledge base can feasibly accommodate. 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 spaces and graphs and fully exploit their potential (Kratochvíl, 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 - efficiently complementing human interpretation, validation, and thinking.

Overall, the LOTUS project is intended to give the pharmacognosy et al. research community access to a greater quantity and better quality of data and, thereby, ultimately pave the way towards a global open electronic NP resource. This reflects the authors’ belief that the integration of NP research results requires a truly open and FAIR knowledge-based in the form of an electronic NP resource structure, and this approach will fuel a virtuous (rather than a vicious) cycle of research habits aiming at a better understanding of Life and its chemistry.

Methods

Data Curation

Gathering

Before their inclusion, the overall quality of 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 29.1) 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 (Fig. 2). 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 written to automatically retrieve entries can be found in the lotusProcessor 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 license 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. 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 (for example, “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 over 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 (2020.03.3) 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 (Feunang et al., 2016) and NPClassifier API.

From the 269,436 initial InChI, 233,652 (87%) sanitized structures were obtained, of which 176,235 (75%) had complete stereochemistry defined. 197,392 (73%) were uploaded to Wikidata. From the 244,724 initial SMILES, 205,244 (84%) sanitized structures were obtained, of which 8,542 (4%) had complete stereochemistry defined. 173,590 (71%) were uploaded to Wikidata. From the 49,315 initial chemical names, 27,780 (56%) sanitized structures were obtained, of which 2,324 (8%) had complete stereochemistry defined. 23,098 (47%) were uploaded to Wikidata. In total, 164,308 structures with fully defined stereochemistry were uploaded as “chemical compounds” (Q11173), and 106,043 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 Verify (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 7070 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 86,076 initial “clean” organism fields, 43,421 (50%) canonical names were obtained, of which 31,965 (37%) were uploaded to Wikidata. From the 294 initial “dirty” organism fields, 241 (82%) canonical names were obtained, of which 198 (67%) were uploaded to Wikidata.

References

The Rcrossref package 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. 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 36,692 initial “original” references, 21,928 (60%) references with sufficient quality were obtained, of which 15,674 (71%) had the organism name in their title. 9,882 (27%) were uploaded to Wikidata. From the 21,306 initial “pubmed” references, 9,015 (42%) references with sufficient quality were obtained, of which 5,695 (63%) had the organism name in their title. 3,013 (14%) were uploaded to Wikidata. From the 35,348 initial “doi” references, 19,682 (56%) references with sufficient quality were obtained, of which 15,522 (79%) had the organism name in their title. 11,847 (34%) were

uploaded to Wikidata. From the 29,584 initial “title” references, 17,410 (59%) references with sufficient quality were obtained, of which 12,732 (73%) had the organism name in their title. 9,638 (33%) were uploaded to Wikidata. From the 11,322 initial “split” references, 5,856 (52%) references with sufficient quality were obtained, of which 3,221 (55%) had the organism name in their title. 2,255 (20%) were uploaded to Wikidata. From the 3,310 initial “publishingDetails” references, 119 (4%) references with sufficient quality were obtained, of which 59 (50%) had the organism name in their title. 16 (0%) 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 over 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, over 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 1.

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

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 a 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 (2021-02-23) version of LOTUS data is also available at https://osf.io/zupqj/. WikidataLotusExporter 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 WikidataLotusImporter 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 documented 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 WikidataLotusImporter 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 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/lotusProcessor. 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/wikidataLotusImporter. 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/wikidataLotusExporter. 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 version used was 4.0.4 (2021-02-15) – "Lost Library Book" (R Core Team, 2020) 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 (2020.09.2) ("RDKit: Open-source cheminformatics," n.d.), 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.0.11.1, GNVerify v.0.1.0, 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 (datasets): https://osf.io/pmgux/.

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. 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 DB used in this work.

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	2	3	3		2	3	3		3		3	3	3	3				1	**32**
CS				1				1						1	1				**4**
DM														1				2	**3**
EW														1				2	**3**
GFP				1				2						2		1			**6**
JB	2	2	3		2	3	2		3	2	2			2		1		2	**26**
JGa														1			2		**3**
JGr														1		1			**2**
J-LW				1				2						2					**5**
JV				1				1									2		**4**
MS									2					2	3				**7**
P-MA	2	2	3	1	2	3	2		2	3	2		3	3				1	**29**
RP																		1	**1**
RS														1				1	**2**
Total	**6**	**7**	**9**	**5**	**6**	**9**	**7**	**6**	**10**	**5**	**7**	**4**	**6**	**21**	**4**	**3**	**4**	**10**	

Supporting Information

SI 1 - Data Sources List

Table SI-1: Data Sources List

database	type	initial retrieved unique entries	cleaned documented structure-organism pairs	pairs validated for wikidata export	website	article	retrieval	license	contact	varia
afrotryp	open	312	135	28	NA	afrotryp_article	afrotryp_download	afrotryp_license	Fidele Ntie-Kang or Ngozi Justina Nwodo	NA
alkamid	open	4434	2582	2092	alkamid_website	alkamid_article	alkamid_script	alkamid_license	Bart De Spiegeleer	NA
biofacquim	open	531	683	534	biofacquim_website (old version)	biofacquim_article_old biofacquim_article_new	biofacquim_download	biofacquim_license	José Medina-Franco	NA
biophytmol	open	546	632	356	biophytmol_website	biophytmol_article	biophytmol_script	biophytmol_license	Anshu Bhardwaj	website often down
carotenoiddb	open	2922	1204	639	carotenoiddb_website	carotenoiddb_article	carotenoiddb_script	carotenoiddb_license	yzjunko@gmail.com	NA
coconut	open	583623	345147	34437	coconut_website	coconut_article	coconut_download	coconut_license	Maria Sorokina	coconut_zenodo
cyanometdb	open	1930	1812	1754	NA	cyanometdb_article	cyanometdb_download	cyanometdb_license	elisabeth.janssen@eawag.ch	NA
datawarrior	open	589	1052	102	datawarrior_website	datawarrior_article	datawarrior_download	datawarrior_license	thomas.sander@idorsia.com	no real link to the dataset inside it
dianatdb	open	290	406	27	dianatdb_website	dianatdb_article	dianatdb_download	dianatdb_license	amadariaga@iquimica.unam.mx or kmtzm@unam.mx	NA
dnp	commercial	210832	258864	NA	dnp_website	NA	dnp_script	TODO	support@taylorfrancis.com	commercial
drduke	open	90675	9052	4266	drduke_website	NA	drduke_download	drduke_license	agref@usda.gov	NA
foodb	restricted	82415	376	NA	foodb_website	NA	foodb_download	foodb_license	jreid3@ualberta.ca (Jennifer)	NA
inflamnat	open	665	656	282	NA	inflamnat_article	inflamnat_download	inflamnat_license	xiaoweilie@ynu.edu.cn	NA

database	type	initial retrieved unique entries	cleaned documented structure-organism pairs	pairs validated for wikidata export	website	article	retrieval	license	contact	varia
knapsack	open	116284	142466	63094	knapsack_website	knapsack_article	knapsack_script	knapsack_license	skanaya@gtc.naist.jp	NA
metabolights	open	29864	29498	4920	metabolights_website	metabolights_article	metabolights_download	metabolights_license	???	NA
mibig	open	1310	1132	548	mibig_website	mibig_article	mibig_download	mibig_license	Tilmann Weber or Marnix Medema	NA
mitishamba	open	1073	1174	369	mitishamba_website	mitishamba_“article”	mitishamba_script	mitishamba_license	???	NA
nanpdb	open	5752	6452	5327	nanpdb_website	nanpdb_article	nanpdb_script	nanpdb_license	ntiekfidele@gmail.com stefan.guenther@pharmazie.uni-freiburg.de	NA
napralert	commercial	681401	380559	263586	napralert_website	napralert_article	TODO	napralert_license	napralert@uic.edu	NA
npass	open	290539	28991	21913	npass_website	npass_article	npass_download	npass_license	phacyz@nus.edu.sg jiangyy@sz.tsinghua.edu.cn jiaochen@163.com	NA
npatlas	open	29006	49861	44456	npatlas_website	npatlas_article	npatlas_download	npatlas_license	rlininget@sfu.ca	NA
npcare	open	7763	4665	2525	npcare_website	npcare_article	npcare_download	npcare_license	choihwanho@gmail.com	NA
npedia	open	82	99	23	npedia_website	npedia_article	npedia_script	npedia_license	hisyo@riken.jp npd@riken.jp	NA
nubbe	open	2189	2618	2463	nubbe_website	nubbe_article	MISSING_nubbe_script	nubbe_license	Vanderlan S. Bolzani	NA
pamdb	open	3061	3198	64	pamdb_website	pamdb_article	pamdb_download	pamdb_license	awilks@rx.umaryland.edu aoglesby@rx.umaryland.edu aoglesby@rx.umaryland.edu	NA
phenolexplorer	open	8968	11779	6755	phenolexplorer_website	phenolexplorer_article	phenolexplorer_download	phenolexplorer_license	scalberta@iarc.fr	NA
phytohub	open	2363	1425	52	phytohub_website	phytohub_“article”	phytohub_script	phytohub_license	claudine.manach@inra.fr	NA

database	type	initial retrieved unique entries	cleaned documented structure-organism pairs	pairs validated for wikidata export	website	article	retrieval	license	contact	varia
procardb	open	6606	9864	70	procardb_website	procardb_article	procardb_script	procardb_license	Anil Kumar Pinnaka Ashwani Kumar	NA
respect	open	2759	601	282	respect_website	respect_article	respect_download	respect_license	ksaito@psc.riken.jp	NA
sancdb	open	861	981	772	sancdb_website	sancdb_article	sancdb_script	sancdb_license	Özlem Tastan Bishop	NA
streptomedb	open	71638	37816	19037	streptomedb_website	streptomedb_article	streptomedb_download	streptomedb_license	stefan.guenther@pharmazie.uni-freiburg.de	NA
swmd	open	1075	1616	1368	swmd_website	swmd_article	swmd_script	swmd_license	Dicky.John@gmail.com	NA
tmdb	open	2116	841	14	tmdb_website	tmdb_article	tmdb_script	tmdb_license	Xiao-Chun Wan Guan-Hu Bao	currently down
tmmc	open	15033	5727	2658	tmmc_website	tmmc_article	tmmc_download	tmmc_license	Jeong-Ju Lee	NA
tppt	open	27182	28355	941	tppt_website	tppt_article	tppt_download	tppt_license	thomas.bucheli@agroscope.admin.ch	NA
unpd	open	340319	550980	353433	unpd_website	unpd_article	TODO	TODO	lirongc@pku.edu.cn xiaojxu@pku.edu.cn	NA
wakankensaku	open	367	49	41	wakankensaku_website	???	wakankensaku_script	TODO	???	NA

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**
Correctd total	-	-	-	-	-	**0.96**	**0.89**	**0.91**	-	-

SI 3 - Wikidata SPARQL Queries

TODO Comment them!

Query 1 - *Arabidopsis thaliana*

This query answers to the following question: `What are the compounds present in Mouse-ear cress (Arabidopsis thaliana)?` Link: https://w.wiki/32y8

```
SELECT DISTINCT ?compound ?compoundLabel WHERE {
  VALUES ?classes {
    wd:Q11173 # chemical compound
    wd:Q59199015 # group of stereoisomers
  }
  ?compound wdt:P31 ?classes.
  ?taxon wdt:P225 "Arabidopsis thaliana".
  {
    ?compound p:P703 ?stmt.
    ?stmt ps:P703 ?taxon;
      prov:wasDerivedFrom ?ref.
    ?ref pr:P248 ?art.
    ?art wdt:P356 ?art_doi.
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Query 2 - β-sitosterol planar structure

This query answers to the following question: `Which organisms are known to contain compounds sharing the planar structure of β-sitosterol?` Link: https://w.wiki/334q

```
SELECT DISTINCT ?taxon ?taxonname WHERE {
  VALUES ?classes {
    wd:Q11173 # chemical compound
    wd:Q59199015 # group of stereoisomers
  }
  ?compound wdt:P31 ?classes;
    wdt:P235 ?inchikey.
  FILTER(REGEX(?inchikey, "KZJWDPNRJALLNS", "i"))
  {
    ?compound p:P703 ?stmt.
    ?stmt ps:P703 ?taxon.
    {
      ?stmt prov:wasDerivedFrom ?ref.
      ?ref pr:P248 ?art.
      ?art wdt:P356 ?art_doi.
    }
  }
  ?taxon wdt:P225 ?taxonname .
}
```

Query 3 - β-sitosterol stereoisomers

This query answers to the following question: `Which organisms are known to contain compounds sharing the planar structure of β-sitosterol?` Link: https://w.wiki/334s

```
# taxa with chemical compounds related to (but different from) beta-sitosterol
SELECT ?compound ?compoundLabel ?InChIKey ?taxonname
WITH {
  SELECT ?compound ?InChIKey WHERE {
    wd:Q121802 wdt:P235 ?queryKey . # beta-sitosterol
    ?compound wdt:P235 ?InChIKey .
    FILTER ( ?InChIKey != ?queryKey )
    FILTER (regex(str(?InChIKey), concat("^", substr($queryKey,1,14), "-")))
  }
} AS %compounds
WHERE {
  INCLUDE %compounds
  ?compound wdt:P703/wdt:P225 ?taxonname .
  ?compound rdfs:label ?compoundLabel.
  FILTER(LANG(?compoundLabel) = "en").
}
ORDER BY ASC(?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/38Rt

```
# Pigments and the taxa they were found in, along with references
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 1000
```

Query 5 - Sister taxon compounds

This query answers to the following question: `What are examples of organisms where compounds were reported to be produced by an organism sharing the same parent taxon, but not the organism itself?` Link: https://w.wiki/3359

```
SELECT ?compound ?compoundLabel ?taxonname_with_compound ?taxonname_without_compound

?parent_taxon

WITH
{
  SELECT DISTINCT ?compound ?taxon_with_compound ?parent_taxon
  WHERE {
    VALUES ?classes {
      wd:Q11173
      wd:Q59199015
    }
    ?compound wdt:P31 ?classes;
              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 ?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 sister species?` Link: https://w.wiki/335x

```
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 {
    ?taxon_YES_1 wdt:P171 target: .
    ?taxon_YES_2 wdt:P171 target: .
    FILTER (?taxon_YES_2 != ?taxon_YES_1)
    ?compound wdt:P703 ?taxon_YES_1 .
    ?compound wdt:P703 ?taxon_YES_2
  }
} 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 .}
  ?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 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 found in.` Link: https://w.wiki/32Qb

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
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".
      ?antibiotic wdt:P233 ?smiles.
    }

    ?compound sachem:similarCompoundSearch [
      sachem:query ?smiles;
      sachem:cutoff "0.9"^^xsd:double ].
  }

  hint:Prior hint:runFirst true. # hint to evaluate the idsm service first

  ?compound wdt:P703 ?taxon.
  ?taxon wdt:P171 ?parent_taxon.
  OPTIONAL { ?parent_taxon wdt:P225 ?parent_taxon_name. }
}
GROUP BY ?parent_taxon ?parent_taxon_name
ORDER BY DESC (?count)
```

Query 8 - Triples

This query answers to the following question: `Which compounds are found in a biological organism, according to which references?` Link: https://w.wiki/335C

```
SELECT DISTINCT ?compound ?compound_inchi ?compound_inchikey ?taxon ?taxon_name ?reference ?reference_doi  WHERE
        {
  ?compound wdt:P235 ?compound_inchikey;
            p:P703 [
              ps:P703 ?taxon ;
                      prov:wasDerivedFrom / pr:P248 [
                        wdt:P356 ?reference_doi
                      ]
            ] .
  ?compound wdt:P234 ?compound_inchi.
  ?taxon wdt:P225 ?taxon_name.
  ?reference wdt:P356 ?reference_doi.
}
LIMIT 1000
```

Query 9 - Indolic scaffold

This query answers to the following question: `Which organisms contain indolic scaffold? Results are grouped by the parent taxon and ordered.` Link: https://w.wiki/32KZ

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
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(?compound) AS ?count) WHERE {
  SERVICE idsm:wikidata {
    ?compound sachem:substructureSearch
      [ sachem:query "NCCC1=CNC2=C1C=CC=C2" ]
  }

  hint:Prior hint:runFirst true. # hint to evaluate the idsm service first

  ?compound p:P703 ?statement.
  ?compound wdt:P235 ?inchikey.
  ?statement ps:P703 ?taxon.
  ?taxon wdt:P171 ?parent_taxon.
  OPTIONAL { ?taxon wdt:P171 ?parent_taxon. }
  OPTIONAL { ?taxon wdt:P225 ?taxon_name. }
  OPTIONAL { ?parent_taxon wdt:P225 ?parent_taxon_name. }
  OPTIONAL { ?taxon wdt:P846 ?taxon_id_gbif. }
  OPTIONAL { ?taxon wdt:P685 ?taxon_id_ncbi. }
}
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 these authors? (Here, two senior natural products chemists and co-authors of this paper are compared to the late Ferdinand Bohlmann).` Link: https://w.wiki/32$m

```
#defaultView:BarChart
SELECT ?authors_namesLabel (COUNT(DISTINCT(?compound)) AS ?count) 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
  }
  ?compound wdt:P31 ?classes. # instance of
  #?taxon wdt:P225 'Arabidopsis thaliana'
  {
    ?compound p:P703 ?stmt. # found in taxon
    ?stmt ps:P703 ?taxon. # found in taxon
    ?stmt prov:wasDerivedFrom ?ref.
    ?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.
  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 documented structure-organism pair to Wikidata

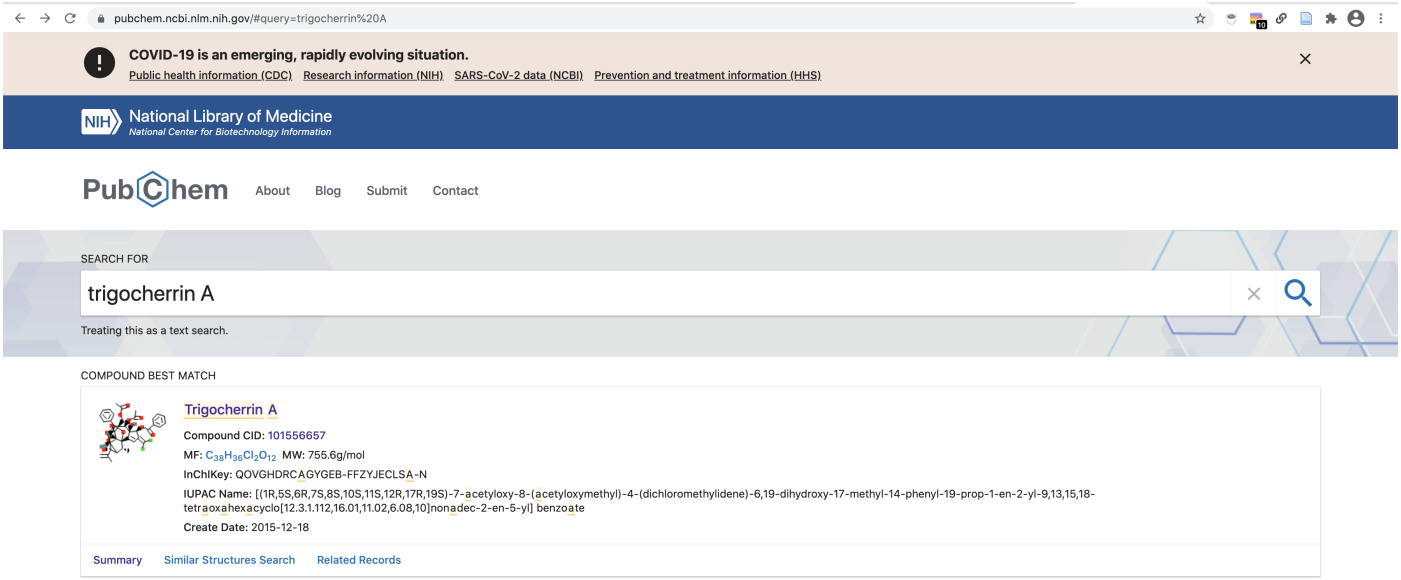
Select a documented structure-organism pair

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 documented structure-organism pair

Structure

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



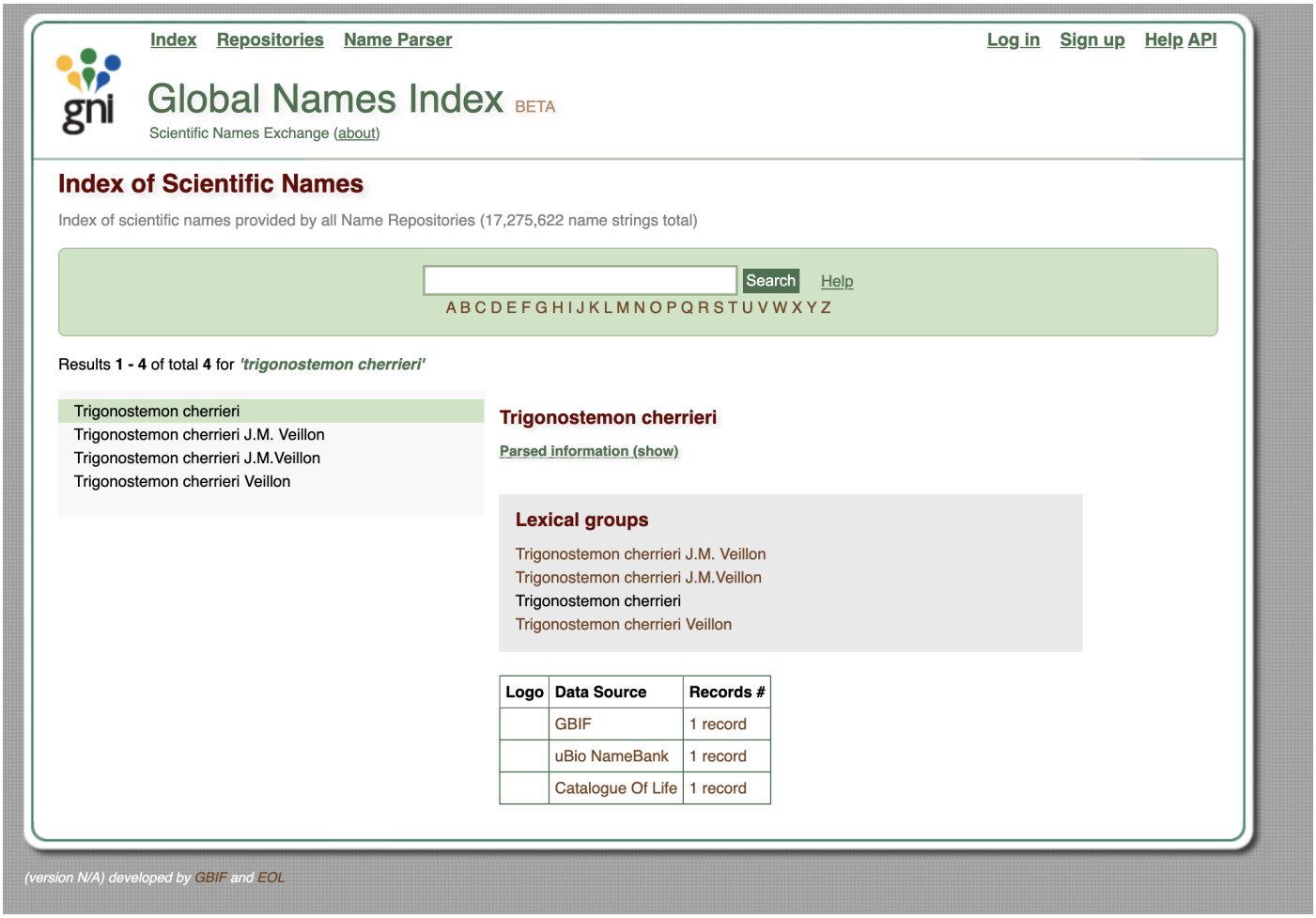
tutorial-image-01

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

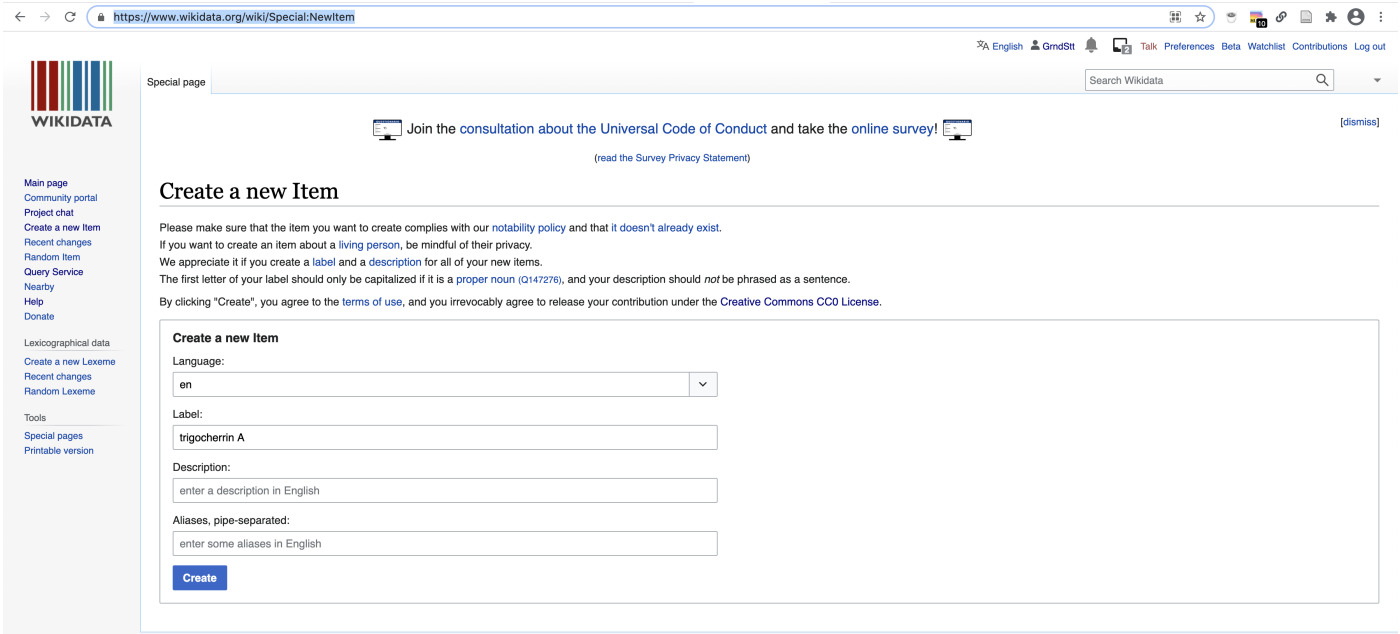
If your compound is already present on Wikidata, you can directly skip to the Add the biological source information section below.

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.

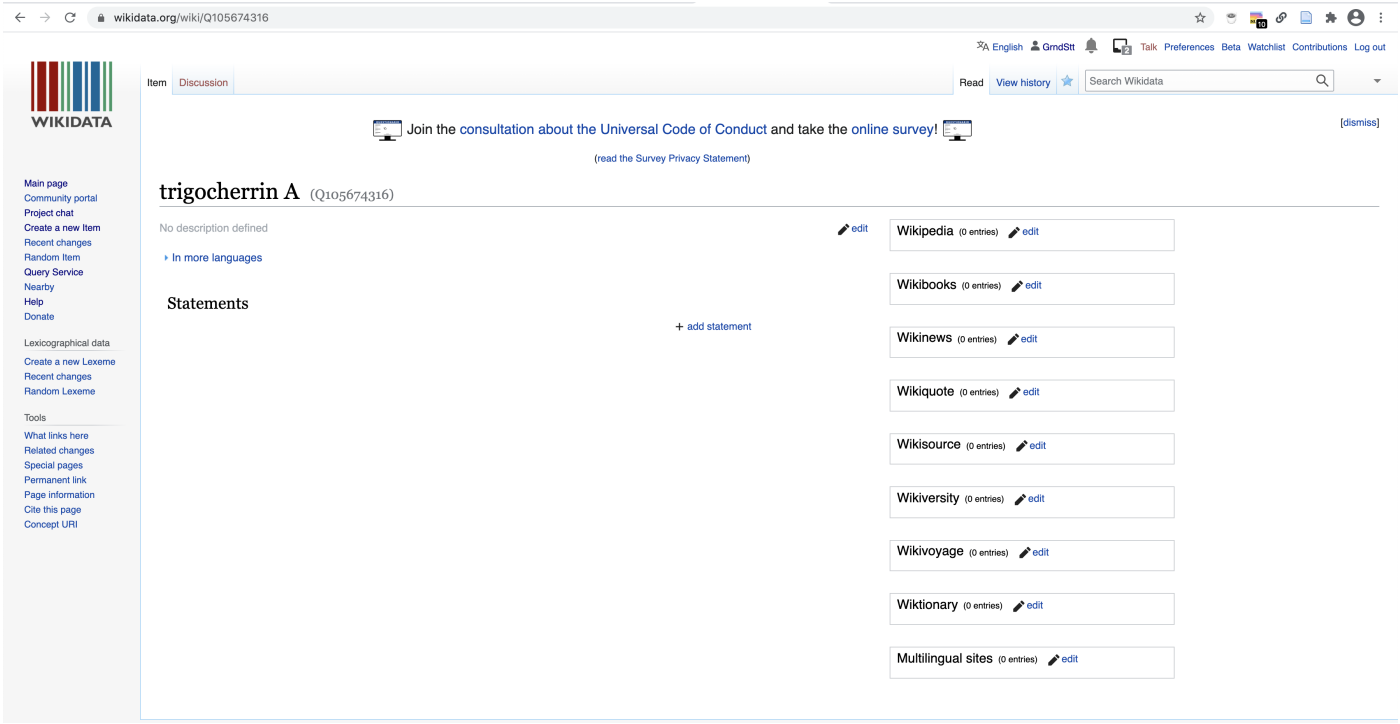
If you are unfamiliar with how Wikidata works, you can start by reading the Wikidata introduction page https://www.wikidata.org/wiki/Wikidata:Introduction and have a look at the Wikidata Tours page https://www.wikidata.org/wiki/Wikidata:Tours.

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

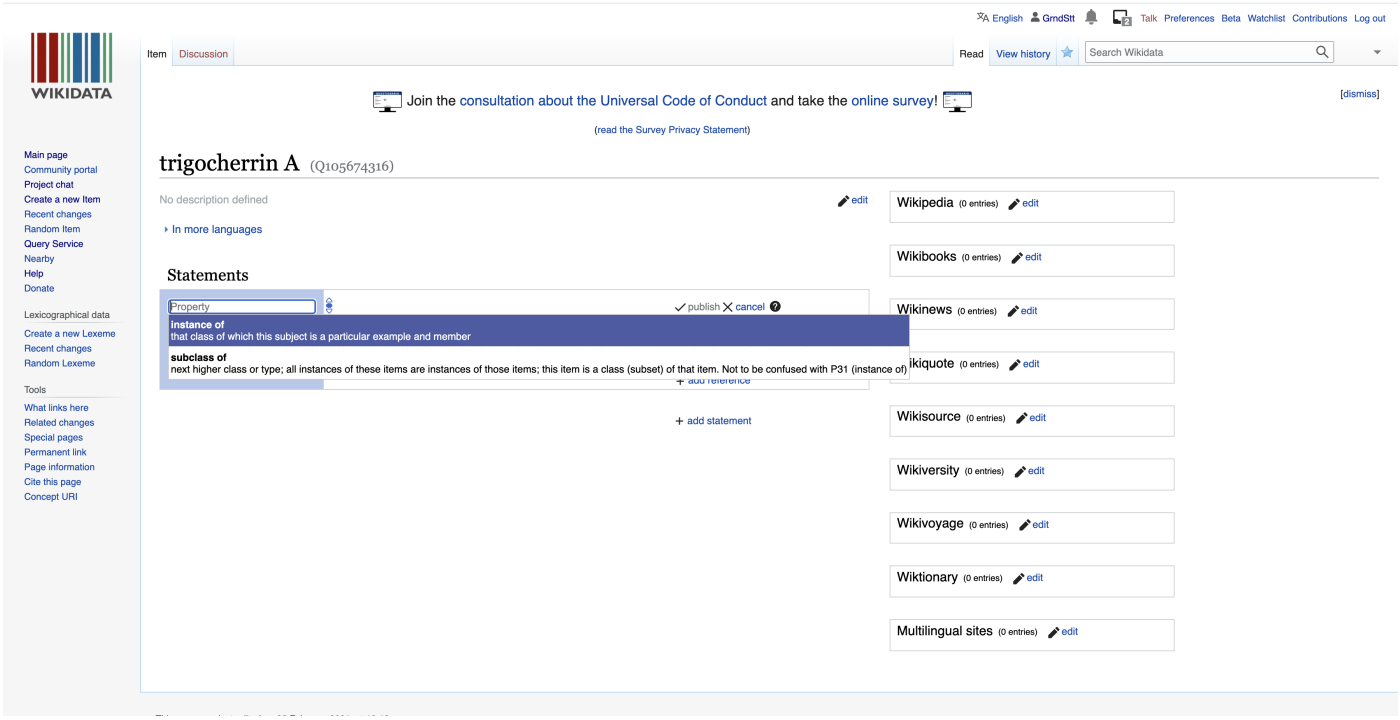
An empty page with a new Wikidata identifier is created



tutorial-image-04

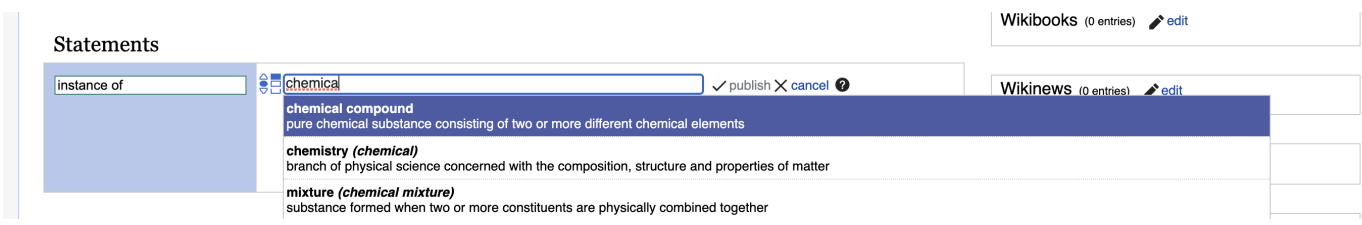
Add the chemical compound information

Create a new statement for `is an instance of`



tutorial-image-05

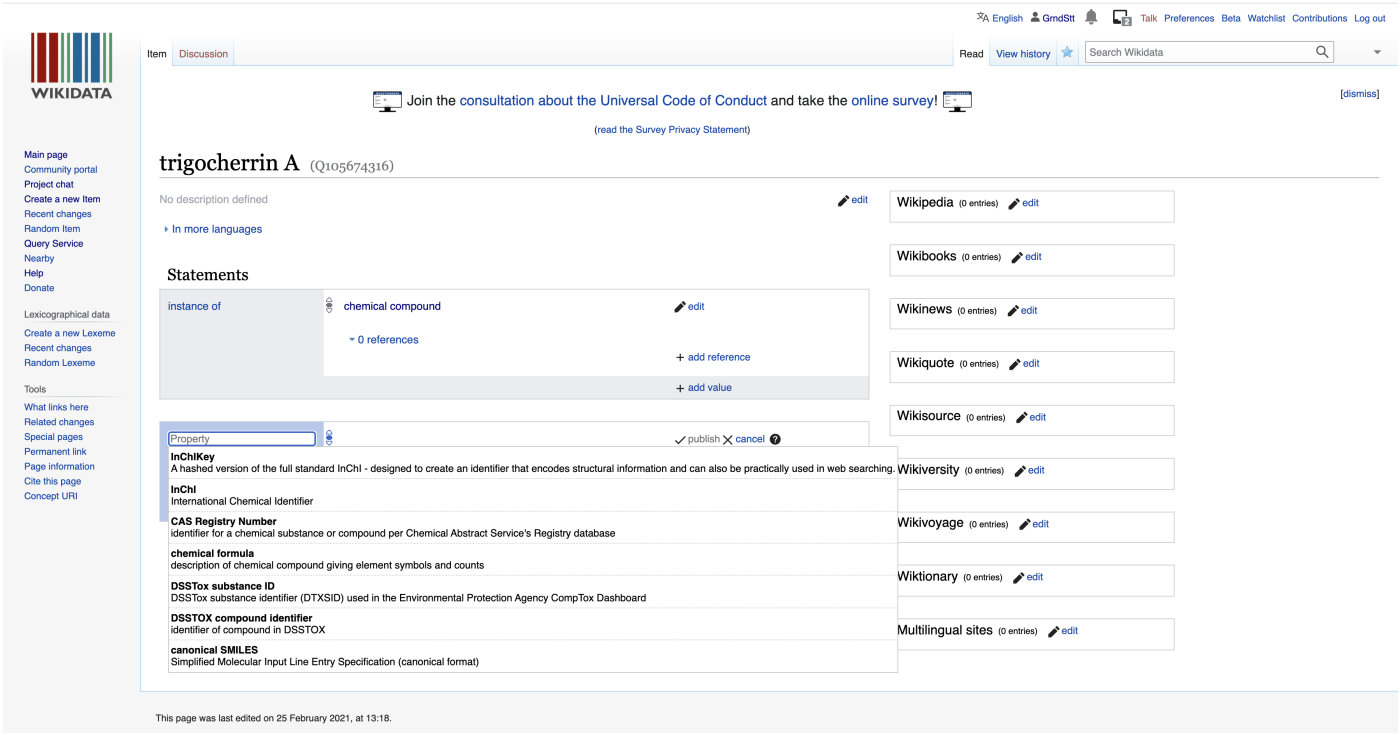
and select chemical compound (i.e. Q11173):



tutorial-image-06

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

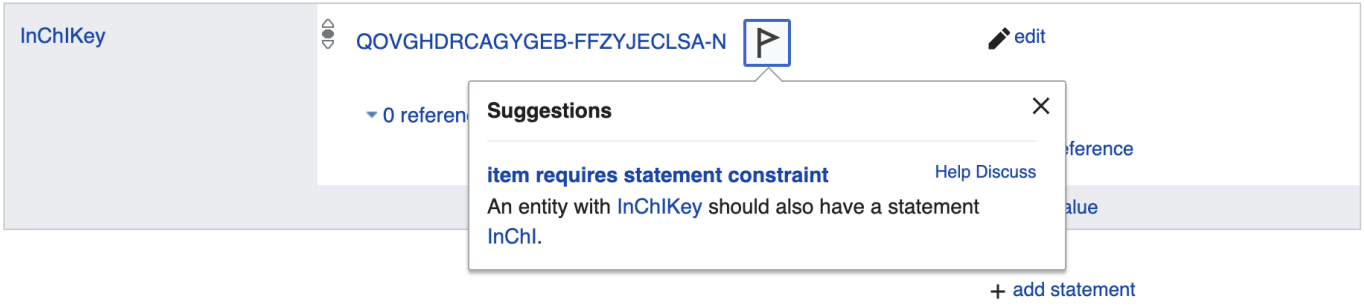
Since you created a new item about an instance of a chemical compound, the user interface will automatically propose to you a set of additional statements commonly found on items about chemical compounds.



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
```

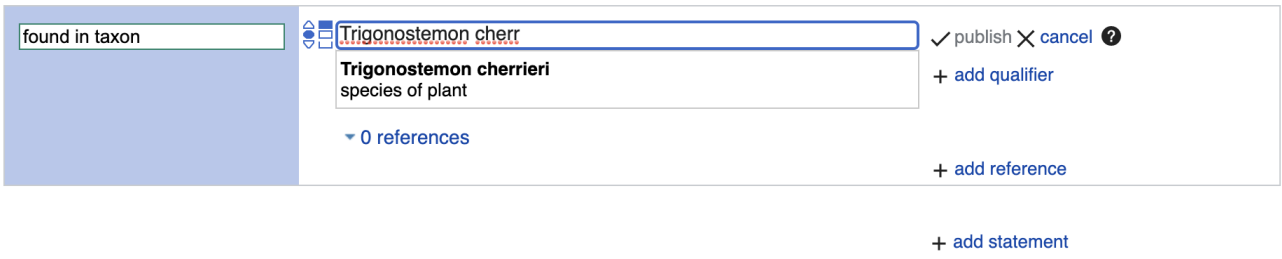
Add the biological source information

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*:



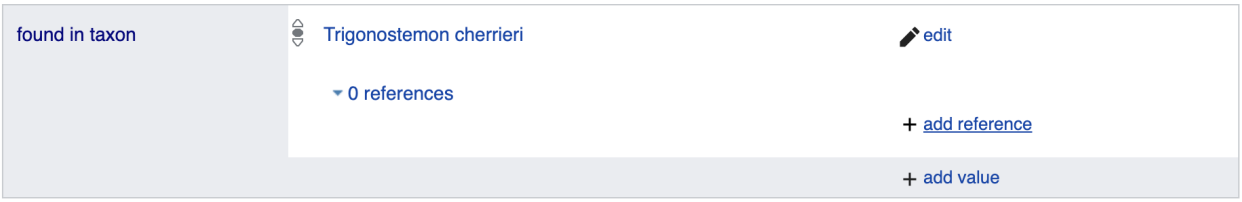
tutorial-image-10

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.

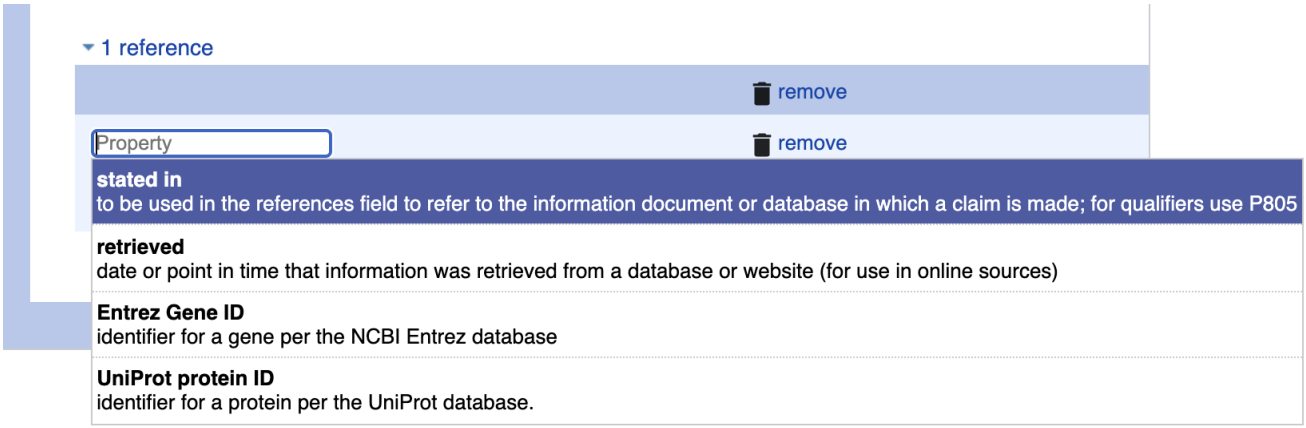
Add the reference documenting the structure-organism pair

Finally, since we report documented 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

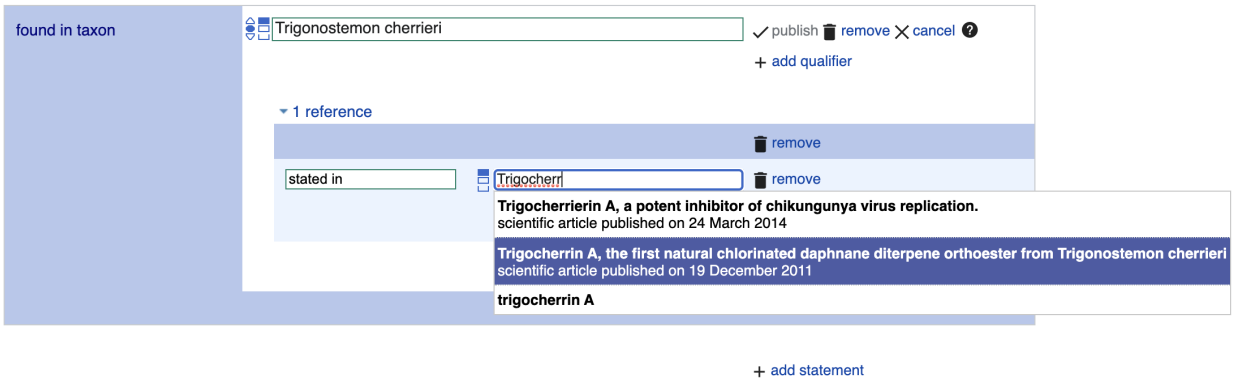
Here, we use the `stated in` property (P248):



tutorial-image-12

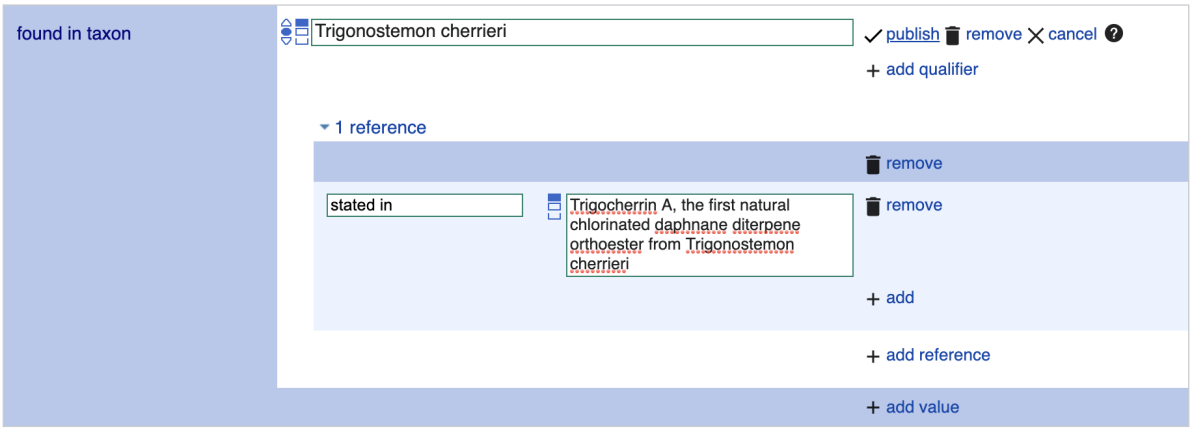
Now, type in the first letters or word of the scientific publication documenting the natural product occurence, autocompletion happens again. Note that multiple publications might have the same title, and that there could be minor differences in punctuation or special characters between the

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 documented 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

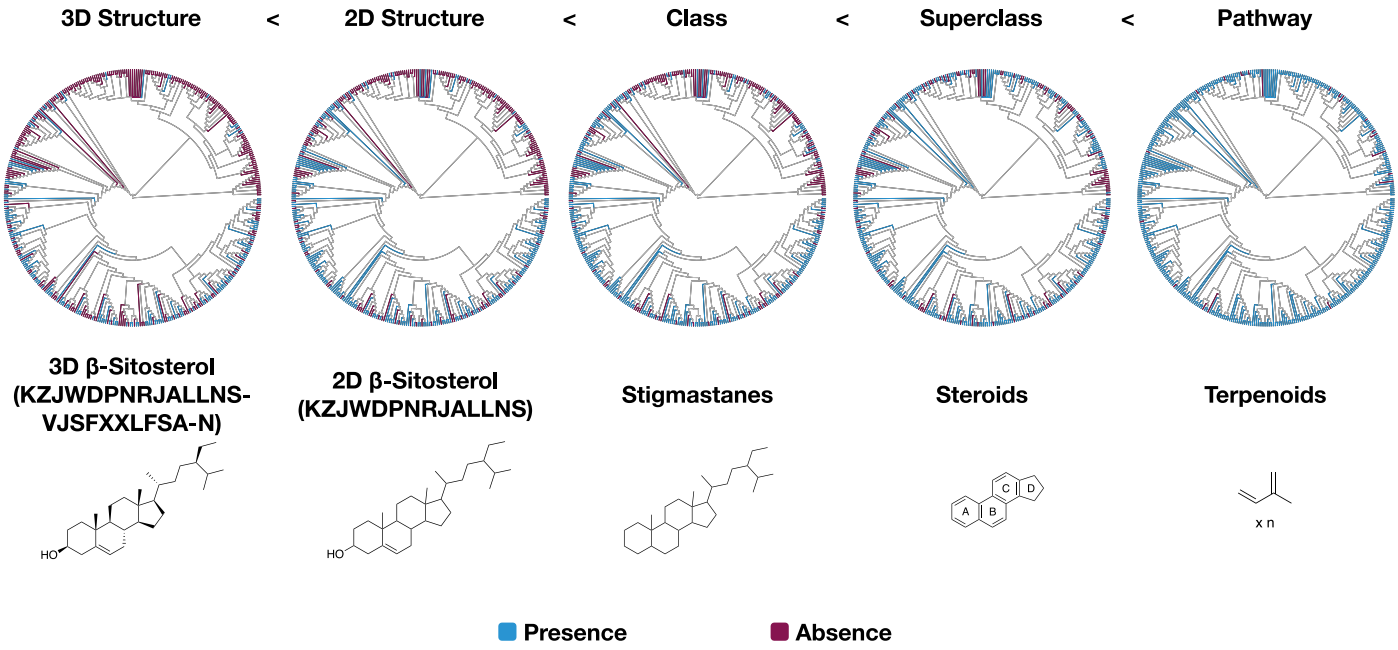


Figure SI-5: Complement to Figure 7: distribution of β-sitosterol and related chemical parents among families with at least 50 reported compounds present in LOTUS. Script used for the generation of each tree in the figure is the same (src/4_visualizing/plot_magicTree.R) as for Figure 7 as both figures are related.

Citations

This list will maagically disappear later on... just here for safety for now!

Afendi (Afendi et al., 2012)

All natural ("All natural," 2007)

Allard (Allard et al., 2018)

Bandyukova (Bandyukova et al., 1983)

Bierer (Bierer et al., 2017)

Bisson IMPS (Bisson et al., 2015)

Bisson NMR (Bisson et al., 2016)

Boonen (Boonen et al., 2012)

Brunson (Brunson, 2020)

Campitelli (Campitelli, 2021)

Cao (Cao et al., 2008)

Capecchi (Capecchi et al., 2020)

Chamberlain (Chamberlain et al., 2020)

Choi (Choi et al., 2017)

Cordell 2001 (Cordell et al., 2001)

Cordell (Geoffrey A. Cordell, 2017a)

Cordell (Geoffrey A. Cordell, 2017b)

Cousijn (Cousijn et al., 2019)

Cousijn (Cousijn et al., 2018)

Davis (Davis and Vasanthi, 2011)

Defossez (Defossez et al., 2021)

Derese (Derese, Solomon et al., 31st August to 3rd September 2015)

Dowle (Dowle and Srinivasan, 2020)

Dührkop (Dührkop et al., 2020)

Feunang (Djoumbou Feunang et al., 2016)

Flor (Flor, 2020)

Gagolewski (Gagolewski, 2020)

GBIF.org ("GBIF.org," 2020)

Gehlenborg (Gehlenborg, 2019)

Giacomoni (Giacomoni et al., 2017)

Graham (Graham and Farnsworth, 2010)

Gu (Gu et al., 2013)

Günthardt (Günthardt et al., 2018)

Hatherley (Hatherley et al., 2015)

Haug (Haug et al., 2019)

Heller (Heller et al., 2013)

Helmy (Helmy et al., 2016)

Hester (Hester and Wickham, 2020)

Horai (Horai et al., 2010)

Huang (Huang et al., 2018)

Hunter (Hunter, 2007)

Ibezim (Ibezim et al., 2017)

Jones (Jones et al., 2021)

Jr (Jr et al., 2020)

Kessler (Kessler and Kalske, 2018)

Kim (kim et al., 2020)

Kim (Kim et al., 2019)

Kim (Kim et al., 2015)

Klementz (Klementz et al., 2016)

Kratochvíl (Kratochvíl et al., 2019)

Kratochvíl (Kratochvíl et al., 2018)

Kuang (Kuang et al., 2019)

Lang (Lang, 2020)

Lin (Lin et al., 2020)

Loo (Loo, 2014)

Lowe (Lowe et al., 2011)

Madariaga-Mazón (Madariaga-Mazón et al., 2021)

Mahto (Mahto, 2019)

McAlpine (McAlpine et al., 2019)

Wakankensaku ("Main Page - WAKANKENSAKU," n.d.)

Martens (Martens et al., 2021)

Michonneau (Michonneau et al., 2016)

Mohamed (Mohamed et al., 2020)

Müller (Müller et al., 2021)

Nielsen (Nielsen et al., 2017)

Noteborn (Noteborn et al., 2000)

Ntie-Kang (Ntie-Kang et al., 2017)

Nupur (Nupur et al., 2016)

Olivon (Olivon et al., 2017)

Ooms (Ooms, 2014)

Pedersen (Pedersen, 2020)

Peroni (Peroni and Shotton, 2012)

Pierce (Pierce et al., 2019)

Pilon (Pilon et al., 2017)

Pilón-Jiménez (Pilón-Jiménez et al., 2019)

Probst (Probst and Reymond, 2020)

Probst (Probst and Reymond, 2018a)

Probst (Probst and Reymond, 2018b)

R (R Core Team, 2020)

R DBI (R Special Interest Group on Databases (R-SIG-DB) et al., 2021)

Rasberry (Rasberry et al., 2019)

Rdkit ("RDKit: Open-source cheminformatics," n.d.)

Reback (Reback et al., 2020)

Rees (Rees and Cranston, 2017)

Rothwell (Rothwell et al., 2013)

Rutz (Rutz et al., 2019)

Sander (Sander et al., 2015)

Sawada (Sawada et al., 2012)

Sedio (Sedio, 2017)

Sharma (Sharma et al., 2014)

Shinbo (Shinbo et al., 2006)

Sievert (Sievert, 2020)

Slenter (Slenter et al., 2018)

Sorokina (Sorokina et al., 2021)

Sorokina (Sorokina and Steinbeck, 2020a)

Sorokina (Sorokina and Steinbeck, 2020b)

Szöcs (Szöcs et al., 2020)

Tomiki (富木 et al., 2006)

Tsugawa (Tsugawa, 2018)

DrDuke ("U.S. Department of Agriculture, Agricultural Research Service. Dr. Duke's Phytochemical and Ethnobotanical Databases." 1992–2016)

van Santen (van Santen et al., 2019)

Virtanen (Virtanen et al., 2020)

Waagmeester (Waagmeester et al., 2020)

Wang (Wang et al., 2020)

Warnes (Warnes et al., 2017)

Weininger (Weininger, 1988)

Wickham (Wickham and Bryan, 2019)

Wickham (Wickham et al., 2019)

Wickham (Wickham, 2020)

Wickham (Wickham et al., 2020)

Wilkins (Wilkins, 2020)

Wilkinson (Wilkinson et al., 2016)

Willighagen (Willighagen et al., 2017)

Winter (Winter, 2017)

Wohlgemuth (Wohlgemuth et al., 2010)

Wohlgemuth (Wohlgemuth et al., 2016)

Wolfender (Wolfender et al., 2020)

Xu (Xu, 2021)

Xu (Xu et al., 2021)

Yabuzaki (Yabuzaki, 2017)

Yu (Yu et al., 2016)

Yue (Yue et al., 2014)

Zeng (Zeng et al., 2018)

Zhang (Zhang et al., 2018)

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

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

Bandyukova VA, Deineko GI, Shapiro DK. 1983. Fatty acid composition of the lipids of pollen (pollen pellets) of some herbaceous plants. III. *Chemistry of Natural Compounds* **19**:97–98. doi:10.1007/bf00579976

Bierer BE, Crosas M, Pierce HH. 2017. Data Authorship as an Incentive to Data Sharing. *New England Journal of Medicine* **376**:1684–1687. doi:10.1056/nejmsb1616595

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

Cordell GA, Quinn-Beattie ML, Farnsworth NR. 2001. The potential of alkaloids in drug discovery. *Phytotherapy Research* **15**:183–205. doi:10.1002/ptr.890

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

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.

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

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).

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

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.

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.

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.

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

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.

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. *Joint Proceedings of the 1st International Workshop on Scientometrics and 1st International Workshop on Enabling Decentralised Scholarly Communication*. 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

Olivon F, Allard P-M, Koval A, Righi D, Genta-Jouve G, Neyts J, Apel C, Pannecouque C, Nothias L-F, Cachet X, Marcourt L, Roussi F, Katanaev VL, Touboul D, Wolfender J-L, Litaudon M. 2017. Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. *ACS Chemical Biology* **12**:2644–2651. doi:10.1021/acschembio.7b00413

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.

Peroni S, Shotton D. 2012. FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics* **17**:33–43. doi:10.1016/j.websem.2012.08.001

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

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

RDKit: Open-source cheminformatics. n.d.

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

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

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

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

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

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

Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O. 2016. SPLASH, a hashed identifier for mass spectra. *Nature Biotechnology* **34**:1099–1101. doi:10.1038/nbt.3689

Wolfender J, Queiroz EF, Allard P. 2020. Massive metabolite profiling of natural extracts for a rational prioritization of bioactive natural products: A paradigm shift in pharmacognosy. *Food Frontiers* **1**:105–106. doi:10.1002/fft2.7

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

富木毅, 斎藤臣, 植木雅, 今野英, 浅岡丈, 鈴木龍, 浦本昌, 掛谷秀, 長田裕. 2006. 理化学研究所・天然化合物バンク(RIKEN NPDepo)の化合物データベース"理化学研究所・天然化合物エンサイクロペディア(RIKEN NPEdia)". *ケモインフォマティクス討論会予稿集*. doi:10.11545/ciqs.2006.0.jl6.0