Open Natural Products Research: Curation and Dissemination of Biological Occurrences of Chemical Structures through Wikidata

Adriano Rutz[1,2], Maria Sorokina[3], Jakub Galgonek[4], Daniel Mietchen[5], Egon Willighagen[6], James Graham[7], Ralf Stephan[8], Roderic Page[9], Jiří Vondrášek[4], Christoph Steinbeck[3], Guido F. Pauli[7], Jean-Luc Wolfender[1,2], Jonathan Bisson[7], and Pierre-Marie Allard[1,2]

[1]School of Pharmaceutical Sciences, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland
[2]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU - Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland
[3]Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07732 Jena, Germany
[4]Institute of Organic Chemistry and Biochemistry of the CAS, Flemingovo náměstí 2, 166 10, Prague 6, Czech Republic
[5]School of Data Science, University of Virginia, Dell 1 Building, Charlottesville, Virginia 22904, United States
[6]Dept of Bioinformatics-BiGCaT, NUTRIM, Maastricht University, Universiteitssingel 50, NL-6229 ER, Maastricht, The Netherlands
[7]Center for Natural Product Technologies, Program for Collaborative Research in the Pharmaceutical Sciences (PCRPS), Pharmacognosy Institute, and Department of Pharmaceutical Sciences, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States
[8]Ontario Institute for Cancer Research (OICR), 661 University Ave Suite 510, Toronto, Canada
[9]IBAHCM, MVLS, University of Glasgow, Glasgow, United Kingdom

February 28, 2021

Abstract

As contemporary bioinformatic and chemoinformatic capabilities are reshaping natural products research, major benefits could result from an open database of referenced structure-organism pairs. Those pairs allow the identification of distinct molecular structures found as components of heterogeneous chemical matrices originating from living organisms. Current databases with such information suffer from paywall restrictions, limited taxonomic scope, poorly standardized fields, and lack of interoperability. To ensure data quality, references to the work that describes the structure-organism relationship are mandatory. To fill this void, we collected and curated a set of structure-organism pairs from publicly available natural products databases to yield **LOTUS** (natura**L** pr**O**duc**T**s occ**U**rrences databa**S**e), which contains over 500,000 curated and referenced structure-organism pairs. All the programs developed for data collection, curation, and dissemination are publicly available. To provide unlimited access as well as standardized linking to other resources, LOTUS data is both hosted on Wikidata and regularly mirrored on https://lotus.naturalproducts.net. The diffusion of these referenced structure-organism pairs within the Wikidata framework addresses many of the limitations of currently-available databases and facilitates linkage to existing biological and chemical data resources. This resource represents an important advancement in the design and deployment of a comprehensive

and collaborative natural products knowledge base.

* Correspondence:

jean-luc.wolfender@unige.ch

bjo@uic.edu

pierre-marie.allard@unige.ch

Graphical abstract

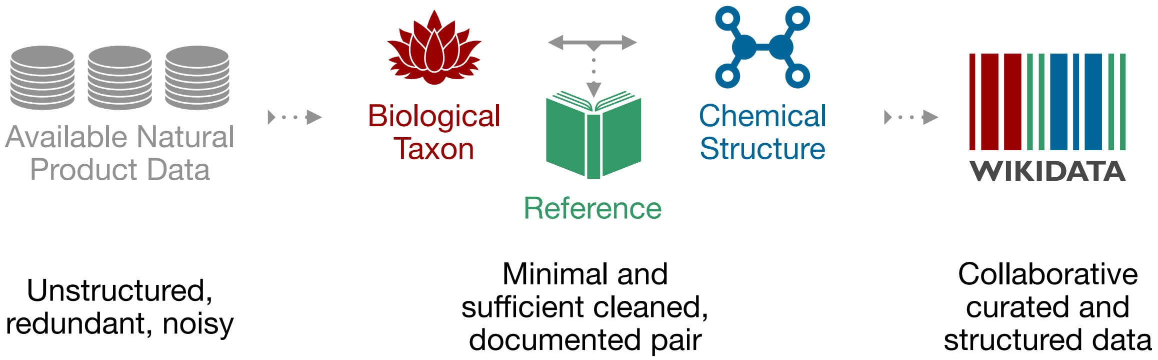
Available Natural Product Data
Unstructured, redundant, noisy
Biological Taxon
Reference
Chemical Structure
Minimal and sufficient cleaned, documented pair
WIKIDATA
Collaborative curated and structured data

Figure 1: Graphical abstract

Introduction

Natural products (NPs) research is a transdisciplinary field with interests ranging from the fundamental structural aspects of molecular entities to their effects on living organisms, or the study of chemically-mediated interactions within entire ecosystems. Recent technological and methodological advancements are currently reshaping the field of NP research. This field, which has a long history also deals with significant traditional elements (Allard et al., 2018). In particular, contemporary bioinformatic approaches enable the (re-)interpretation and (re-)annotation of datasets originating from complex biological matrices (Olivon et al., 2017). To efficiently annotate previously-reported NPs, or to identify new entities, these tools rely on properly maintained NP databases (DBs) (Tsugawa, 2018). Assuming that a NP is a chemical entity found in or produced by a living organism ("All natural", 2007), a NPs DB should at least contain a list of chemical entities, organisms, and the reference to the work describing the established links between them. However, most DBs favor the chemical objects or the biological ones and just a few report the links between these objects. Large and well-structured DBs, composed only of chemical structures (PubChem (Kim et al., 2018), over 100M entries) or biological organisms (GBIF ("GBIF.org", 2020), over 1,400M entries) are freely accessible but operate independently. Currently, no open, cross-kingdom and comprehensive DB links NPs and their producing organisms, along with information about the experimental works describing those links. It is precisely those referenced structure-organism pairs that are critical for NP and related research but which are scarcely accessible (Cordell, 2017). Pioneering efforts led by Shinbo et al. led to the establishment of KNApSAck (Shinbo et al., n.d.), likely the first public curated DB of referenced structure-organism pairs. KNApSAck currently contains over 50,000 structures and over 100,000 structure-organism pairs. However, its organism field is not standardized and download is complicated. The NAPRALERT dataset (Graham and Farnsworth, 2010), compiled by Farnsworth and colleagues 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 compound/species pairs, with over 900,000 records of compound species pairs due to multiple reports of equivalent compound/species pairs from different citations. NAPRALERT is searchable, but the data are not openly available online. Finally, the NPAtlas (van et al., 2019) is a more recent project aimed at complying with the FAIR (Findability, Accessibility, Interoperability, and Reuse) guidelines for digital assets (Wilkinson et al., 2016) and offering web access. While the NPAtlas encourages submission of new compounds with their biological source, it focuses at the moment on microbial NPs and ignores a wide range of biosynthetically active organisms, such as the plant kingdom.

Most of the available NPs DBs provide entries without referencing their origin, thus breaking the precious link for tracing information back to the original data and assessing its quality. Even valuable efforts for compiling NP data made by commercial DB distributors, such as the Dictionary of Natural Products (DNP) are missing documentation of this informational pair, precluding further computational use or exhaustive review. To compensate for these shortcomings, our project aims at curating and disseminating a structured natura**L** pr**O**duc**T**s occ**U**rrence databa**S**e (**LOTUS**). Taking FAIR principles as guidance, we selected Wikidata (WD) for disseminating this resource as it was the best candidate with its focus on cross-disciplinary and multilingual support. It agglomerates referenced structure-organism pairs from publicly available data. After collection and harmonization, each documented structure-organism pair has been curated at the chemical, biological, and reference level, resulting in atomic and computer-interpretable identifiers. It is curated and governed collaboratively by a global community of volunteers, about 20,000 of which are contributing monthly. While closely integrated with Wikipedia and serving as its source for its infoboxes, WD represents information as machine-interpretable statements in the form of subject-predicate-object triples, which can be enriched with qualifiers and references. WD currently contains more than 1 billion statements covering ~90 million entries. Entries can be grouped into classes such as countries, songs, disasters or chemical compounds. Workflows have been established for the reporting of such classes, particularly those of interest to the life sciences, such as genes, proteins, diseases, drugs, or biological taxa (Waagmeester et al., 2020).

Building on the above principles and experiences, this report introduces the development and implementation of workflows for NPs occurrence curation and dissemination using **TRUST** (**T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability and **T**echnology) principles (Lin et al., 2020). The presented data upload and retrieval procedures ensure optimal data accessibility and foster reuse by the research community, by allowing any researcher to contribute and reuse the data with a clear and open license (Creative Commons 0). Despite all these advantages, the WD hosting of the LOTUS project presents some drawbacks. While the SPARQL query language offers a powerful way to interrogate available data, it can also appear intimidating at first for the inexperienced user. Furthermore, some typical queries of molecular DBs such as structural or spectral search are not yet available in WD. To bridge this gap, we decided, in parallel, to host LOTUS in a more traditional format in the naturalproducts.net ecosystem of databases and tools as the LNPN project (`https://lotus.naturalproducts.net`) (LNPN), This repository is periodically updated with the latest LOTUS data. The advantages of this dual hosting are the production of both a community-curated and vast knowledge-based integrated DB (via WD) and a NP community-oriented product, including tailored search modes as described above.

We expect that the LOTUS project and its multiple data interaction possibilities will provide a solid basis to establish transparent and sustainable ways to access, share and create knowledge on NPs occurrence and, more widely, participate in the cross-fertilization of the fields of chemistry and biology. Hereafter, we present an overview of the LOTUS blueprint, as a snapshot of where it stands at the time of this document writing. We detail the central collection, curation and dissemination stages. We then expose possibilities for the end user to interact with LOTUS, whether by retrieving, adding or editing data. We finally illustrate the dimensions and qualities of the current LOTUS dataset from the chemical and biological perspectives.

Results & Discussion

Outline of the LOTUS blueprint

To avoid classical pitfalls of public scientific DB creation (Helmy et al., 2016), and to enhance current and future dissemination, WD appears as an ideal repository. Building on the standards established by three existing WD projects in chemistry (Wikidata:WikiProject Chemistry), taxonomy (Wikidata:WikiProject Taxonomy), and source metadata (Wikidata:WikiProject Source MetaData), we created a NPs chemistry oriented subproject (Wikidata:WikiProject Chemistry/Natural products) that inherited the data formats employed in the parent and related WD projects. The central data was constituted of 3 minimal sufficient objects, allowing retrieval of all associated information:

- A chemical structure, defined by an International Chemical Identifier (InChI) (Heller et al., 2013), a Simplified Molecular Input Line Entry System (SMILES)(Weininger, 1988), and an InChIKey (a hashed version of the InChI) to avoid any possible collisions.
- A biological organism, defined by its taxon name, taxon ID, and the associated taxonomic DB.
- A reference describing the structure-organism pair, defined by its corresponding title and a Digital Object Identifier (DOI), a PubMed ID (PMID), or a PubMed Central ID (PMCID).

As data formats are inhomogeneous among existing NP DBs, fields related to chemical structure, biological organism, and literature reference are variable and essentially unstandardized. Therefore, LOTUS implements multiple steps of collection, harmonization, curation, and dissemination. Figure 2, stage 1 to 3. LOTUS is elaborated with a Single Source of Truth (SSOT, Single_source_of_truth) to ensure data reliability and always propose the latest curated version of the LOTUS data both at WD and LNPN. The SSOT consists of a PostgreSQL DB structuring links and data schema so that every data element is in a single place. By accommodating further data addition (directly as new data sources or at the WD level) the LOTUS processing pipeline is tailored to efficiently include and diffuse novel or curated data. Figure 2, stage 4. This iterative workflow relies both on data addition and data retrieval actions described in the Data interaction section. The overall process leading to referenced and curated structure-organisms pairs is illustrated in Figure 2, and detailed below.

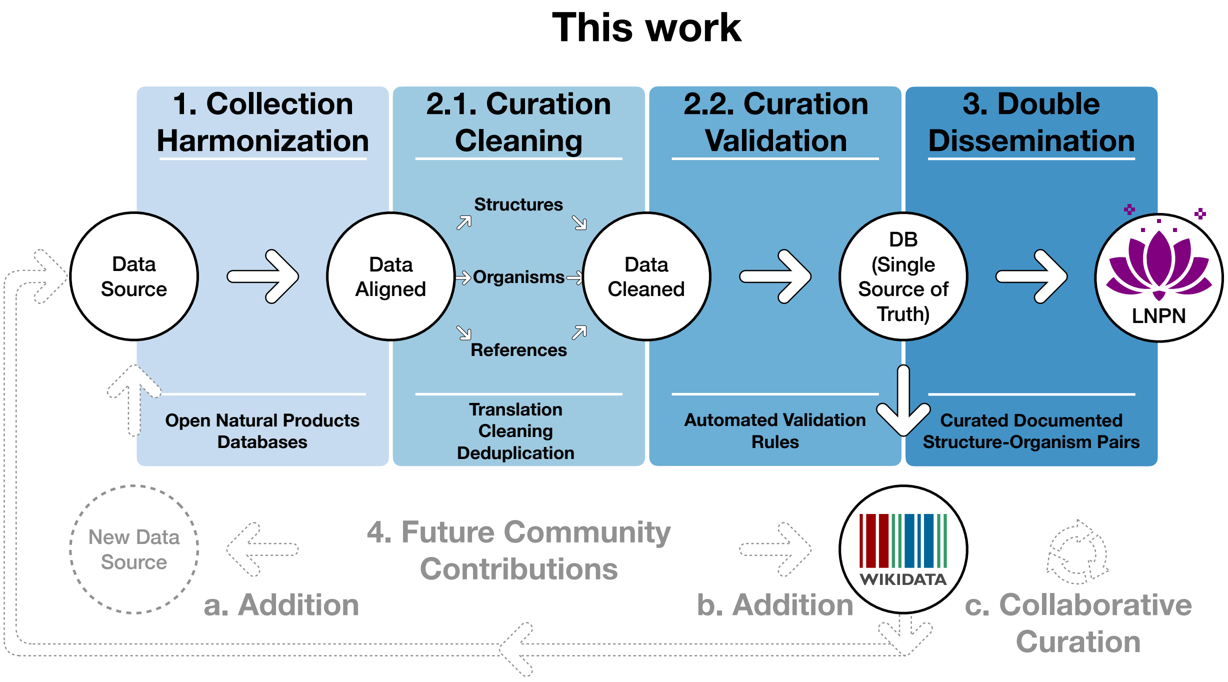
This work
1. Collection Harmonization
2.1. Curation Cleaning
2.2. Curation Validation
3. Double Dissemination
Data Source
Data Aligned
Structures
Organisms
References
Data Cleaned
DB (Single Source of Truth)
LNPN
Open Natural Products Databases
Translation Cleaning Deduplication
Automated Validation Rules
Curated Documented Structure-Organism Pairs
New Data Source
a. Addition
4. Future Community Contributions
b. Addition
WIKIDATA
c. Collaborative Curation

Figure 2: LOTUS process overview. The process consists of three main steps: Collection and Harmonization (1), Curation (2) and Dissemination (3). It was conceived to integrate Future contributions (4) either as new data addition (a. and b.) or as curation of existing data (c.) and thus build an iterative virtuous circle empowering the community to participate in the global NPs occurrences documentation effort.

All the steps of the process can be found in the https://gitlab.com/lotus7 project and https://github.com/mSorok/LOTUSweb. At the time of submission, this leads to 742,041 entries consisting of a curated chemical structure, a curated biological organism, and a curated reference available on WD and LNPN. Since 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 available at https://osf.io/hgjdb/. This frozen dataset is the one that has been used to generate Figures 4 and 6.

Collection and harmonization

Initial data were collected from the recently published COlleCtion of Open NatUral producTs (COCONUT)(Sorokina et al., 2021). All DBs referred to as open-access in COCONUT and containing referenced structure-organism pairs were used. They were complemented with COCONUT's own structure-organism documented pairs (Sorokina and Steinbeck, 2020) and the following additional DBs: Dr. Duke (Duke, 2016), Cyanometdb (Jones et al., 2020), Datawarrior (Sander et al., 2015), a subset of NAPRALERT (Graham and Farnsworth, 2010), Wakankensaku ("Wakankensaku", n.d.), and DiaNat-DB (Madariaga-Mazón et al., 2021) . The list of data sources is available in Supplementary Table S1. All necessary scripts for collection and harmonization can be found in the lotusProcessor repository in the src/1_gathering directory and the process is detailed in the corresponding Methods section (Ibezim et al., 2017), (Boonen et al., 2012)(Pilón-Jiménez et al., 2019), (Sharma et al., 2014), (Yabuzaki, 2017), (Sorokina and Steinbeck, 2020), (Zhang et al., 2019), (Afendi et al., 2012), (Haug et al., 2020), (Kautsar et al., 2020), (Derese, Solomon et al., 2015), (Ntie-Kang et al., 2017), (Zeng et al., 2018), (Choi et al., 2017), (Tomiki et al., 2006), (Pilon et al., 2017), (Huang et al., 2018), (Rothwell et al., 2013), (Giacomoni et al., 2017), (Nupur et al., 2016),(Sawada et al., 2012), (Hatherley et al., 2015), (Klementz et al., 2016), (Davis and Vasanthi, 2011), (Yue et al., 2014), (Kim et al., 2015), (Günthardt et al., 2018), (Gu et al., 2013). All subsequent iterations with additional data sources (either the updated versions of the same data sources or new ones), will first compare the new data sources with previously collected ones at the SSOT level in order to curate data only once.

Curation

As described in Figure 2, the data curation process was divided into alignment, cleaning, and validation stages. Cleaning of each of the three central objects (the chemical, the biological, and the reference object) of the referenced pairs was performed before realignment. The overall process is detailed in the corresponding Methods section. Given the size of the data (more than 2.5M initial entries), manual validation was not possible. An especially problematic point of the curation process was encountered while treating the references. If organisms are often reported at least by their canonical name, structures by their SMILES, InChI or InChIkey, references suffer from insufficient reporting standards. The major inconvenience is poor information retrieval26. Better reporting together with new tools such as Scholia (`https://scholia.toolforge.org/`) (Nielsen et al., 2017), relying on Wikidata, Fatcat (`https://fatcat.wiki/`), or Semantic Scholar (`https://www.semanticscholar.org/`) should allow improved information retrieval in the future. Despite the poor standardization of the initial reference field, this last object is crucial to establish the validity of the structure-organism pair. After the curation of the chemical and biological objects, the references were thus exploited to assess the quality of the documented structure-organism pair. In addition to the entries we curated as we processed the data, we also manually analyzed 420 referenced structure-organism pairs to establish rules for automatic filtering of the curated entries. This filter (detailed in the corresponding Methods section) was then applied to all entries. To confirm the efficacy of the filtering process, a second representative set of 100 entries was subsetted and its manual validation led to a rate of 97% of true positives. See results of the two manual validation steps in Supporting Information S2. Resulting data are available in the dataset shared at `https://osf.io/hgjdb/`. In Table 1, we show an example of a referenced structure-organism pair before and after the curation process, which resolved the structure to an InChIKey, the organism to a valid taxonomic name and the reference to a Digital Object Identifier (DOI). Challenging examples encountered during the curation process development were compiled in an edge case table (tests/tests.tsv), which allowed automatic unit testing (see corresponding Methods section). These tests allow a continuous revalidation of any change made to the code, making sure no corrected error can reappear.

	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	10.1/NP50117A020

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

The alluvial plot in Figure 3 illustrates the individual contribution of each curated DBs and original subcategory to the final structure, organism and reference categories. For example, the high contribution of NAPRALERT and UNPD (Gu et al., 2013) is highlighted. The important contribution of the DOI category of references to the validated references is also clearly visible. Combining the results of the automatic curation pipeline and our manually curated entries, led to the establishment of four categories (manually validated, manually rejected, automatically validated, and automatically rejected) of documented structure-organism pairs that constituted the SSOT DB. Out of a total of more than 2M pairs, manually and automatically validated pairs constituted over 740,000 pairs, or circa 30 %, which were selected for dissemination on WD. They were constituted from over 250,000 structures, over 30,000 organisms, and over 75,000 references. The technical details of the curation cleaning and validation processes are described in the corresponding methods section. All necessary programs for curation can be found in the lotusProcessor repository under src/2_curating and src/3_analysing.

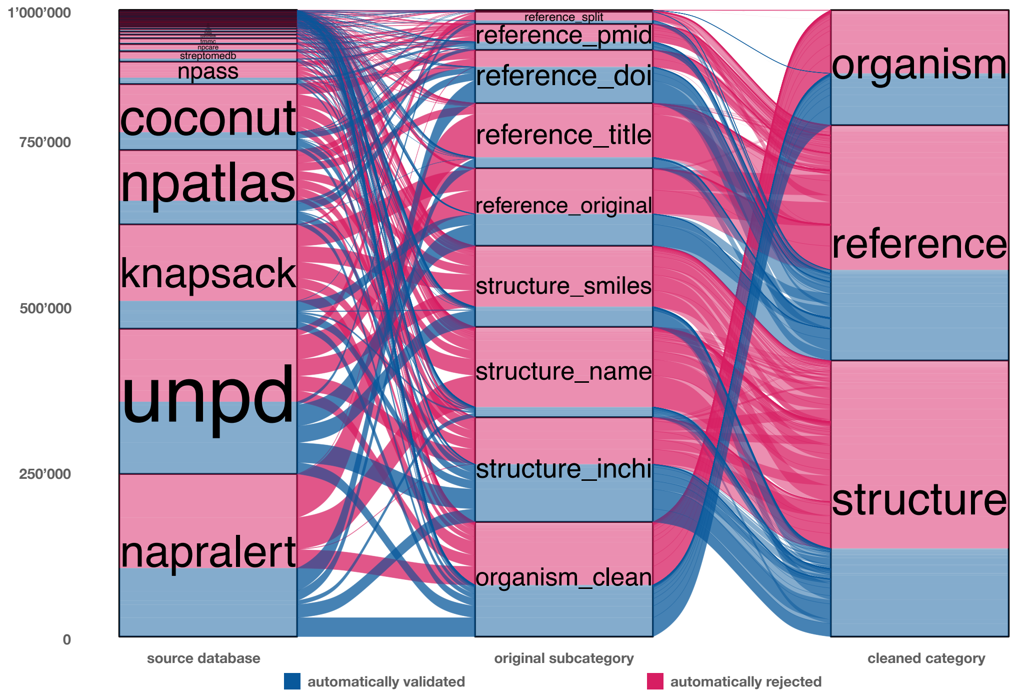

Figure 3: Alluvial plot of the LOTUS data flux during the automated curation and validation process: it represents the relative proportion of individual entries' data streams and repartition by database (first block), their harmonized subcategories for curation (second block), and their final validation status (third block). Automatically validated entries are represented in blue and rejected ones in red.

Dissemination

Ideally, researchers should benefit from all results of studies in their field and adjacent areas immediately upon publication. This is considered as the foundation of scientific investigation and a prerequisite for effectively directing new research efforts based on available prior information. To achieve this, research results have to be made publicly available and reusable. As computers are now the main instrument of any scientist, this data needs to be computer interpretable, publications should contain structured data to be efficiently organized and summarized in a DB-compatible form. Following the FAIR guidelines, we chose WD as the repository for the referenced structure-organism pairs. Hosting on WD enables the documented research data on NPs to be integrated with the pre-existing body of chemical and biological knowledge. The flexibility of SPARQL queries, the language used to query WikiData and many other resources, allows users to efficiently retrieve and query data. Besides, by being hosted on a dynamic platform, the quality of the data is expected to evolve continuously, benefiting from the curation by the different user communities of WD.

Despite the numerous advantages of the WD-based hosting of LOTUS, in particular its independence from individual lab funding, some limitations were anticipated. SPARQL queries, despite their power, are complex and often require a good understanding of the models and the structure of the data. This can discourage some end users as it requires there is a steep learning curve. Furthermore, traditional ways to query NP DBs such as structural or spectral searches are currently not within the scope of WD. Based on the pre-existing COCONUT DB template, LOTUS data is also hosted at https://lotus.naturalproducts.net (LNPN) to facilitate such structure-based searches. The double diffusion of LOTUS on WD and at LNPN addresses these shortcomings: the WD hosting allows to benefit from the integration of the uploaded data within the whole WD knowledge base and elaborated SPARQL queries to explore this dataset under various angles. The WD hosting also opens the community curation possibilities which will guarantee a dynamic and evolving data repository. On the other hand, LNPN hosting allows the user to perform structural searches more classically (e.g., by drawing the molecule). In the future, versions of LOTUS and COCONUT augmented by predicted MS spectra are expected to be hosted at the naturalproducts.net portal and should allow mass, fragment and spectral based queries. To facilitate queries focussed on specific taxa(e.g., "return all molecules found in the Asteraceae family"), a unified taxonomy is paramount. As taxonomy is a complex and always evolving field, we decided to keep all the taxon identifiers from all accepted taxonomic DB for a given taxon name. This implies that for a given name, multiple taxonomies, coming from different taxonomic DB, are allowed. We expect initiatives such as the Open Tree of Life (OTL) (https://tree.opentreeoflife.org/) (Rees and Cranston, 2017) to gradually reduce those discrepancies and WD to efficiently help in this direction. OTL also benefits from regular expert curation and new data. As the taxonomic identifier property for this database did not exist in WD, we requested and obtained its creation (P9157). After the previously described curation process, all validated entries were thus made available through WD and LNPN. LNPN will be regularly mirroring WD LOTUS through the SSOT as described in Figure 2. Below, we will describe the various ways to interact with data hosted at WD and LNPN.

Data interaction from the user point of view

The data being available in multiple formats, the possibilities to interact with the LOTUS data are numerous. We provide hereafter some basic and more advanced examples on how to retrieve, add and edit LOTUS data.

Data retrieval

LOTUS data can be queried and retrieved either on WD directly or on LNPN. Both of these options present unique advantages. Wikidata offers modularity at the cost of a 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/hgjdb/. Hereafter we detail finer approaches to directly interrogate the up-to-date LOTUS data both in WD and LNPN.

Wikidata

The simplest way to search for NPs occurrence information in WD is by directly typing the name of a chemical structure in the "Search Wikidata" lookup field. For example by typing erysodine the user lands on the WD page for 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.For more elaborated queries, the usual way is to write SPARQL queries in the Wikidata Query Service. Below are some examples of simple or more elaborated requests which can be done using this service. 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 group of stereoisomers (Q59199015) found in taxon (P703) taxon stated in bibliographic reference (Q10358455) is available here: 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 angle of each one of the three objects constituting the referenced structure-organism pairs can be 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 beta-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 DBs such as structural search are not yet available in WD. It is a general issue, as the SPARQL language does not support a simple integration

Questions	Wikidata SPARQL query
What are the compounds present in Mouse-ear cress (Arabidopsis thaliana)?	https://w.wiki/32y8
Which organisms are known to contain the 2D structure of beta-sitosterol?	https://w.wiki/334q
Which taxa have chemical compounds related to (but different from) beta-sitosterol?	https://w.wiki/334s
What are examples of organisms where compounds were reported to be produced by a sister organism but not the organism itself?	https://w.wiki/3359
Which Zephyranthes species lack compounds known from at least two sister species?	https://w.wiki/335x
How many compounds are structurally similar to compounds labelled 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 compounds have an indolic scaffold?	https://w.wiki/32KZ
How many structure-organism pairs have been referenced by these authors? (Here, we compare two senior natural products chemists and co-authors of this paper with the late Ferdinand Bohlmann).	https://w.wiki/32$m

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

of such queries. To address this issue, Galgonek et al. developed an in-house SPARQL engine that allows utilization of Sachem, a high-performance chemical DB cartridge for fingerprint-guided substructure and similarity search (Kratochvíl et al., 2018). The engine is used by the Integrated Database of Small Molecules (IDSM) that operates, among other things, several dedicated endpoints allowing structural search in selected small-molecule datasets via SPARQL (Kratochvíl et al., 2019). To allow substructure and similarity searches via SPARQL also on compounds from WD, we created a dedicated IDSM/Sachem endpoint for WD as well. The endpoint indexes isomorphic (P2017) and canonical (P233) SMILES code available in WD. To ensure that data are kept up-to-date, SMILES codes are downloaded from WD automatically daily. The endpoint allows users to run federated queries (https://www.w3.org/TR/sparql11-federated-query/) 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 the indolic scaffold. The list is aggregated at the parent taxa level of the containing organisms and ordered by the number of scaffold occurrences.

LNPN

In the search field of the LNPN interface, simple queries can be achieved by typing in the molecule name (e.g. protopine), pasting a SMILES string or an return all compounds found in a given organism by typing the organism name at the species or any higher taxa level (e.g. Tabernanthe iboga). Alternatively, structure can be directly drawn in the Structure search interface (https://lotus.naturalproducts.net/search/structure). Refined search mode combining multiple search criteria 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 MongoDB dump via https://lotus.naturalproducts.net/download.

Data addition

A strong advantage of LOTUS is that the possibility is given for users to contribute to the NPs occurrences documentation effort by adding new data or editing uploaded data.

All of the data managed by LOTUS is stored in the SSOT. The SSOT is also used to avoid reprocessing elements that have already been previously obtained such as a structure from a name, a bibliographical reference from a citation or a taxonomic identifier from a taxon name. However, at the moment, we are not opening the SSOT for direct write access to the public in order to maintain its coherence and allow us to make the schema evolve. To add or modify data in LOTUS, the users can employ the following approaches.

Source databases

The LOTUS process will regularly re-import both the current source DBs and new ones. New and modified information from those DBs will be checked against the SSOT and if not present or updated they will follow the curation pipeline and will be further stocked into SSOT. Any researcher can, thus, contribute to these DBs as a 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 directly on Wikidata. This data will then be imported into the SSOT database. 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 NPs occurrence reporting will be able to manually add the data directly in WD, without programming language barriers of any kind. The only prerequisite is to create a Wikidata account and follow the general object editing guidelines (https://www.wikidata.org/wiki/Wikidata:Tours). Regarding the addition of NPs centered objects (documented structure-organisms pairs) please refer to the WikiProject Chemistry/Natural products group page https://www.wikidata.org/wiki/Wikidata:WikiProject_Chemistry/Natural_products.

A tutorial for the manual creation and upload to WD of a documented structure-organism pairs is available in Supplementary Information . While direct WD upload is possible, future contributors are still encouraged to use the LOTUS curation pipeline as a preliminary step to strengthen initial data quality. The added data will then benefit from the curation and validation stages implemented in the LOTUS processing pipeline.

Batch and automated upload

At the end of the previously described curation process, more than 500,000 referenced structure-organisms were validated for WD addition. To automate the WD upload process, we wrote a set of scripts that automatically process the curated outputs, group references, organisms, and compounds together, check if they are already present in WD (using SPARQL and direct connection to WD), and insert or update the entities as needed (upserting). These scripts can be used for batch upload of properly curated and referenced structure-organism pairs to WD. Scripts for data addition on WD can be found in the repository Wikidata-LotusImporter. The Xtools page offers an overview of the latest WikidataLotusImporter activity.

Data editing

Even if correct at a given time point, scientific advances can invalidate the data later on. Thus, possibilities to continuously edit the data are desirable and guarantee data quality and sustainability. Community-maintained knowledge bases such as WD allows such a process. WD presents the advantage of allowing both manual and automated correction. Field-specific robots (SuccuBot, KrBot, Pi_bot, and ProteinBoxBot), (SuccuBot, KrBot, Pi_bot), (SuccuBot, KrBot, Pi_bot), (SuccuBot, KrBot, Pi_bot), which have gone through an approval process, can make multiple thousands of edits without the need for human decision-making. This helps in automatically reducing the amount of incorrect data but remains incomplete. However, manual curation by experts remains the highest standard. Valuing this quality, we suggest interested users to follow the manual curation tutorial in Supplementary Information .

The Scholia platform offers an example of a powerful and user-friendly edition interface for scientific references. The adaptation of such a framework to edit the LOTUS documented structure-pairs could facilitate the collection of manual experts curation inputs in the future.

Data interpretation

To illustrate the nature and dimensions of the LOTUS dataset we showcase hereafter selected data interpretation examples. We first describe the distribution of biological organisms according to the number of related chemical structures and likewise the distribution of chemical structures across biological organisms (Figure 4). We then picture individual DBs coverage using upset plot depiction, which allows the visualization of multiple intersecting datasets (Figure 5). In these two previous interpretations we take the cases of sitosterol, for the chemical structure and of *Arabidopsis thaliana*, for the biological organisms, to provide well documented entries to the reader. Finally, we present a chemically-informed taxonomical tree qualitatively illustrating the overall chemical and biological coverage of LOTUS by linking family-specific classes of chemical structures to their taxonomic position (Figure 6). Figure 4 and 6 were generated using the frozen table (at the 23.02.21) available here `https://osf.io/hgjdb/`. Figure 5 required a dataset containing information from a commercial DB (DNP) and is thus not available for public distribution. All scripts used for figures generation are available in the lotusProcessor repository in the src/4_visualizing folder.

Organisms per structure and structure per organisms distribution

As depicted in Figure 4, on average, 3 organisms are reported per chemical structure and 11 structures per organism. Half of the structures are reported in only 1 organism and 5 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). Such numbers suggest that a better reporting of the metabolites during a phytochemical investigation could greatly improve coverage. A bias partly explaining this incomplete coverage may come from the fact that, usually, only newly described or bioactive structures are accepted for publication in classical NPs research journals.

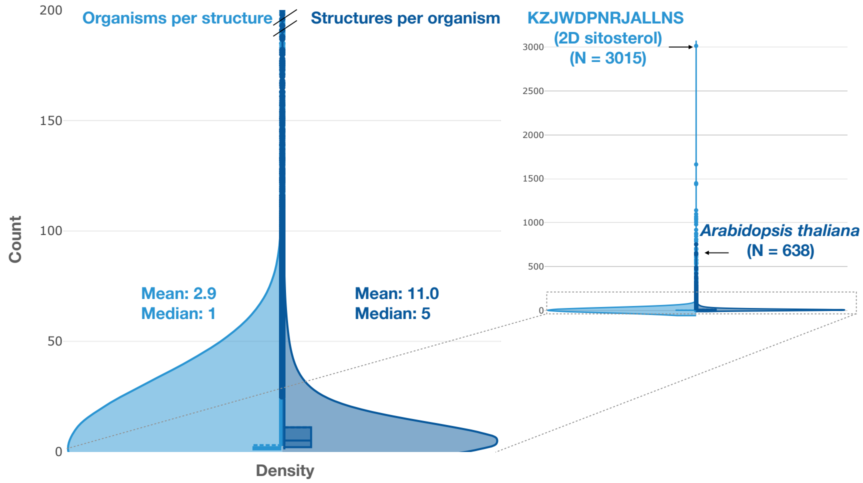

Figure 4: Violin plot representing the density of structures found in organisms and organisms containing structures. Number of organisms linked to the 2D structure of sitosterol (KZJWDPNRJALLNS) and the chemical diversity of Arabidopsis thaliana are highlighted as two notable examples.

Individual DB contribution to LOTUS

The added value of assembling all available NPs DBs in WD is illustrated in Figure 5, showing the individual DBs contribution to all chemical structures found in *Arabidopsis thaliana* (“Mouse-ear cress”; Q147096) (A) and all taxa containing the two-dimensional structure corresponding to sitosterol (Q121802) and (Q63409374) (B), a compound of ubiquitous occurrence in higher plants.

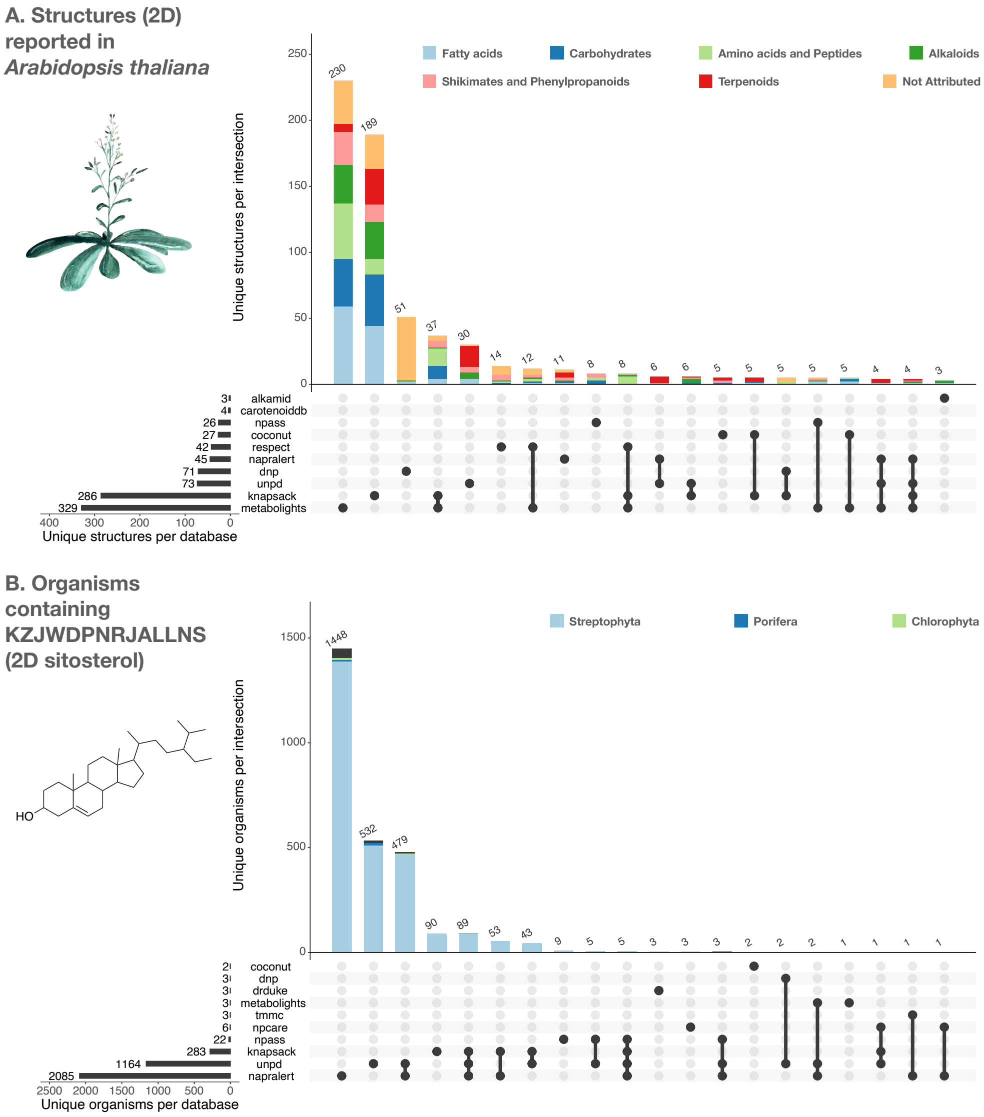

Figure 5: Upset plots of the individual DBs contribution to 2D structures found in Arabidopsis thaliana (A) and organisms containing the 2D structure of sitosterol (KZJWDPNRJALNS) (B). Upset plots are evolved Venn diagrams, allowing to represent intersections between multiple sets. The horizontal bars on the lower left represent the number of corresponding entries per database. 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 sitosterol are present in both UNPD and NAPRALERT, which in turn, respectively report 1164 and 2085 organisms containing the structure of sitosterol.

Figure 5.A shows that the chemical pathways distribution (according to NPClassifier (Kim et al., n.d.) across DBs is not conserved. Note that being specially tailored for NPs, NPClassifier was prefered over ClassyFire (Feunang et al., 2016) but both chemical taxonomies are available as metadata in the frozen LOTUS export (`https://osf.io/hgjdb/`) and in 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 NP DBs 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 sitosterol (KZJWDPNRJALLNS) has been reported for each intersection is shown. NAPRALERT has by far the highest number of entries (2085 in total), while other DBs complement this well (UNPD, for example, has 532 organisms where sitosterol is reported that are not overlapping with the ones reported in NAPRALERT). Interestingly, 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.

Chemically-informed taxonomic tree

A summary of the biological and chemical diversity covered by LOTUS is illustrated in Figure 6. To limit biases due to underreporting while keeping a reasonable display size, only families with at least 50 reported structures 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 belonging. 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.

In Figure 6, it is possible to spot highly specific compound classes such as trinervitane terpenoids in the Termitidae, rhizoxin macrolides in Rhizopodaceae or typical quassinoids and limonoids from Simaroubaceae and Meliaceae, respectively. More generic tendencies can also be observed. For example, within the fungal kingdom, Basidiomycotina appears to have a higher biosynthetic specificity toward terpenoids than the rest of the members, 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 beta-carotene (OENHQHLEOONYIE-JLTXGRSLSA-N). We specifically looked at the repartition of the sitosterol scaffold (KZJWDPNRJALLNS) within the overall biological taxonomy. For this we plotted the presence/absence of the sitosterol scaffold, and its two superior chemical classification, namely stigmastane and steroid derivatives, over the taxonomic tree used in Figure 6. The comparison of these three chemically-informed taxonomic trees clearly highlighted the increasing speciation of the sitosterol biosynthetic pathway in the Archaeplastida kingdom, while the upper classes were distributed across all kingdoms. See Supplementary Information . As illustrated, the possibility to interrogate data at multiple precision levels is valuable. As recently shown in the frame 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 involved pathways and organisms carrying the genes responsible for the biosynthesis of these structures would be of interest to confirm the previous observations. These selected data interpretations establish the importance of reporting not only new structures but also novel occurrences of known structures in organisms. Then only, comprehensive chemotaxonomic studies will allow a better understanding of living organisms' metabolomes.

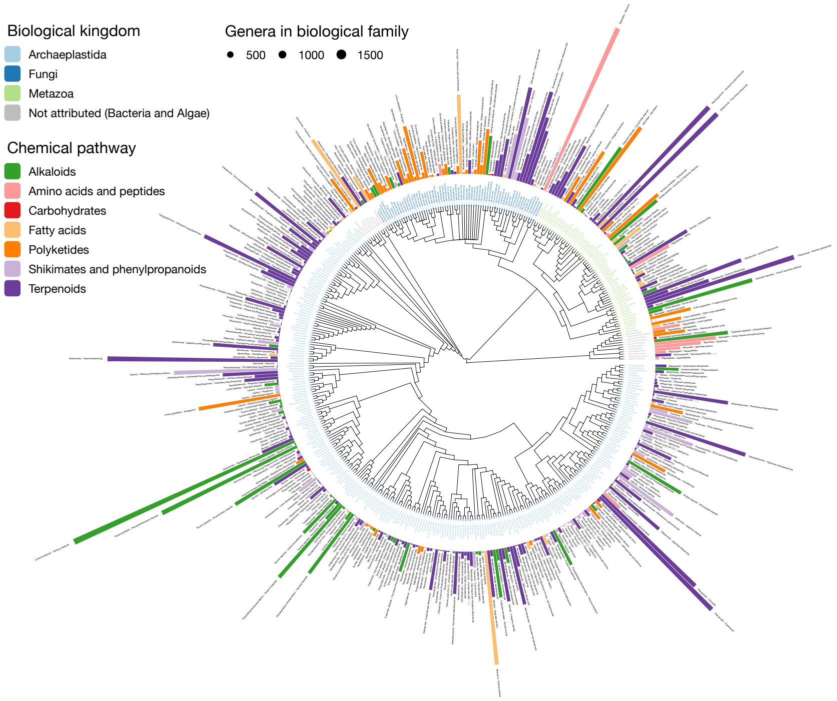

Figure 6: The chemical and biological diversity within LOTUS. The tree corresponds to the biological taxonomy, with kingdom as label color. The size of the leaves node corresponds to the number of genera reported in the family. The outer bars correspond to the most specific chemical class found in the 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 in the chemical class with the number of reported structures in the biological family. The bar colors correspond to the chemical pathway (NPClassifier classification system) of the most specific class.

Conclusion and perpectives

As it stands, the compiled data are still imperfect and partly biased. Indeed, and as discussed, in the context of bioactive NPs research, published data tend to highlight novel structures or compounds for which an interesting bioactivity has been measured. Ubiquitous compounds are poorly documented. This gives, for the time being, a partial view of the actual metabolome of the organisms, with maybe the exception of thoroughly studied model organisms. With LOTUS and the associated WD distribution and editing possibilities, we anticipate community collaboration to correct such bias. The dissemination of referenced structure-organism pairs through WD together with harmonized data format makes it possible to query NPs research generated knowledge from a radically novel perspective. Researchers involved in NPs research and specialized metabolism should benefit from it, whether in the fields of ecology and evolution, chemical ecology, drug discovery, biosynthesis pathway elucidation, chemotaxonomy or similar research thematics. However, to incentivize community efforts, data contribution to open repositories should also be better acknowledged in academia and data re-use should be acknowledged (Cousijn et al., 2019; Cousijn et al., 2018; Pierce et al., 2019).

The possibilities for expansion and future applications of WD hosted LOTUS data are significant. For example, properly formatted spectral data (e.g. obtained by mass spectrometry or nuclear magnetic resonance) can be linked to the WD entries for chemical compounds. MassBank (Horai et al., 2010) and SPLASH

(Wohlgemuth et al., 2016) identifiers are already reported in Wikidata, and this can be used to report MassBank records for *Arabidopsis thaliana* compounds: `https://w.wiki/335H`. Such possibilities should help to bridge experimental data results obtained early in the research process to previously reported and formatted data and thus open exciting perspectives in the fields of dereplication and NPs annotation. We previously demonstrated that taxonomically-informed metabolite annotation critically improves the NPs annotation process (Rutz et al., 2019). The availability of an open repository linking chemical objects to both their spectral information and biological occurrences will facilitate and improve such applications.

As shown in Fig. S1, observing the chemical and biological diversity at various granularity levels offers clear advantages. Regarding the chemical objects it will be important to implement chemical taxonomies annotations of the entries in WD. However this is not a straightforward task and stability and coverage issues will have to be addressed. Existing chemical taxonomies such as ChEBI, ClassyFire or NPClassifier are evolving and we need to make sure tools using those annotations are updated accordingly. Repositioning NPs in their biosynthetic context is also a major challenge. The fact that LOTUS is disseminated on WD should facilitate its integration to projects such as WikiPathways and help in this complex task (Martens et al., 2021).

In the field of ecology, molecular traits are gaining increased attention (Sedio, 2017; Kessler and Kalske, 2018). Classical plant traits (e.g. leaf surface area, photosynthetic capacities, etc.) could perfectly be associated with WD biological organisms entries, and thus, allow the integration and comparison with organisms' associated chemicals. Likewise, the association of biogeography data documented in repositories such as GBIF could be further exploited in WD to pursue the exciting but understudied thematic of "chemodiverse hotspots" (Defossez et al., 2021). Other NPs related information are of great interest but very poorly formatted. For example, traditional medicine is the historical and empiric approach of mankind to encounter bioactive products from Nature. The amount of knowledge generated in our history of the use of medicinal substances represents a fascinating sum of information which could be valued and conserved in our digital era if appropriately formatted and shared (Cordell, 2017; Allard et al., 2018).

As seen, all these future developments could be accommodated in the WD knowledge base. Behind the scenes, all these resources are representing data as graphs that can be interconnected. The craft of appropriate federated queries will allow to navigate these graphs and fully exploit their potential (Waagmeester et al., 2020; Kratochvíl et al., 2018). The development of interfaces such as RDFFrames (Mohamed et al., 2020) should also facilitate the use of the wide arsenal of existing machine learning approaches to automate reasoning on these knowledge graphs.

Overall, the LOTUS project is expected to efficiently allow access to greater quality and quantity of data and ultimately pave the way towards a global open natural products database. We believe that the integration of NPs research results in such open knowledge DB can only help to fuel a *virtuous cycle of research habits aiming to better understand Life and its chemistry.*

Methods

Data collection and harmonization

Before their inclusion, source DBs overall quality was manually assessed to evaluate the quality of referenced structure-organism pairs and lack of ambiguities in the links between data and references. This led to thirty-six DBs identified 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 (https://foodb.ca/) was also curated but not publicly disseminated since its license did not allow sharing in WD. Supplementary Table S1 gives all necessary details regarding DBs access and characteristics.

Manual inspection of each DB 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 the writing of a tailored script for the extraction of relevant data and their categorization from each DB. It 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 DB, providing an initial harmonized format for each table before alignment.

For all thirty-eight DBs, if a single file or multiple files were accessible via a download option or FTP, data was collected that way. For some DB, data was scraped (cf. Supplementary Table S1). All scraping scripts written to automatically retrieve entries can be found in the lotusProcessor repository in the src/1_gathering folder (under each respective DB subfolder). Data extraction scripts for the DNP are available and should allow license owners 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 DBs (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.

Data curation

Alignment

To perform the alignment of all previously collected and harmonized DB, sixteen columns were chosen as described above. Upon DBs alignment, 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 the minimal sufficient object for WD dissemination (an unambiguously referenced structure-organism pair), the initial sixteen columns had to be translated and cleaned into three fields: the reported structure, the canonical name of the producing organism, and the reference describing the occurrence. The structure was reported as InChI, together with its SMILES and InChIKey3D 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 in 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 DBs, the following procedure was followed. Allowed structural fields for the DBs 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 OPSIN30, a fast Java-based translation open-source solution (https://github.com/dan2097/opsin). If no translation was obtained, chemical names were then submitted to the CTS31, 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 (https://cactus.nci.nih.gov) 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 are available in src/r/preparing_name.R.

The chemical sanitization step sought to standardize the representation of a collection 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 here (https://molvs.readthedocs.io/en/latest/guide/standardize.html). 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 (https://www.chemaxon.com). 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 (link).

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) (https://github.com/gnames/gnfinder) and Global Names Verify (GNV) (https://github.com/gnames/gnverify), both tools coming from the Global Names Architecture (GNA) a system of web-services which helps people to register, find, index, check and organize biological scientific names and interconnect on-line information about species (http://globalnames.org). 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 according to the parameters described in the methods. When a single entry led to multiple canonical names (accepted synonyms), all of them were kept (cf. Discussion). 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 (https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c). The script used for this was src/1_gathering/translation/common.R. When the canonical translation of a common name contained a specific epithet which 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 WD, 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 DBs are listed at http://index.globalnames.org/datasource. To retrieve as much information as possible from the original organism field of each of the DB, 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.

References

The Rcrossref package (`https://cran.r-project.org/web/packages/rcrossref/`) interfacing with the Crossref (`https://www.crossref.org`) 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 solr score (see: `https://lucene.apache.org/core/8_8_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`). 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 (`https://cran.r-project.org/web/packages/rentrez/`). Scripts used for all subcategories of references are available in the folder 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)

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 DB 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 objects translations was written for each object to avoid those translations to be 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 references translations were 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 Table S2. 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 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) (cf. Methods). 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 the DOI not coming from COCONUT. 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.

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

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 Supplementary Table S2.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, a system of unit tests and partial data full-tests were written. They can run on the developer machine but also on the CI/CD system (GitLab) for each commit in the codebase.

Those tests assess that the functions are providing results coherent with what is expected and especially for edge cases that have been detected along with 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 WD under a Creative Commons 0 license according to https://www.wikidata.org/wiki/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 publication15 i.e. the data is 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 in the “Documentation” page of the website. LNPN is part of the NaturalProducts.net portal, an initiative aiming to gather diverse open NP collections and open tools in the same place.

Data interaction

Data retrieval

Bulk retrieval of a frozen (2021-02-23) version of LOTUS data is also available at https://osf.io/hgjdb/.

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 a 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 WD. The importer processes the curated outputs grouping references, organisms and compounds together. It then checks if they already exist in WD (using SPARQL or a direct connection to WD depending on the kind of data). It then 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 utilize the Crossref REST API (https://www.crossref.org/education/retrieve-metadata/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 are dropped if 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 InChI-Key and upserts the compound with the *found in taxon* statement and the associated organisms and references.

LNPN

At the moment LNPN has been importing data directly from the frozen tabular data of the LOTUS dataset (https://osf.io/hgjdb/). Later on, LOTUS will directly feed from the SSOT.

Data edition

We adapted the bot framework WikidataLotusImporter so that, in addition to batch upload, it could also edit erroneously created entries on WD. As massive edits have a large potential to disrupt otherwise good data, we are always using a progressive deployment of this script where it starts by editing progressively 1, 10, 100 entries that are manually checked. Once we get those 100 entries validated, we run the full script and check its behavior at regular intervals.Here is an example of a corrected entry https://www.wikidata.org/w/index.php?title=Q105349871&type=revision&diff=1365519277&oldid=1356145998

Curation interface

We are currently working on a web-based (Kotlin, Spring Boot for the back-end and TypeScript with Vue for the front-end) curation interface that will allow us to mass-edit entries and navigate quickly in the SSOT to curate new or existing entries. We are thinking about making that interface open to the public so they can curate the entries of the database in yet another way. As with the rest of our approach, any modification made in this curation interface will be mirrored on WD and LNPN.

Code availability

All programs written for this work can be found in the following group: https://gitlab.com/lotus7. 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. In the first iteration, the source data corresponds to all mentioned open natural products DBs. Afterward, data uploaded to Wikidata (and thus potentially corrected) is integrated as additional source data.

The first step of the process is to check 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.

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 on Wikidata. As a first step, 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. It finally updates the chemical compound page with a “found in taxon” statement complemented with a “stated in” reference.

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 format for further use. Then, two options are possible:

The end-user can directly use the exported data.

The exported data, which can be new or modified since the last iteration is used as new source data in lotusProcessor.

The LNPN website and processing system is available at https://github.com/mSorok/LOTUSweb. This project 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 from Pr. Steinbeck and Dr. Sorokina. The website allows searching the data from different points of views, complemented with taxonomies for both on chemical and biological sides. Many chemical molecular properties and molecular descriptors not available in Wikidata are also given.

A special *preprint* branch with code at the time of publication is available.

A frozen version of the code is also available in the LOTUS OSF repository (https://osf.io/pmgux/).

R version used was 4.0.4 (2021-02-15) – “Lost Library Book”32. Packages used were, in alphabetical order:

ChemmineR (3.42.1) (Cao et al., 2008), chorddiag (0.1.2) (Flor, 2020), ClassyfireR (0.3.6) (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 (Yu et al., 2017), ggtreeExtra (1.0.1) (Xu and Yu, 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 version used was 3.8.6 (https://www.python.org/). Packages used were, in alphabetical order:

Molvs (0.1.1) (https://github.com/mcs07/MolVS), pandas (1.1.4) (Reback et al., 2020), rdkit (2020.09.2) ("RDKit: Open-source cheminformatics", n.d.)

Kotlin packages used were:

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:, WDTK 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

Data availability

This manuscript has been released as a pre-print at bioRxiv.

A snapshot of the obtained data at the time of publication is available at the following OSF repository (datasets): https://osf.io/pmgux/.

Acknowledgements

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 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 CRC1127 ChemBioSys. 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.

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	LNPN Website	NAPRALERT	Sachem/ IDSM	Wikidata
AR	X	X	X		X	X	X		X		X	X	X	X				X
CS				X				X						X	X			
DM														X				X
EW														X				X
GFP				X				X						X		X		
JB	X	X	X	X	X	X	X		X	X	X			X		X		X
JGa														X			X	
JGr														X		X		
J-LW				X				X		X				X				
JV				X				X									X	
MS									X					X	X			
P-MA	X	X	X	X	X	X	X		X	X	X		X	X				X
RP																		X
RS														X				X

Table 3: Author contributions

Competing interests

The authors declare no competing interest.

Supplementary Information

Supplementary Table S1

Available at https://gitlab.com/lotus7/lotusProcessor/-/blob/d8e4bf34761da454dac6880f0b3398bb0965e03b/docs/dataset.csv

Table S1: Natural Products databases curated within LOTUS. Commercial and restricted databases are not disseminated (except for NAPRALERT subset, in accordance with owners).

Supplementary Table S2

reference type	true positives	false positives	false negatives	true negatives	relative abundance	precision	recall	F0.5 score	true positives Validation	false positives Validation
original	80	6	7	11	0.31	0.93	0.92	0.92	26	1
pubmed	37	1	5	6	0.3	0.97	0.88	0.92	3	1
doi	115	6	0	6	0.19	0.95	1	0.97	39	1
title	38	2	0	16	0.12	0.95	1	0.97	6	0
split	8	0	15	27	0.08	1	0.35	0.52	4	0
publishingDetails	1	0	1	32	0.01	1	0.5	0.67	NA	NA
Total	279	15	28	98	1	NA	NA	NA	78	3
Corrected total	NA	NA	NA	NA	NA	0.96	0.89	0.91	NA	NA

Table S2: Summary of training and validation statistics of the database curation

Supplementary File S1

Available at https://osf.io/7dk8h/

Supplementary Figure S1

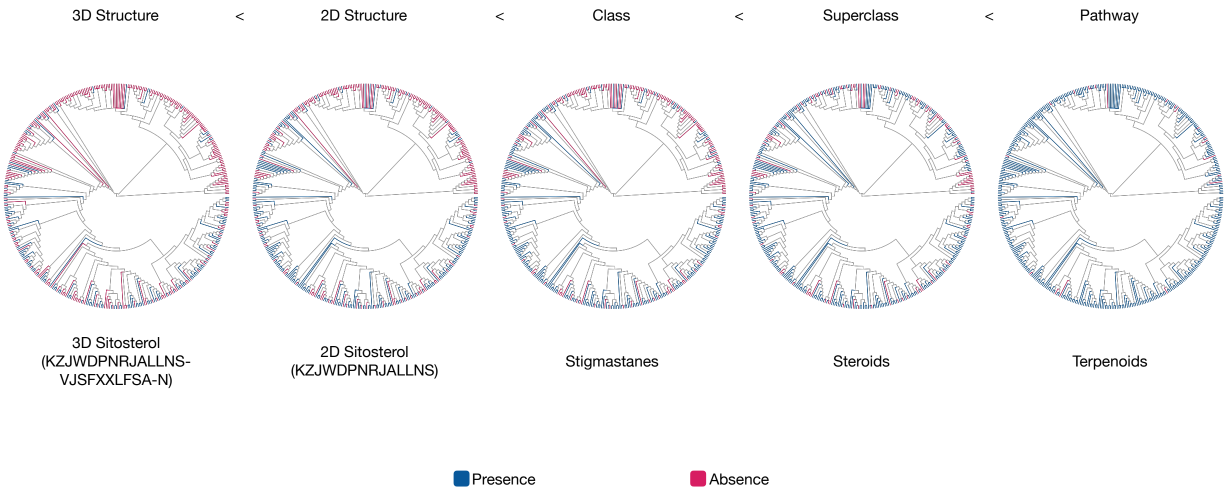

Figure S1: Complement to Figure 6

References

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

2020. . *GBIF Home Page.*

n.d.

n.d.

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 Cell Physiol* **53**:e1.

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

Allard PM, Bisson J, Azzollini A, Pauli GF, Cordell GA, Wolfender JL. 2018. Pharmacognosy in the digital era: shifting to contextualized metabolomics.. *Curr Opin Biotechnol* **54**:57–64.

Boonen J, Bronselaer A, Nielandt J, Veryser L, De TG, De SB. 2012. Alkamid database: Chemistry, occurrence and functionality of plant N-alkylamides.. *J Ethnopharmacol* **142**:563–90.

Brunson JC. 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 LC, Jiang T, Girke T. 2008. ChemmineR: A compound mining framework for R. *Bioinformatics* **24**:1733–1734. doi:10.1093/bioinformatics/btn307

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 JY, 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.. *J Cheminform* **9**:2.

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

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

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**. 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**. doi:10.1038/sdata.2018.259

Davis GD, Vasanthi AH. 2011. Seaweed metabolite database (SWMD): A database of natural compounds from marine algae.. *Bioinformation* **5**:361–4.

Defossez E, Pitteloud C, Descombes P, Glauser G, Allard PM, Walker TWN, Fernandez-Conradi P, Wolfender JL, Pellissier L, Rasmann S. 2021. Spatial and evolutionary predictability of phytochemical diversity.. *Proc Natl Acad Sci U S A* **118**.

Dowle M, Srinivasan A. 2020. data.table: Extension of ‘data.frame‘.

Duke JA. 2016. Dr. Duke's Phytochemical and Ethnobotanical Databases. doi:10.15482/USDA.ADC/1239279

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. doi:10.1101/2020.04.17.046672

Feunang YD, 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**. doi:10.1186/s13321-016-0174-y

Flor M. 2020. chorddiag: Interactive Chord Diagrams.

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

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 CompoundsComprehensive Natural Products II. Elsevier. pp. 81–94. doi:10.1016/b978-008045382-8.00060-5

Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X. 2013. Use of natural products as chemical library for drug discovery and network pharmacology.. *PLoS One* **8**:e62839.

Günthardt BF, Hollender J, Hungerbühler K, Scheringer M, Bucheli TD. 2018. Comprehensive Toxic Plants-Phytotoxins Database and Its Application in Assessing Aquatic Micropollution Potential.. *J Agric Food Chem* **66**:7577–7588.

Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, Tastan BÖ. 2015. SANCDB: a South African natural compound database.. *J Cheminform* **7**:29.

Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C. 2020. MetaboLights: a resource evolving in response to the needs of its scientific community.. *Nucleic Acids Res* **48**:D440–D444.

Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. 2013. InChI - the worldwide chemical structure identifier standard.. *J Cheminform* **5**:7.

Helmy M, Crits-Christoph A, Bader GD. 2016. Ten Simple Rules for Developing Public Biological Databases.. *PLoS Comput Biol* **12**:e1005128.

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.. *J Mass Spectrom* **45**:703–14.

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 Res* **46**:D575–D580.

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. 2020. Comprehensive database of secondary metabolites from cyanobacteria. doi:10.1101/2020.04.16.038703

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

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der HJJJ, van SJA, Tracanna V, Suarez DHG, Pascal AV, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function.. *Nucleic Acids Res* **48**:D454–D458.

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 HW, Wang M, Leber CA, Nothias L-félix. n.d. NPClassifier : deep neural structural classification tool for natural products.

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. 2018. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**:D1102–D1109. doi:10.1093/nar/gky1033

Kim SK, Nam S, Jang H, Kim A, Lee JJ. 2015. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine.. *BMC Complement Altern Med* **15**:218.

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 Res* **44**:D509–14.

Kratochvíl M, Vondrášek J, Galgonek J. 2018. Sachem: a chemical cartridge for high-performance substructure search.. *J Cheminform* **10**:27.

Kratochvíl M, Vondrášek J, Galgonek J. 2019. Interoperable chemical structure search service.. *J Cheminform* **11**:45.

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 GM, 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.. *Sci Data* **7**:144.

Loo MPJvan der. 2014. The stringdist package for approximate string matching. *The R Journal* **6**:111–122. doi:10.32614/RJ-2014-011

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

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

Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A MR, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. 2021. WikiPathways: connecting communities.. *Nucleic Acids Res* **49**:D613–D621.

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/f9jgkm

Mohamed A, Abuoda G, Ghanem A, Kaoudi Z, Aboulnaga A. 2020. RDFFrames: Knowledge Graph Access for Machine Learning Tools. *arXiv:200203614 [cs]*.

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

Nielsen FÅ, Mietchen D, Willighagen E. 2017. Scholia, Scientometrics and Wikidata In: Blomqvist E, Hose K, Paulheim H, Ławrynowicz A, Ciravegna F, Hartig O, editors. The Semantic Web: ESWC 2017 Satellite Events. Cham: Springer International Publishing. pp. 237–259.

Noteborn HP, Lommen A, van der JRC, Weseman JM. 2000. Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops.. *J Biotechnol* **77**:103–14.

Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, A MAF, Malange YI, Njume LE, Yong JN, Sippl W, Günther S. 2017. NANPDB: A Resource for Natural Products from Northern African Sources.. *J Nat Prod* **80**:2067–2076.

Nupur LN, Vats A, Dhanda SK, Raghava GP, Pinnaka AK, Kumar A. 2016. ProCarDB: a database of bacterial carotenoids.. *BMC Microbiol* **16**:96.

Olivon F, Allard PM, Koval A, Righi D, Genta-Jouve G, Neyts J, Apel C, Pannecouque C, Nothias LF, Cachet X, Marcourt L, Roussi F, Katanaev VL, Touboul D, Wolfender JL, Litaudon M. 2017. Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks.. *ACS Chem Biol* **12**:2644–2651.

Ooms J. 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:14032805 [statCO].*

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

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

Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS. 2017. NuBBE¡sub¿DB¡/sub¿: an updated database to uncover chemical and biological information from Brazilian biodiversity.. *Sci Rep* **7**:7215.

Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. 2019. BIOFACQUIM: A Mexican Compound Database of Natural Products.. *Biomolecules* **9**.

Reback J, McKinney W, jbrockmendel, Bossche JVden, Augspurger T, Cloud P, gfyoung, Sinhrks, Hawkins S, Roeschke M, Klein A, Petersen T, Tratner J, She C, Ayd W, Naveh S, Garcia M, Schendel J, Hayden A, Saxton D, Jancauskas V, McMaster A, Battiston P, Seabold S, chris-b1, h-vetinari, Dong K, Hoyer S, Overmeire W, Gorelli M. 2020. pandas-dev/pandas: Pandas 1.1.4. 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-Remón A, M'hiri N, García-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 (Oxford)* **2013**:bat070.

Rutz A, Dounoue-Kubo M, Ollivier S, Bisson J, Bagheri M, Saesong T, Ebrahimi SN, Ingkaninan K, Wolfender JL, Allard PM. 2019. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation.. *Front Plant Sci* **10**:1329.

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.

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 Phytol* **214**:952–958.

Sharma A, Dutta P, Sharma M, Rajput NK, Dodiya B, Georrge JJ, Kholia T, Bhardwaj A. 2014. BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts.. *J Cheminform* **6**:46.

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D, Kanaya S. n.d. KNApSAcK: A Comprehensive Species-Metabolite Relationship DatabasePlant Metabolomics. Springer-Verlag. pp. 165–181. doi:10.1007/3-540-29782-$0_1$3

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

Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. 2021. COCONUT online: Collection of Open Natural Products database.. *J Cheminform* **13**:2.

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

Sorokina M, Steinbeck C. 2020. Review on natural products databases: where to find data in 2020.. *J Cheminform* **12**:20.

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

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**:1–17. doi:10.18637/jss.v093.i13

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

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

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**.

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/ggwr93

Wang S, Alseekh S, Fernie AR, Luo J. 2019. The Structure and Function of Major Plant Metabolite Modifications.. *Mol Plant* **12**:899–919.

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 LDA, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, 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 JW, da SSLB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't HPA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van SR, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der LJ, van ME, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship.. *Sci Data* **3**:160018.

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.. *J Cheminform* **9**:33.

Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API. *The R Journal* **9**:520–526. doi:10.32614/RJ-2017-058

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.. *Nat Biotechnol* **34**:1099–1101.

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

Xu S, Yu G. 2021. ggtreeExtra: An r package to add geom layers on circular or other layout tree of ggtree (manual).

Yabuzaki J. 2017. Carotenoids Database: structures, chemical fingerprints and distribution among organisms.. *Database (Oxford)* **2017**.

Yu G, Smith D, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an R 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/f9qv8x

Yue Y, Chu GX, Liu XS, Tang X, Wang W, Liu GJ, Yang T, Ling TJ, Wang XG, Zhang ZZ, Xia T, Wan XC, Bao GH. 2014. TMDB: a literature-curated database for small molecular compounds found from tea.. *BMC Plant Biol* **14**:243.

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 Res* **46**:D1217–D1222.

Zhang R, Lin J, Zou Y, Zhang XJ, Xiao WL. 2019. Chemical Space and Biological Target Network of Anti-Inflammatory Natural Products.. *J Chem Inf Model* **59**:66–73.

van SJA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Neto FC, Castaño-Espriu L, Chang C, Clark TN, Cleary LJL, Delgadillo DA, Dorrestein PC, Duncan KR, Egan JM, Galey MM, Haeckl FPJ, Hua A, Hughes AH, Iskakova D, Khadilkar A, Lee JH, 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 HJJJ, 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 Cent Sci* **5**:1824–1833.

Derese, Solomon, Ndakala, Albert, Rogo, Michael, Maynim, Cholastica, Oyim, James. 2015. Mitishamba database: a web based in silico database of natural products from Kenya plants.

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