Terpenes, Evidence, and Infrastructure

How Terpedia turns fragmented terpene science into data industry can use

Terpedia, LLC

Abstract

Terpenes are the largest family of natural products and the working vocabulary of the cannabis, hemp, flavor, fragrance, essential-oil, functional-food, and natural-product-pharmaceutical industries. The information needed to make terpene decisions is scattered across dozens of databases that do not share identifiers, names, licenses, or standards of evidence, and laboratory terpene profiles vary more by laboratory than by product. Terpedia is a provenance-first knowledge platform that reconciles more than thirty sources on structural chemical identity, attaches evidence type and license to every statement, and exposes the result through a browsable encyclopedia, a documented API, domain agents, and a product-and-certificate catalog. This paper explains the chemistry, the commercial landscape and its evidence, the data problem, the platform, and concrete industry workflows, with runnable examples executed against the live Terpedia Knowledge API.

Keywords terpenes, terpenoids, knowledge graph, provenance, certificate of analysis, cannabis, flavor and fragrance, essential oils

Published Sep 10, 2026

1. Executive summary

Terpenes are the largest family of natural products and the working vocabulary of several industries at once. They are the volatile signature of a cannabis cultivar, the top notes of a fragrance, the citrus in a beverage, the active principle in a chest rub, and the scaffold of two of the most important drugs of the last century. Product decisions worth billions of dollars a year turn on which terpenes are present, at what levels, and what can truthfully be said about them.

The information needed to make those decisions is scattered across dozens of public chemistry, biology, and literature databases that do not share identifiers, naming conventions, licenses, or standards of evidence. A single molecule can carry fifty synonyms. Cultivar names carry almost no chemical information. Laboratory terpene profiles vary more by which lab produced them than by what they claim to measure. Promotional “effects” propagate through the market faster than the evidence behind them. Anyone who has tried to build a terpene-aware product, quality program, or content library knows the cost: weeks of manual reconciliation, brittle spreadsheets, and claims that cannot be defended when a regulator or a customer asks for the source.

Terpedia is a knowledge platform built to remove that cost. It ingests more than thirty public and licensed sources into a single provenance-preserving knowledge base, resolves names to chemical identities, links compounds to organisms, proteins, products, measurements, and literature, and exposes the result through a browsable encyclopedia (kb.terpedia.com), a documented API, a suite of domain agents (chat.terpedia.com), a product-and-certificate-of-analysis catalog (Terproduct), and a research program that publishes its methods.

Terpedia's differentiator is not that it has the most records. It is that every record carries its source, its version, its license, and its evidence type, and that the platform is engineered to keep claims smaller than the data. A database record is not a validated natural product. A receptor association is not efficacy. A docking score is not affinity. A cultivar name is not a chemistry. Those distinctions are enforced in the data model and surface in every API response, which is what makes the output usable in regulated settings.

1.a. What this paper covers

1. **The chemistry.** What terpenes are, how plants make them, why there are so many, and which physical properties matter for products.
2. **The business landscape.** Where terpenes create value in cannabis and hemp, flavor and fragrance, essential oils and aromatherapy, functional foods, pharmaceuticals, and industrial biotechnology, and what the evidence actually supports.
3. **The data problem.** Fragmentation, identity, counting, measurement variability, claims inflation, and licensing.
4. **The platform.** Terpedia's architecture, sources, scale, access surfaces, partnerships, and research portfolio.
5. **Live examples.** Runnable queries against the public Terpedia Knowledge API, executed when this document was built and re-runnable in the browser.
6. **Use cases.** Concrete workflows for formulation, quality, regulatory, content, sourcing, IP, and data science teams.
7. **Evidence rules and limitations.** What Terpedia will and will not assert, and where coverage is still incomplete.
8. **Getting started.** Integration paths and the roadmap.

1.b. Key numbers

All counts are dated, scoped observations from Terpedia's own counting reports, not marketing totals; the scope and identity key for each is given in Scale, stated carefully.

Measure	Value (September 2026)
Candidate terpene/terpenoid identities in the two-source (COCONUT + TeroKit) census union	268,924, of which 225,905 have an exact PubChem CID match
Source-declared terpenoid identities from COCONUT	199,234 unique identities

Confirmed terpene/terpenoid identities under the TeroKit classification	145,354 unique identities
LOTUS occurrence triples (compound reported in organism, each with a DOI)	674,422 triples over 227,316 structures, 37,468 organisms, 91,379 papers
Normalized cannabis laboratory terpene measurements	409,655 non-null measurements from 43,018 source rows
CannabisDatabase.ca records and relations	12,039 source records; 54,251 relation records
Public sources in the knowledge base catalog	More than 30 (see Appendix A: Source inventory)

1.c. Who should read this

This paper is written for the people who have to make terpene decisions stick: R&D and formulation leads, quality and laboratory managers, regulatory and compliance officers, product and content teams, sourcing managers, and the data and AI groups that support them. It assumes technical literacy but not a chemistry degree. Sections 1 and 2 can be read as a primer; Sections 3 to 7 are the case for Terpedia; the live-examples section is for anyone who wants to see the data rather than read about it.

2. What terpenes are

2.a. Five-carbon bricks

Every terpene is assembled from the same five-carbon unit. Two activated isoprene equivalents, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are joined head-to-tail by prenyltransferases into linear chains of ten, fifteen, twenty, or more carbons. Those chains are then folded, cyclized, and rearranged by a family of enzymes called terpene synthases into the thousands of ring systems that give the class its diversity. The carbon count fixes the top-level vocabulary:

Class	Carbons	Familiar examples
Hemiterpenes	C5	Isoprene; prenol
Monoterpenes	C10	Limonene, myrcene, α- and β-pinene, linalool, geraniol, 1,8-cineole (eucalyptol), menthol, camphor, thujone
Sesquiterpenes	C15	β-Caryophyllene, α-humulene, farnesene, valencene, nootkatone, bisabolol, nerolidol, artemisinin
Diterpenes	C20	Paclitaxel (Taxol), retinol precursors, gibberellins, phytol, cannabis-relevant sclareol
Triterpenes	C30	Squalene, sterol precursors, ursolic acid, glycyrrhizin
Tetraterpenes	C40	Carotenoids: β-carotene, lycopene, lutein

In strict usage a **terpene** is a hydrocarbon (only carbon and hydrogen) and a **terpenoid** is a terpene that has been oxidized, rearranged, or otherwise functionalized. Linalool, an alcohol, is technically a monoterpenoid; limonene, a hydrocarbon, is a true terpene. Industry and most of the literature use "terpenes" for both, and so does this paper except where the distinction changes a count or a regulatory status.

2.b. How plants make them

Plants run two independent routes to the C5 precursors (Christianson, 2017; Pichersky & Raguso, 2018). The **mevalonate (MVA) pathway** in the cytosol supplies precursors mainly for sesquiterpenes (C15) and triterpenes (C30). The **methylerythritol phosphate (MEP) pathway** in the plastids supplies precursors mainly for monoterpenes (C10), diterpenes (C20), and carotenoids (C40). Prenyltransferases condense IPP and DMAPP into geranyl diphosphate (GPP, C10), farnesyl diphosphate (FPP, C15), and geranylgeranyl diphosphate (GGPP, C20). Terpene synthases then act on those diphosphates.

Terpene synthases are the source of the class's astonishing diversity, and of much of the confusion around it. A single synthase frequently yields a dozen or more products from one substrate, because the carbocation chemistry it initiates can branch at several points before the reaction is quenched (Christianson, 2017). Small changes in the enzyme's active site change the product spectrum. Plants also carry large, rapidly evolving synthase gene families; *Cannabis sativa* alone has dozens of terpene synthase genes, and its inflorescence has been reported to accumulate well over a

hundred distinct terpenes (Booth & Bohlmann, 2019). Downstream cytochrome P450s, dehydrogenases, and transferases decorate the skeletons further. The ecological logic is that a cheap, combinatorial system for making many volatile and semi-volatile molecules lets plants signal to pollinators, deter herbivores, recruit the enemies of herbivores, and protect tissues from oxidative and thermal stress (Pichersky & Raguso, 2018).

For a product team the practical lessons are these. First, the terpene profile of a plant is a *population* of molecules produced by a *network* of enzymes, so it varies with genotype, tissue, development, environment, and post-harvest handling. Second, minor constituents are not noise; a synthase's side products are real, reproducible, and often odor-active. Third, "the same terpene" from two botanical sources may differ in enantiomer ratio, and enantiomers can smell and behave differently.

2.c. How many terpenes are there?

The honest answer is that there is no defensible single number without a counting rule (Terpedia, LLC, 2026a). Reviews have variously reported more than 20,000, more than 30,000, roughly 55,000, and more than 80,000 known terpenes or terpenoids. These figures count different chemical universes: some include only plant compounds, some include stereoisomers as separate entries, some include glycosides and conjugates, some count database records rather than characterized structures, and some quietly conflate terpenes with terpenoids.

Terpedia's own census takes a different approach. Rather than asserting a total, it unions the source-declared terpenoid records of two independent open databases, COCONUT (Sorokina et al., 2021) and TeroKit (Zeng et al., 2020), on exact standard InChI, and reports the result as a dated, scoped tuple: **268,924 candidate terpene/terpenoid identities** in the September 2026 snapshot, of which 225,905 have an exact PubChem CID match. An independent audit against the ChEBI ontology confirms 6,585 of those by exact InChIKey, and adding 54 directly characterized terpene-synthase products from the MARTS database brings the ontology-confirmed set to 6,605 (Terpedia, LLC, 2026a). Those are candidate identities, not independently validated natural terpenes. The gap between the 268,924 candidates and the 6,605 ontology-confirmed identities is itself informative: it is the amount of terpene chemistry that exists in databases but has not yet been organized into a curated classification.

The point for industry is not the number. It is that a supplier, a lab, a regulator, and a marketing team who each quote a different "number of known terpenes" are not disagreeing about facts; they are using different counting rules, and a platform that wants to be trusted has to make the rule explicit.

2.d. Where terpenes occur

Terpenes are produced by plants, fungi, bacteria, marine organisms, and insects. Commercially, the plant kingdom dominates:

- **Conifers** supply turpentine and its fractions (α- and β-pinene, camphene, 3-carene), the historical bulk source of monoterpene feedstocks.

- **Citrus** supplies d-limonene at industrial scale as a by-product of juice production, together with valencene, γ-terpinene, and citral.
- **Lamiaceae herbs** (mint, lavender, rosemary, thyme, oregano, basil, sage) supply menthol, linalool, linalyl acetate, 1,8-cineole, camphor, thymol, carvacrol, and thujone.
- **Myrtaceae** (eucalyptus, tea tree, clove, allspice) supply 1,8-cineole, terpinen-4-ol, and β-caryophyllene.
- **Hops and cannabis** share a characteristic profile dominated by myrcene, β-caryophyllene, α-humulene, and limonene, with linalool, α-pinene, β-pinene, ocimene, terpinolene, and nerolidol as frequent secondary constituents (Booth & Bohlmann, 2019).
- **Spices** such as black pepper, cinnamon, cardamom, ginger, and nutmeg supply β-caryophyllene, sabinene, zingiberene, and myristicin-adjacent aromatics.

Occurrence is a *report*, not a *concentration*. The LOTUS initiative, an open compound-organism occurrence resource that Terpedia ingests, records which molecules have been reported in which organisms and cites the paper that reported each pair (Rutz et al., 2022). It is invaluable for asking "where else has this molecule been found?" but it does not tell you that an organism is a meaningful commercial source. Terpedia keeps that distinction explicit in its data model.

2.e. Physical chemistry that matters for products

Four properties govern how terpenes behave in a product and how reliably they can be measured.

Volatility. Monoterpenes have boiling points around 150–190 °C and appreciable vapor pressure at room temperature; sesquiterpenes are heavier and less volatile. This is why terpene profiles change during drying, curing, storage, and heating, why headspace and vapor composition differ from the bulk, and why a certificate of analysis is a snapshot of a moment rather than a property of a product.

Lipophilicity. Terpenes are hydrophobic. Limonene's measured logP is about 4.4 and it is practically insoluble in water. This drives their partitioning into oils, membranes, and packaging, their need for emulsification in aqueous products, and their tendency to be lost to plastics.

Oxidation. Unsaturated terpenes autoxidize on exposure to air and light. For limonene and linalool the primary oxidation products are allylic hydroperoxides, which are potent skin sensitizers even though the parent molecules are weak or non-sensitizing (Dittmar & Schuttelaar, 2019). Oxidative aging also changes aroma. Stability, antioxidant systems, and storage conditions are therefore quality and safety questions, not just shelf-life questions.

Stereochemistry. Many terpenes are chiral, and enantiomers can differ in odor and biology. (R)-(+)-limonene is the orange note; (S)-(-)-limonene reads as pine or turpentine. Analytical methods that do not resolve enantiomers, and databases that do not distinguish them, lose information that a perfumer or a flavorist relies on.

Each of these properties creates a specific data-management requirement: stereo-aware identity, time-stamped measurements, oxidation-product tracking, and matrix-aware concentration units. Those requirements are the reason Terpedia's data model is built around exact chemical identity and provenance rather than around names.

3. Why terpenes matter to industry

Terpenes are one of the few chemical families that sit at the center of several distinct industries at once. This section surveys where the value is, what the evidence actually supports, and what is at stake when the data behind a terpene decision is wrong.

3.a. Market context

Market-research estimates for the standalone terpenes market vary widely with scope. Published 2025 valuations range from roughly USD 0.7 billion to USD 1.6 billion, with forecasts approaching USD 3 billion by 2035 (Market.us, 2025; Precedence Research, 2025); flavors and fragrances are consistently reported as the largest application segment. The adjacent essential-oils market is estimated at roughly USD 10–14 billion in 2025 (IMARC Group, 2025), and aromatherapy at about USD 10 billion (Grand View Research, 2025). These are third-party estimates with different methodologies and should be read as an order of magnitude rather than as a precise size. The more useful observation is structural: terpenes are a comparatively small ingredient market that gates decisions in several very large product markets (beverages, personal care, home care, cannabis, supplements, pharmaceuticals). Their information value exceeds their ingredient value.

3.b. Flavor and fragrance

Terpenes are the backbone of natural aroma chemistry. Citrus, pine, mint, floral, herbal, spicy, and woody notes are largely terpene-driven, and terpene-derived molecules (menthol, linalool, geraniol, citral, citronellol, nerol, and their esters) are among the highest-volume flavor and fragrance ingredients in the world. Myrcene is not used much directly but is the industrial intermediate for many of them.

Safety and regulatory status in this sector is comparatively mature. The Flavor and Extract Manufacturers Association (FEMA) Expert Panel has evaluated aliphatic and aromatic terpene hydrocarbons and reaffirmed their “generally recognized as safe” (GRAS) status for use as flavoring substances in food, based in part on their self-limiting organoleptic properties (Adams et al., 2011). On the fragrance side, the picture is more nuanced: the autoxidation products of limonene and linalool are among the most frequently detected contact allergens in dermatology clinics, with one consecutive patch-test series reporting positive reactions to limonene hydroperoxides in 9.4% and to linalool hydroperoxides in 11.7% of patients tested (Dittmar & Schuttelaar, 2019). European cosmetic regulation has responded with labeling requirements and ongoing scientific-committee review. For a formulator, that means the *oxidation state* of a terpene ingredient is a compliance variable, not just a quality variable.

The data requirement here is identity and stereochemistry at the level a perfumer works at, together with regulatory status (FEMA number, GRAS determination, EU flavouring or allergen listing, IFRA standard) attached to the right isomer.

3.c. Cannabis and hemp

Terpenes are how cannabis products differentiate. Cannabinoid content is largely commoditized; the aroma and the perceived character of a cultivar are terpene-driven, and consumers, budtenders, and brands increasingly talk in terpene language. Regulators have followed. California limits the terpenes that may be added to inhalable cannabis products to those naturally occurring in cannabis that contribute to its natural flavor and aroma (California Department of Cannabis Control, 2023), and Oregon's review of non-cannabis additives stresses that natural occurrence or food-GRAS status does not establish safety for inhalation (Oregon Liquor and Cannabis Commission, 2020). Terpene profiles on certificates of analysis are now a routine part of product release in most regulated markets.

The scientific story is more careful than the marketing story, and the gap matters commercially. The influential 2011 review by Russo proposed that terpenes could modulate the effects of cannabinoids, the so-called entourage effect (Russo, 2011). Fifteen years later, the evidence is mixed. Preclinical work has shown that several cannabis terpenes are cannabimimetic in mice and can enhance cannabinoid activity in a receptor-dependent way (LaVigne et al., 2021). β-Caryophyllene is a genuine, selective CB2 receptor agonist (Ki $\approx$ 155 nM) and is a common dietary constituent, which makes it the best-characterized "dietary cannabinoid" in the class (Gertsch et al., 2008). In humans, a 2024 double-blind, placebo-controlled crossover trial in twenty healthy adults found that vaporized d-limonene dose-dependently reduced the anxiety and paranoia induced by inhaled THC without altering THC's other subjective effects (Spindle et al., 2024). That is a real, specific, clinically measured interaction. It is not evidence that every terpene enhances every cannabinoid. A comprehensive 2024 review concluded that synergistic or additive enhancement of cannabinoid efficacy by terpenes "remains unproven" and called for clinical trials (André et al., 2024).

For a brand, the implication is that "myrcene is sedating" and "limonene is uplifting" are hypotheses with some preclinical support, not substantiated claims, and that the regulatory and reputational risk of stating them as facts is rising. For a lab or a producer, the implication is that the terpene profile itself, measured well and reported honestly, is the defensible asset.

3.d. Essential oils and aromatherapy

Essential oils are, chemically, terpene mixtures with a minority of phenylpropanoids and other volatiles. The sector is large, consumer-facing, and lightly regulated, which makes it both an opportunity and a liability. There is credible mechanistic work: linalool vapor produces anxiolytic behavior in mice that is abolished in anosmic animals and blocked by the benzodiazepine antagonist flumazenil, indicating an olfactory route to $GABA_A$-mediated effects (Harada et al., 2018). There is also a long history of claims that outrun evidence, and enforcement actions against disease claims on essential-oil products in several jurisdictions.

The data requirement is composition: what is actually in this oil, at what percentage, from which botanical and geographic source, and how variable is that across batches. Terpedia ingests EssoilDB and related composition data precisely so that "lavender

oil" can be resolved into a distribution of linalool, linalyl acetate, camphor, and 1,8-cineole rather than a name.

3.e. Functional foods and beverages

The fastest-growing terpene application is arguably the one with the least settled vocabulary: foods and beverages that carry terpenes for both flavor and an implied function. Hemp-derived terpene blends in chews, seltzers, and gummies; citrus terpenes positioned for mood; hop terpenes in non-alcoholic beer. The regulatory line between a flavor and a structure/function claim is sharp in the United States and the European Union, and the substantiation standard for the latter is high.

Terpedia's partnership with MONDAYS (described in Partnership example: MONDAYS) is a worked example of doing this responsibly: each product page publishes the measured terpene profile of the product, every compound at or above 0.05% of total volatiles with the milligrams it contributes to a 20 mg terpene dose, and links each molecule to the underlying science on a separate reference site where laboratory-dose findings are labeled as such. Composition is stated; effects are not.

3.f. Pharmaceuticals

Terpenoids are among the most successful natural-product scaffolds in medicine. Artemisinin, a sesquiterpene lactone from *Artemisia annua*, and its derivatives are the backbone of modern antimalarial therapy. Paclitaxel, a diterpenoid originally isolated from *Taxus brevifolia*, is a first-line anticancer agent. Both illustrate the drug-discovery value of the class and its formulation challenges: poor water solubility, restricted bioavailability, and the need for semisynthetic optimization.

β-Caryophyllene's selective CB2 agonism (Gertsch et al., 2008) has made it a candidate for inflammatory and neuropathic pain and is the subject of ongoing preclinical and early clinical work. More broadly, the terpene chemical space is enormous, largely uncharacterized biologically, and increasingly accessible to computational target prediction. That is the space Terpedia's research program on protein docking and terpene–protein interaction ranking is designed to organize (see Research portfolio).

3.g. Industrial biotechnology

Engineered microbes now produce terpenes at commercial scale. Amyris demonstrated yeast-based production of the sesquiterpene farnesene from sugar feedstocks; valencene, historically distilled from citrus oil in low yield and at the mercy of seasonal supply, is produced in engineered *Saccharomyces cerevisiae* and converted to the high-value grapefruit aroma nootkatone (Song et al., 2024). Global valencene production has been estimated at roughly 10,000 kg per year, about half of it consumed in nootkatone manufacture (Song et al., 2024). As biosynthetic routes proliferate, the questions a buyer must answer change: is this molecule "natural" under the relevant flavor regulation, what is its isomer purity, and what trace by-products does the production host contribute? Those are, again, identity and provenance questions.

3.h. Agriculture and pest management

Many terpenes are insect repellents, attractants, or antifeedants in their native ecological roles, and several are registered biopesticide active ingredients. Nootkatone is registered in the United States as an insect repellent. Limonene, pine oil, and thyme oil appear in minimum-risk pesticide formulations. This sector is small relative to the others but is growing as synthetic-pesticide restrictions tighten, and it draws on the same occurrence, composition, and safety data.

3.i. What is at stake

Across all of these sectors, the cost of bad terpene data falls into four buckets:

1. **Formulation errors** when a supplier's name, a lab's name, and a regulatory list's name refer to different isomers or different compounds.
2. **Quality failures** when batch-to-batch or lab-to-lab variability is misread as product variability or, worse, hidden.
3. **Compliance exposure** when marketing copy states effects that the literature treats as hypotheses, or when an additive that is food-GRAS is assumed to be inhalation-safe.
4. **Wasted expert time** reconciling sources by hand: the same reconciliation done again in every company, every quarter.

The next section explains why the public data landscape makes these failures so easy, and the section after it explains what Terpedia does about it.

4. The data problem

The scientific and regulatory information about terpenes is, in aggregate, rich. In practice it is nearly unusable without heavy engineering, because it is fragmented across sources that were never designed to work together. This section names the specific failure modes. Each of them corresponds to a design decision in Terpedia described in the next section.

4.a. Fragmentation

A team that wants a complete picture of a single terpene must consult, at a minimum:

- **Chemical registries** for identity and structure: PubChem, ChEBI, ChEMBL, the FDA's UNII substance registry, CAS.
- **Natural-product databases** for occurrence and classification: COCONUT, LOTUS, TeroKit, KNApSAcK, SuperNatural, Dr. Duke's phytochemical database.
- **Biochemical databases** for enzymes and reactions: Rhea, BRENDA, UniProt.
- **Composition databases** for essential oils and cannabis: EssoilDB, CannabisDatabase.ca, commercial and state laboratory result sets.
- **Assay and target data**: PubChem BioAssay, ChEMBL, and terpene–protein interaction sets.
- **Regulatory lists**: FEMA GRAS, FDA SCOGS and food-additive listings, EU flavouring and cosmetic-allergen annexes, state cannabis rules.
- **Literature and patents**: PubMed, patent full-text indexes, and grey literature such as monographs and ethnobotanical compendia.
- **Ontologies** for biology and disease: MeSH, MONDO, NCBI Taxonomy, Cellosaurus.

Each has its own identifier scheme, update cadence, file format, and license. None of them is terpene-focused. Several of them disagree with one another about names, stereochemistry, and even which compounds count as terpenoids.

4.b. Identity

Names do not identify molecules. The Terpedia record for limonene carries **58 aliases**, including (+)-limonene, (R)-(+)-p-mentha-1,8-diene, d-limonene, dipentene, and the FEMA and CAS designations. Certificates of analysis, ingredient declarations, supplier specifications, and papers each use whichever name their authors were trained on. Stereochemistry is often dropped: "limonene" on a COA might be the (R)-enantiomer, the (S)-enantiomer, or an unresolved mixture, and only enantioselective chromatography can tell. Salts, hydrates, glycosides, and esters of the same core appear as separate entries in some databases and are folded together in others.

The only robust identity key is a structure-based one: the standard InChI and its hashed InChIKey. Terpedia unions its sources on exact standard InChI and reports coverage in those terms (Terpedia, LLC, 2026a). Anything less produces counts and joins that silently mix compounds.

4.c. Counting

Because identity is unstable, counts are unstable. As How many terpenes are there? described, published totals for "known terpenes" span an order of magnitude, and even a single well-curated database can be counted several ways. In Terpedia's own counting report, the same TeroKit source yields 145,354 confirmed identities under its explicit terpene/terpenoid categories and a further 23,061 in categories labeled "Others" and "Steroids" that are deliberately excluded as ambiguous (Terpedia, LLC, 2026b). A general natural-product table is not a terpene table just because some of its entries are terpenes. Any vendor who quotes a single large "compounds in our database" figure without a scope is asking you to trust an unstated counting rule.

4.d. Measurement variability

Laboratory terpene profiles are the most commercially important terpene data and the least reliable. Terpenes are volatile, oxidize on standing, and are lost to sample preparation; laboratories differ in extraction solvent, instrument (GC-MS versus GC-FID), calibration standards, reporting thresholds, and the list of analytes they quantify. Inter-laboratory comparisons in the cannabis sector routinely show divergent results on split samples, and proficiency-testing programs remain immature relative to established food and pharmaceutical testing.

Terpedia's own research quantifies how bad this is. In a provenance audit of a commercial cannabis terpene archive, the identity of the source laboratory database was more strongly associated with a sample's assigned terpene profile class (normalized mutual information 0.324) than the cultivar's commercial name was (normalized mutual information 0.112), with two laboratories' profiles mapping almost entirely to a single class (McShan & Trapp, 2026a). In plain terms: knowing which lab produced a profile told you more about the profile than knowing which cultivar it claimed to describe. This finding led Terpedia to downgrade its own earlier seven-class "terpotype" classifier from a biological claim to a provisional archive partition (McShan & Trapp, 2026b), an example of the platform applying its evidence rules to itself.

The practical consequences for industry are direct. Cultivar names are not a chemistry. Reference profiles must be built from many samples and must carry laboratory identity as a covariate. And a product's terpene claims can only be as good as the measurement pipeline behind them.

4.e. Claims inflation

Effect claims about terpenes propagate through the market by repetition. A receptor-binding result in a cell assay becomes "activates CB2"; that becomes "anti-inflammatory"; that becomes label copy. Each step drops a qualifier. Terpedia's claims investigation catalogs this drift explicitly: it treats every promotional effect statement as a hypothesis, joins it to whatever receptor or assay evidence exists, and reports two separate fields, whether the compound has any linked mechanism evidence and whether the effect itself is supported by appropriately scoped studies (Terpedia, LLC, 2026c). A compound can have the first without the second. Most do.

The failure mode is not dishonesty; it is the absence of a data model that distinguishes evidence types. When "associated with CB2" and "clinically shown to reduce anxiety" are stored in the same free-text field, the distinction is gone by the time the content team sees it.

4.f. Provenance and licensing

Public databases carry heterogeneous licenses. Wikidata is CC0. COCONUT and LOTUS are open. CannabisDatabase.ca is CC BY-NC 4.0, which restricts commercial reuse. BRENDA is license-scoped. Some regulatory and monograph sources are redistributable only as documents. A pipeline that merges these into one undifferentiated table has destroyed the information a legal team needs to approve commercial use. It has also destroyed the information a scientist needs to judge reliability: a source-curated association from a community database and a reviewed UniProt entry are not the same grade of evidence.

Provenance also has a time dimension. Sources refresh; records change; a number that was correct in June may be wrong in September. Without release identifiers, retrieval timestamps, and content hashes on every row, there is no way to reproduce a result or to know whether a discrepancy is a data error or a data update.

4.g. The cost of doing it yourself

Every organization that works seriously with terpenes has, at some point, built an internal spreadsheet or database that tries to solve the problems above. These efforts share a life cycle: they are built by one motivated scientist, they encode that person's naming conventions, they fall out of date when sources refresh, and they cannot be defended when a regulator or a customer asks for the source of a specific value. The work is not differentiating; it is the same reconciliation done again in every company.

Commercial chemistry platforms exist but are priced and designed for pharmaceutical discovery, are not terpene-focused, do not carry cannabis or essential-oil composition data, and do not link to product-level certificates of analysis. The gap between "general chemistry database" and "usable terpene knowledge for a product team" is exactly the gap Terpedia fills.

5. Terpedia: an evidence-first terpene knowledge platform

Terpedia, LLC is a Colorado company founded by a terpene biochemist and a knowledge-systems engineer to build the reference infrastructure the terpene economy lacks. The platform has three layers: a **knowledge base** that ingests and reconciles sources with provenance intact; a set of **access surfaces** (encyclopedia, API, agents, product catalog) that put that knowledge where people and systems work; and a **research program** that tests the platform's own claims and publishes the results.

5.a. Design principles

Four principles govern every part of the system.

1. **Identity is structural.** Compounds are keyed on standard InChI and InChIKey; names, synonyms, and source identifiers are attributes of the identity, never the key.
2. **Provenance travels with the record.** Every promoted row carries, where applicable, its source release, source file URI, manifest URI, content hash, ingestion run identifier, load timestamp, and license. Every statement in the semantic graph carries its dataset, retrieval time, evidence type, and (where available) PubMed identifier or DOI.
3. **Evidence types are explicit.** A source-curated association, a reviewed ontology term, a measured concentration, an occurrence report, and a computational prediction are different kinds of statement and are stored and displayed as such.
4. **Claims stay smaller than the data.** The platform does not assert effects. It reports what sources say, at the evidence grade they say it, and it says so in every API response.

5.b. Architecture

The shared knowledge base runs on Google Cloud (project `terpedia-489015`) as five deliberately separate layers, each chosen for what it is good at.

Layer	Technology	Role
Raw lake	Cloud Storage bucket	Immutable, content-addressed source files, manifests, and derived exports. Nothing is overwritten; refreshes are new snapshots.
SQL store	BigQuery (`terpedia_raw`, `terpedia_ops`)	Typed relational tables and source-specific projections for bulk analytics; ingestion-run and validation state.

Semantic store	Apache Jena Fuseki (RDF/ SPARQL)	Ontologies (ChEBI, Rhea, MeSH, MONDO), PubChem-derived RDF, and Terpedia's generated semantic bundles; graph queries and reasoning.
Semantic/API store	Firestore	Versioned source records, resolved entities, statements with sources, and the metadata behind the public API.
Document index	Full-text index over PDFs and monographs	Regulatory documents, ethnobotanical compendia, patents, and other sources that are documents rather than tables.

A synchronization layer ("tonsync") manages immutable snapshots, manifests, validation, and promotion between layers, so that any served record can be traced to the exact source bytes it was derived from. Scheduled jobs refresh sources on their natural cadence; relation sets from CannabisDatabase.ca, for example, refresh weekly.

5.c. Sources

The source catalog currently covers more than thirty public and licensed databases across chemistry, biology, composition, regulation, and literature. Appendix A: Source inventory gives the full inventory with ingestion status. The principal groups are:

- **Chemical identity and classification**: PubChem, ChEBI, the FDA UNII registry (159,376 records, 986,598 names), COCONUT (Sorokina et al., 2021), TeroKit (Zeng et al., 2020), SuperNatural II, KNApSAcK.
- **Occurrence**: LOTUS (Rutz et al., 2022), Dr. Duke's phytochemical and ethnobotanical databases, the Native American Ethnobotany Database (NAEB), TCMID.
- **Biochemistry**: Rhea, BRENDA, UniProt (adapter planned), and terpene–protein interaction sets (SAIR).
- **Composition**: EssoilDB essential-oil compositions; CannabisDatabase.ca (12,039 source records, 54,251 relations); normalized cannabis laboratory profiles (43,018 source rows, 409,655 measurements).
- **Assays**: PubChem BioAssay.
- **Regulatory and pharmacopoeial**: FDA and SCOGS records, HPUS, and monograph collections.
- **Literature and IP**: PubMed (via an automated terpene-focused collection pipeline), patent indexes and documents.
- **Ontologies**: MeSH, MONDO, Cellosaurus.

5.d. Scale, stated carefully

Terpedia publishes its counts as versioned observations with an explicit identity key and classification rule (Terpedia, LLC, 2026b). The following figures were retrieved from the production counting endpoint in September 2026.

Dataset / scope	Rows	Unique identities	Classification
COCONUT source-declared terpenoids	199,405	199,234	Source-declared terpenoid
COCONUT terpenoids with PubChem link	177,449	177,313	Derived projection of the above
TeroKit, confirmed terpene/terpenoid categories	145,779	145,354	Source-declared
TeroKit, “Others” and “Steroids”	23,173	23,061	Ambiguous; excluded from confirmed totals
Two-source census union (COCONUT + TeroKit, exact InChI)	—	268,924	Candidate identities; 225,905 with exact PubChem CID
LOTUS occurrences (April 2026 release)	674,422 triples	227,316 structures	37,468 organisms; 91,379 papers

These figures are not additive; records overlap across datasets, and the general COCONUT, UNII, and SuperNatural tables are deliberately *not* presented as terpene totals because they have not been through a classification pass. That restraint is the point. A number with a scope can be reproduced, audited, and updated; a number without one cannot.

5.e. Access surfaces

5.e.i. *kb.terpedia.com: the open biochemical graph:*

The encyclopedia site is a browsable, searchable view of the resolved graph. Users can search any name and browse by category: terpenes, molecules, proteins, enzymes, reactions, pathways, products, ingredients, diseases, and organisms. Each entity page shows identity, structure, cross-references, classification, related entities, and every statement with its source and license. The site is server-rendered from the same API described next, so what a person sees and what a program receives are the same data.

5.e.ii. *The Terpedia Knowledge API:*

The Knowledge API is the programmatic surface. Its **open encyclopedia resolver**, `GET /v1/cyc/{keyword}`, accepts a compound, cultivar, or protein name and returns the resolved entry, related entities, typed statements with sources, citations, and a plain-text evidence context. It requires no key and permits cross-origin requests, which is why the live examples in this paper can run in a browser. Keyed **partner endpoints** add:

- `POST /v1/tabular/search`: search a named source table or the unified source union (for example, all records mentioning "limonene" across sources).
- `POST /v1/tabular/count`: scoped, classified counts with identity keys, suitable for audit.
- `POST /v1/semantic/search`: semantic search with source citations and graph context.
- `GET /v1/entities/{id}` and `/v1/entities/{id}/neighbors`: entity and graph-neighborhood retrieval.

Partners with analytical needs can also be granted direct BigQuery access to the promoted tables and SPARQL access to the semantic store. A documented strain-profile endpoint distinguishes normalized measurements from wide source rows for structured consumers such as Colab notebooks.

5.e.iii. *chat.terpedia.com: domain agents with a knowledge boundary:*

The chat surface runs a hub-and-spoke multi-agent system over the knowledge base. As of this writing the live roster includes twenty-two specialist agents and eleven Tersona personas. Specialists include a **Terpene Scientist** (chemistry, sensory profiles, formulation implications, analytical interpretation), a **COA Analyst** (certificate interpretation, lab-result review, specifications), a **Compliance Agent** backed by state-specific regulatory datasets, **Formulation**, **Sourcing**, **Patent**, **Spectra**, **Cultivation**, **Medical Literature**, **Functional Products**, **Product Intelligence**, and a **Publication Editor** for template-governed documents. The backend also exposes OpenAI-compatible `/v1` endpoints so that existing tools can point at Terpedia as a drop-in model provider with domain retrieval behind it.

What distinguishes these agents from a general chatbot with a good prompt is where the boundary lives. Retrieval returns statements with their evidence type and source, and the knowledge API's `context` block begins with an explicit instruction not to extend claims beyond the cited evidence. The knowledge layer sets the boundary; the language model works inside it.

5.e.iv. *Tersona: the terpenes speak for themselves:*

Tersonae are entity personas, one per terpene (Limonene, Myrcene, α-Pinene, Linalool, β-Caryophyllene, Humulene, Terpinolene, Ocimene, Bisabolol, Geraniol), hosted by **TerpeneQueen**, a persona voiced by Terpedia's founding scientist. Each Tersona is grounded in its molecule's record and evidence, and the system supports speech in and out, so a consumer, a retail associate, or a student can *ask limonene about itself* and get an answer that stays inside what the graph can support. Tersonae also subscribe to literature and news feeds for their own name, which is the basis of the planned feed-first terpedia.com experience.

5.e.v. *Terproduct: products, ingredients, certificates, compounds:*

Terproduct is the system of record for product-level chemistry: **products arrow.r ingredients arrow.r certificates of analysis arrow.r compound results**. A

product's label lines are parsed into ingredients; ingredients are linked to organisms and to compounds; certificates are stored as documents with their parsed compound results; and every compound is joined to the knowledge base by InChIKey so that identity, bioactivity, disease associations, literature, and LOTUS occurrence records attach automatically. An installable web app supports barcode scanning to look up or ingest products in the field.

Two data-model decisions in Terproduct show the evidence rules at work. Disease associations carry a `kind`: a *reported association* (the compound was detected or studied in that condition) is distinguished from an *occupational exposure* hazard, and the molecule page must render the kind alongside the condition, because a bare disease name next to a consumer product reads as a health claim regardless of what the record says. And LOTUS occurrences are described in the interface as reports of detection, not as concentrations or sourcing recommendations.

5.e.vi. *Terports and the encyclopedia:*

Terports are Terpedia's short-form research reports: structured, cited summaries on a compound, a question, or a product class, generated from the knowledge base and reviewed before publication. The terpedia.com encyclopedia exposes compound lookups (`/cyc/{compound}`) and text-processing endpoints that automatically link scientific keywords in content to their entries, so that a partner's own site can carry Terpedia definitions without leaving the page.

5.f. Partnership example: MONDAYS

MONDAYS, a hemp-derived terpene chew brand, publishes a measured terpene profile on each product page: every compound at or above 0.05% of total volatiles, the aggregate trace fraction, and the milligrams each compound contributes to a 20 mg terpene serving. Terpedia builds the co-branded catalog at mondays.terpedia.com from that data and maintains the molecule reference pages that each compound links to. A product marked as *measured* carries laboratory-quantified compounds from its certificate; the catalog shows composition only. The protein-assay, disease-association, and literature records for each molecule live on Terproduct, where they are labeled as laboratory-dose findings, and each molecule page links out to them rather than restating them next to the product.

The arrangement is a template for any brand that wants to talk about its terpenes truthfully: publish the measurement, link to the science, keep the two visibly separate.

5.g. Research portfolio

Terpedia runs a source-auditable research program whose purpose is to test the platform's own assumptions and to publish reusable methods. Fourteen projects are active as of September 2026; the index at terpedia.github.io/research tracks status and target journals. Several are directly relevant to industry:

- **Census** (Terpedia, LLC, 2026a): how many terpenes there are, and why the answer depends on the counting rule.

- **Claims** (Terpedia, LLC, 2026c): promotional terpene effects recast as hypotheses and joined to receptor evidence without confusing mechanism with efficacy; includes a prioritized register of falsifiable tests.
- **Cannabis terpene chemistry**: the Terpotype classifier (McShan & Trapp, 2026b) and the provenance audit that demoted it (McShan & Trapp, 2026a), together with a record-level variability analysis and a causal-network model of terpene profile variation.
- **Classifier and TerpNet**: auditable molecular classification using canonical identity and fingerprints, and a general terpene pathway graph framework for genome, transcript, and docking evidence.
- **Absinthe** (Terpedia, LLC, 2026d): a critical evidence map testing whether absinthe should be described as psychedelic, mapping thujone, fenchone, and trans-anethole against 5-HT2A and $GABA_A$ pharmacology; the companion public site is at nowhere.terpedia.com.
- **Artemisia** (Terpedia, LLC, 2026e): a frozen scoping review of terpene diversity, biosynthetic evolution, and bounded antiparasitic evidence across the genus that gave the world artemisinin.
- **Terpene PubMed**: a nightly search matrix that makes literature density visible across terpene and topic combinations.
- **NAuTILUS**: a GPU-accelerated design for large-scale terpene–protein interaction ranking and docking, positioned explicitly as a hypothesis-ranking tool rather than a binding oracle.
- **Functional Flavors** and **ADHD attention**: research surfaces for flavor chemistry and for the attention literature in which terpenes are sometimes invoked.

The portfolio's stated evidence rule is worth quoting in full, because it is also the platform's:

> A database record is not a validated natural product. A receptor association is not efficacy. A docking score is not affinity. A graph path is not proof of in-vivo production.

The next section shows the platform doing what this section describes, against the live service.

6. Live examples against the Terpedia Knowledge API

Every result on this page came from the public Terpedia Knowledge API at build time. Nothing is mocked. The same service powers kb.terpedia.com and chat.terpedia.com.

> **Run it yourself**
>
> On the web version of this paper, click the **power button** at the top of the page to start an in-browser Python kernel (JupyterLite, no install). Then edit any keyword below and re-run the cell. The cells use only the standard library, so they run identically in CPython and in the browser.

The endpoint used throughout is the open **encyclopedia resolver**:

`GET https://terpedia-knowledge-nanrsdlaoa-uc.a.run.app/v1/cyc/{keyword}`

It accepts a compound name, synonym, cultivar name, or protein name and returns:

Key	What it holds
`entry`	The resolved entity: label, type, identifiers, structure, provenance
`entities`	Related entities in the graph (up to 30)
`statements`	Typed relationships (`found_in_taxon`, `contains`, `associated_with` …), each with sources
`citations`	The literature and database records behind those statements
`context`	A plain-text evidence summary that begins with the instruction *do not extend claims beyond the cited evidence*

Partner endpoints (`/v1/tabular/search`, `/v1/tabular/count`, `/v1/semantic/search`, `/v1/entities/{id}/neighbors`) require an API key and are described in the white paper.

```
import json, sys

API = "https://terpedia-knowledge-nanrsdlaoa-uc.a.run.app"

def kb(keyword: str) -> dict:
    """Resolve a keyword with the Terpedia Knowledge API.

    Works in CPython (urllib) and in the browser under Pyodide (fetch).
    Returns {} when the API has no entry for the keyword (HTTP 404).
    """
    from urllib.parse import quote
    url = f"{API}/v1/cyc/{quote(keyword)}"
    if sys.platform == "emscripten":                        #
JupyterLite / Pyodide
        from pyodide.http import open_url
        text = open_url(url).read()
```

```
        data = json.loads(text)
        return {} if "error" in data else data
    from urllib.request import urlopen
    from urllib.error import HTTPError
    try:
        with urlopen(url, timeout=60) as r:
            return json.load(r)
    except HTTPError as e:
        if e.code == 404:
            return {}
        raise

print("helper ready ->", API)
```

```
helper ready -> https://terpedia-knowledge-nanrsdlaoa-uc.a.run.app
```

6.a. 1. Resolve a name to a chemical identity

The first thing any downstream system needs is an unambiguous identity. A trade name, a lab-report label and a literature synonym must all land on the same record. Terpedia returns the structure (InChI, InChIKey, SMILES) and the cross-references that let you join to PubChem, ChEBI, ChEMBL, CAS, KEGG, FooDB and others.

```
d = kb("limonene")
e = d["entry"]

print("label      :", e["label"])
print("type       :", e["type"])
print("formula    :", e.get("molecularFormula"))
print("SMILES     :", e.get("smiles"))
print("InChIKey   :", e["identifiers"].get("inchiKey"))
print("class      :", e.get("taxonomy", {}).get("direct_parent"))
print()
print("cross-references:")
for k, v in sorted(e["identifiers"].items()):
    print(f"  {k:12s} {v}")
```

```
label      : (-)-Limonene
type       : chemical
formula    : C10H16
SMILES     : CC(=C)[C@@H]1CCC(C)=CC1
InChIKey   : XMGQYMWWDOXHJM-JTQLQIEISA-N
class      : Menthane monoterpenoids

cross-references:
  biocyc       CPD-4886
  cannabisdb   CDB005271
  cas          5989-27-5
  chebi        15382
  chembl       CHEMBL15799
  chemspider   389747
  drugbank     DB08921
  foodb        FDB006329
  inchiKey     XMGQYMWWDOXHJM-JTQLQIEISA-N
```

```
  kegg          C06099
  knapsack      C00010868
  pubchem       440917
  pubchemCid    22311
  wikidata      Q278809
  wikipedia     Limonene
```

Every record carries its own provenance: which dataset it came from, the source record id, when it was retrieved, and under which license it may be reused. That last field matters for commercial users; it is how Terpedia keeps license-scoped sources visible instead of laundering them into one undifferentiated table.

```
p = e["provenance"]
for k in ("datasetId", "datasetVersion", "sourceRecordId",
"curationStatus",
          "licenseTerms", "license", "retrievedAt"):
    print(f"{k:16s} {p.get(k)}")
```

```
datasetId        cannabisdb
datasetVersion   1.0-a7d05067088ee5b8
sourceRecordId   CDB005271
curationStatus   source_curated
licenseTerms     CC BY-NC 4.0
license          https://creativecommons.org/licenses/by-nc/4.0/
retrievedAt      2026-09-06T11:18:33.538Z
```

6.b. 2. Why identity resolution is hard: one molecule, dozens of names

Limonene alone arrives under many names in COAs, ingredient lists, and papers. The alias list is what lets a search for `d-limonene`, `(+)-limonene`, `dipentene`, or a CAS number resolve to the same node.

```
aliases = e["aliases"]
print(len(aliases), "aliases on record. A sample:")
for a in aliases[:18]:
    print("  -", a)
```

```
58 aliases on record. A sample:
  - (+)-(4R)-Limonene
  - (+)-(R)-Limonene
  - (+)-4-Isopropenyl-1-methylcyclohexene
  - (+)-Limonene
  - (4R)-1-Methyl-4-isopropenylcyclohex-1-ene
  - (4R)-4-Isopropenyl-1-methylcyclohexene
  - (R)-(+)-Limonene
  - (R)-(+)-p-Mentha-1,8-diene
  - (R)-1-Methyl-4-(1-methylethenyl)cyclohexene
  - (R)-4-Isopropenyl-1-methyl-1-cyclohexene
  - (R)-p-Mentha-1,8-diene
  - 4BetaH-p-mentha-1,8-diene
  - D-(+)-Limonene
  - D-Limonen
  - AISA 5203-L (+)limonene
  - Dipentene
```

```
  - (-)-Limonene
  - 1-Methyl-4-(1-methylethenyl)cyclohexene
```

6.c. 3. Statements with sources: what is this molecule associated with?

Beyond identity, the graph holds typed statements. For a compound these include taxa it has been reported in (`found_in_taxon`), protein associations from the source database (`associated_with`), and class membership. Each statement is a claim with a source, not a fact about biology. Note the `evidenceType` field: `source_curated_protein_association` tells you exactly what kind of evidence you are looking at.

```
from collections import Counter

stmts = d["statements"]
print("statement predicates for limonene:")
for pred, n in Counter(s["predicate"] for s in stmts).most_common():
    print(f"  {n:3d}  {pred}")

print()
print("protein associations (source-curated, not efficacy):")
for s in stmts:
    if s["predicate"] == "associated_with":
        src = (s.get("sources") or [{}])[0]
        print(f"  {s['subjectLabel']} -> {s['objectLabel']:<32s} "
              f"[{src.get('evidenceType')}] {src.get('url')}")
```

```
statement predicates for limonene:
   20  found_in_taxon
    3  part_of
    2  associated_with
    2  has_part
    1  classified_as
    1  described_as
    1  has_molecular_formula
    1  image
    1  mass
    1  canonical_smiles
    1  inchi
    1  inchi_key
    1  chemical_formula
    1  subclass_of
    1  instance_of
    1  chembl_id
    1  pubchem_cid

protein associations (source-curated, not efficacy):
  (-)-Limonene -> Cytochrome P450 2C9
[source_curated_protein_association] https://cannabisdatabase.ca/
proteins/CDBP00973/compound_protein_links
  (-)-Limonene -> Cytochrome P450 2C19
[source_curated_protein_association] https://cannabisdatabase.ca/
proteins/CDBP00975/compound_protein_links
```

6.d. 4. Resolve a protein target

The same resolver handles proteins. `CB2` lands on the reviewed UniProt entry P34972 from Terpedia's curated core graph; `CYP2C9` and `TRPV1` come from the CannabisDatabase.ca protein set and bring their compound associations with them.

```
for name in ("CB2", "CYP2C9", "TRPV1"):
    r = kb(name)
    if not r:
        print(f"{name}: no entry")
        continue
    ent = r["entry"]
    assoc = [s for s in r.get("statements", []) if s["predicate"] ==
"associated_with"]
    print(f"{name:7s} -> {ent['label']:<48s} id={ent['id']:<16s} "
          f"source={ent['provenance']['datasetId']:<24s}
associations={len(assoc)}")
```

```
CB2     -> Cannabinoid receptor 2
id=uniprot:P34972   source=terpedia-curated-core    associations=1
```

```
CYP2C9  -> Cytochrome P450 2C9                                id=CDBP00973
source=cannabisdb               associations=16
```

```
TRPV1   -> Transient receptor potential cation channel subfamily V
member 1 id=CDBP01848        source=cannabisdb
associations=6
```

```
r = kb("CYP2C9")
print("compounds associated with", r["entry"]["label"], "in the source
graph:")
seen = set()
for s in r["statements"]:
    if s["predicate"] == "associated_with" and s["subjectLabel"] not in
seen:
        seen.add(s["subjectLabel"])
        print("  -", s["subjectLabel"])
```

```
compounds associated with Cytochrome P450 2C9 in the source graph:
  - Delta-9-tetrahydrocannabinol
  - Limonene
  - Formic acid
  - NADP
  - NADPH
  - Oxygen
  - Salicylic acid
  - Water
  - Rutin
  - Hydrogen sulfide
  - Heme
  - (-)-Limonene
  - Dopamine
  - Paraxanthine
  - Nicotine
  - Ethanol
```

6.e. 5. Compare a panel of terpenes in one pass

A formulator or QA lead rarely wants one molecule. This loop pulls a small panel and shows how much the graph currently knows about each: formula, mass, number of reported taxa, number of statements, and which dataset supplied the primary record. Where a record is thin, the table says so; a missing value is reported as missing, not filled in.

```
panel = ["myrcene", "limonene", "linalool", "alpha-pinene",
"caryophyllene",
         "humulene", "valencene", "nootkatone"]

rows = []
for name in panel:
    r = kb(name)
    if not r:
        rows.append((name, "not found", "", "", "", "", ""))
        continue
    ent = r["entry"]
    st = r.get("statements", [])
    taxa = sum(1 for s in st if s["predicate"] == "found_in_taxon")
    mass = ""
    for s in st:
        if s["predicate"] == "mass":
            raw = str(s.get("objectLabel") or s.get("objectValue") or
s.get("value") or "")
            mass = raw.split()[0] if raw.split() else ""
            break
    rows.append((name, ent["type"], ent.get("molecularFormula") or "",
mass,
                 str(taxa), str(len(st)), ent["provenance"]
["datasetId"]))

hdr = ("keyword", "type", "formula", "mass", "taxa", "stmts", "primary
source")
w = [max(len(str(x[i])) for x in rows + [hdr]) for i in range(len(hdr))]
print("  ".join(h.ljust(w[i]) for i, h in enumerate(hdr)))
print("  ".join("-" * w[i] for i in range(len(hdr))))
for row in rows:
    print("  ".join(str(c).ljust(w[i]) for i, c in enumerate(row)))
```

```
keyword        type                    formula  mass        taxa  stmts
primary source
-------------  ----------------------  -------  ----------  ----  -----
------------------------
myrcene        chemical                C10H16   136.125201  25    40
cannabisdb
limonene       chemical                C10H16   136.125201  20    40
cannabisdb
linalool       terpene                 C₁₀H₁₈O  154.135765  27    40
terpedia-terpene-registry
alpha-pinene   terpene                 C₁₀H₁₆   136.125     27    40
terpedia-terpene-registry
```

```
caryophyllene  essential_oil_compound  C₁₅H₂₄   204.187801  28    40
terpedia-terpene-registry
humulene       chemical                C15H24   204.187801  24    40
cannabisdb
valencene      essential_oil_compound  C₁₅H₂₄   204.188     29    40
terpedia-terpene-registry
nootkatone     essential_oil_compound  C₁₅H₂₂O  218.167065  25    35
terpedia-terpene-registry
```

6.f. 6. From a cultivar name to quantified chemistry, with the paper behind it

Cultivar names are notoriously unreliable, so Terpedia never treats a name as chemistry. What it can do is return the *measured* records that exist for a name, each with its quantification, status, and the publication it came from. Here the `contains` statements for *Blue Dream* carry concentrations in mg/g dry weight and resolve to a PubMed identifier. A compound appears once per measured sample, so repeated rows are separate measurements, not duplicates.

```
r = kb("Blue Dream")
ent = r["entry"]
print(ent["label"], "-", ent["type"], "-", ent["id"])
print()
print(f"{'compound':<36s} {'value':<26s} status")
for s in r["statements"][:14]:
    if s["predicate"] != "contains":
        continue
    q = s.get("qualifiers", {})
    print(f"{s['objectLabel']:<36s} {q.get('value',''):<26s}
{q.get('status','')}")

src = (r["statements"][0].get("sources") or [{}])[0]
print()
print("source:", src.get("title", "")[:140], "...")
print("pmid  :", src.get("pmid"), "->", src.get("url"))
print("evidence type:", src.get("evidenceType"))
```

```
Blue Dream - cannabis_cultivar - STRAIN0007

compound                             value                      status
Cannabichromene                      1.3 +/- 0.1 mg/g dry wt    Detected
and Quantified
Cannabidiol                          0.5 +/- 0.1 mg/g dry wt    Detected
and Quantified
Delta-9-tetrahydrocannabinol         109.9 +/- 24.1 mg/g dry wt Detected
and Quantified
Delta-9-tetrhydrocannabivarin        0.3 +/- 0.1 mg/g dry wt    Detected
and Quantified
Delta-9-cis-tetrahydrocannabinol     109.9 +/- 24.1 mg/g dry wt Detected
and Quantified
(+)-nerolidol                        0.353 mg/g dry wt          Detected
and Quantified
(+)-nerolidol                        0.414 mg/g dry wt          Detected
and Quantified
```

```
(+)-nerolidol                          0.794 mg/g dry wt          Detected
and Quantified
(+)-nerolidol                          0.8 mg/g dry wt            Detected
and Quantified
beta-Myrcene                           2.146 mg/g dry wt          Detected
and Quantified
beta-Myrcene                           2.405 mg/g dry wt          Detected
and Quantified
beta-Myrcene                           5.113 mg/g dry wt          Detected
and Quantified
beta-Myrcene                           6.512 mg/g dry wt          Detected
and Quantified
beta-Myrcene                           7.5 +/- 2.8 mg/g dry wt    Detected
and Quantified

source: Richins RD, Rodriguez-Uribe L, Lowe K, Ferral R, O'Connell MA:
Accumulation of bioactive metabolites in cultivated medical Cannabis.
PLoS On ...
pmid  : 30036388 -> https://pubmed.ncbi.nlm.nih.gov/30036388/
evidence type: curated_cultivar_concentration
```

6.g. 7. Where the compound has been reported: occurrence statements

`found_in_taxon` statements link a compound to organisms in which it has been reported. An occurrence is a report that a compound was detected in an organism in a cited work. It is not a concentration and not evidence that the organism is a meaningful commercial source. The statement tells you where the record came from and under what license. Where a taxon label has not yet been resolved, the bare Wikidata identifier is shown; that coverage gap is visible rather than papered over.

```
r = kb("linalool")
occ = [s for s in r["statements"] if s["predicate"] == "found_in_taxon"]
print(len(occ), "occurrence statements returned for linalool (page 1).
First ten:")
for s in occ[:10]:
    src = (s.get("sources") or [{}])[0]
    print(f"  {s['objectLabel']:<32s} {src.get('datasetId'):<10s}
{src.get('license')}")
```

```
27 occurrence statements returned for linalool (page 1). First ten:
  Baccharis dracunculifolia        wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Pilocarpus grandiflorus          wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Myrcianthes cisplatensis         wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q15583465                        wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q15586596                        wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q13939086                        wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q536839                          wikidata   https://creativecommons.
```

```
org/publicdomain/zero/1.0/
  Q15583174                          wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q3595850                           wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
  Q28801933                          wikidata   https://creativecommons.
org/publicdomain/zero/1.0/
```

6.h. 8. The evidence boundary is part of the payload

Every response includes a `context` block written for downstream language models. It opens with an explicit instruction not to extend claims beyond the cited evidence, then lists the statements with their sources. This is how the chat agents at chat.terpedia.com are kept honest: the knowledge layer, not the prompt, sets the boundary.

```
ctx = kb("limonene")["context"]
print(ctx[:1100])
print("...")
```

```
Verified Terpedia semantic graph (do not extend claims beyond the cited
evidence):
- (-)-Limonene --associated_with--> Cytochrome P450 2C9 [https://cannabi
sdatabase.ca/proteins/CDBP00973/compound_protein_links]
- (-)-Limonene --associated_with--> Cytochrome P450 2C19 [https://cannab
isdatabase.ca/proteins/CDBP00975/compound_protein_links]
- (-)-Limonene --classified_as--> Menthane monoterpenoids [https://canna
bisdatabase.ca/compounds/CDB005271]
- (-)-Limonene --described_as--> d-Limonene, also known as dipentene,
belongs to the class of organic compounds known as menthane
monoterpenoids. These are monoterpenoids with a structure based on the
o-, m-, or p-menthane backbone. P-menthane consists of the cyclohexane
ring with a methyl group and a (2-methyl)-propyl group at the 1 and 4
ring position, respectively. The o- and m- menthanes are much rarer, and
presumably arise by alkyl migration of p-menthanes. Thus, D-limonene is
considered to be an isoprenoid lipid molecule. d-Limonene is a very
hydrophobic molecule, practically insoluble (in water), and relatively
neutral. (-)-Limonene is e
...
```

6.i. 9. Honest misses

A keyword the graph cannot resolve returns HTTP 404 and `{"error": "entry not found"}` rather than a fuzzy guess. That is deliberate. Coverage is versioned and expanding; the counting reports in the white paper describe exactly what is and is not in scope today.

```
for k in ("thujone", "artemisinin", "not-a-real-compound"):
    print(f"{k:22s} ->", "found" if kb(k) else "no entry (404)")
```

```
thujone                -> no entry (404)
artemisinin            -> no entry (404)
not-a-real-compound    -> no entry (404)
```

6.j. 10. The same data, rendered

The entity pages on kb.terpedia.com are built from the same responses shown above.

Figure 1: The live limonene entity page at kb.terpedia.com.

Ask a question in natural language at chat.terpedia.com, or browse the graph by category (terpenes, molecules, proteins, enzymes, reactions, pathways, products, ingredients, diseases, organisms) at kb.terpedia.com.

7. Use cases for industry

This section translates the platform into workflows. Each use case names the team, the job, what Terpedia supplies, and, just as importantly, what Terpedia deliberately does not do. Buyers should evaluate a knowledge platform on both.

7.a. Formulation and product development

Team: R&D, flavorists, formulators. **Job:** Design a terpene blend to a target profile, source compliant ingredients, and document the rationale.

With Terpedia. Start from a reference: a measured cultivar profile, an essential-oil composition, or a competitor's published COA. Resolve every listed compound to an InChIKey so that "β-caryophyllene", "caryophyllene", and "(E)-BCP" are one line, not three. Pull physical properties (mass, logP, volatility class) and occurrence data to identify natural sources and plausible botanical carriers. Check each compound against regulatory listings (FEMA number, GRAS status, UNII) and, for inhalable products, against state additive rules. Use the **Formulation Agent** and **Terpene Scientist** to reason about sensory interactions, stability, and dosing under the platform's evidence rules, and the **Functional Products** agent to map the boundary between flavor and claim before the concept reaches marketing.

What Terpedia does not do. It does not predict organoleptic outcomes of a blend, and it does not certify a formulation as compliant. It gives the formulator the identities, data, and regulatory pointers to do both with a paper trail.

7.b. Quality and laboratory management

Team: QA/QC, laboratory directors, supply-chain quality. **Job:** Interpret incoming certificates of analysis, detect drift, and maintain specifications.

With Terpedia. Ingest COAs into Terproduct so that every compound result is normalized to a structural identity and stored with its document, batch, and report date. Compare a new batch against the product's own history and, where relevant, against the normalized cannabis measurement set (409,655 measurements across 43,018 profiles) to see whether a value is unusual for the compound and matrix. Because Terpedia's own audit shows that laboratory identity dominates profile shape (McShan & Trapp, 2026a), treat the lab as a covariate: a change in lab is a change in the measurement system, not necessarily in the product. Use the **COA Analyst** agent to draft result reviews that cite the specification and the record.

What Terpedia does not do. It does not replace proficiency testing or accredit a laboratory. It makes the data from many laboratories comparable enough that drift becomes visible.

7.c. Regulatory affairs and claims substantiation

Team: Regulatory, legal, compliance. **Job:** Approve label and marketing language; respond to regulator and retailer questions with sources.

With Terpedia. Every effect statement a marketing team proposes can be checked against the claims register, which separates *mechanism evidence* (is there any linked

receptor or assay evidence for this compound?) from *effect support* (is the effect itself supported by appropriately scoped human or preclinical studies?) (Terpedia, LLC, 2026c). The knowledge API returns the evidence type on every statement, so the difference between a source-curated protein association and a controlled human trial is machine readable, not a matter of interpretation. The **Compliance Agent** is backed by state-specific regulatory datasets rather than a general corpus, and the **Medical Literature** agent is scoped to clinical evidence and safety.

What Terpedia does not do. It does not give legal advice or pre-clear claims. It ensures that when a claim is made, the evidence behind it is visible, graded, and citable, and that when a claim cannot be supported, that is visible too.

7.d. Product content, education, and marketing

Team: Content, brand, retail education. **Job:** Talk about terpenes accurately and engagingly at scale.

With Terpedia. Molecule pages provide sourced summaries, structure images, occurrence, and cross-references that can be embedded or linked. The text-processing endpoints on terpedia.com link scientific keywords in a partner's own content to encyclopedia entries automatically. Tersonae give consumers and retail associates a conversational interface to each terpene that stays inside the evidence. The MONDAYS catalog (Partnership example: MONDAYS) is the template: publish composition, link to science, keep them visibly separate. The **Marketing** and **Ad** agents draft within those constraints rather than outside them.

What Terpedia does not do. It will not generate effect copy the graph cannot support. Brands that want "relaxing" and "energizing" on the label will find the platform an honest friction, which is its value in a market where enforcement is increasing.

7.e. Sourcing and procurement

Team: Procurement, supply chain, ingredient sourcing. **Job:** Find and qualify sources of a compound or a profile.

With Terpedia. Occurrence data (LOTUS, Dr. Duke, EssoilDB) answers "which organisms has this compound been reported in, and with what literature?" so that alternative botanical sources can be identified. TeroKit's purchasable-molecule tables and UNII registrations help qualify synthetic and biosynthetic supply. Essential-oil composition data lets a buyer compare a supplier's specification against typical ranges for the botanical. The **Sourcing Agent** frames the search in supply-chain terms.

What Terpedia does not do. An occurrence record is not a yield estimate and not a sourcing recommendation; the platform says so on the record.

7.f. Intellectual property and competitive intelligence

Team: IP counsel, strategy, product intelligence. **Job:** Understand the prior art and the competitive landscape for a terpene-based concept.

With Terpedia. Patent indexes and documents are part of the knowledge base and are searchable alongside the literature. The terpene PubMed matrix shows where literature is dense and where it is thin for any terpene–topic combination, which is where defensible novelty tends to be. The **Patent** and **Product Intelligence** agents work over these sources with the same provenance discipline.

What Terpedia does not do. It does not perform freedom-to-operate analysis. It shortens the research phase that precedes one.

7.g. Data science and AI teams

Team: Data engineering, ML, internal tools. **Job:** Build terpene-aware applications, models, and retrieval systems.

With Terpedia. Bulk access to promoted BigQuery tables, SPARQL access to the semantic store, a documented REST API with provenance on every record, an MCP tool server for agent frameworks, and OpenAI-compatible chat endpoints that can be dropped into existing pipelines. Structure-based identity means Terpedia records join cleanly to internal chemistry systems; provenance fields mean a model's training data can be audited; explicit evidence types mean a retrieval-augmented system can be told to cite only human evidence, or only measured concentrations, or only open-licensed sources.

What Terpedia does not do. It does not hide license constraints. Some sources (CannabisDatabase.ca, BRENDA) are license-scoped, and the license is on the record so that a team can filter by it.

7.h. A note on who this is not for

Terpedia is not a consumer wellness app, a medical-advice service, or a substitute for a certified laboratory or a licensed regulatory consultant. It is infrastructure for the people who do those jobs. Organizations that want a platform to tell them what their product *does* will be disappointed; organizations that want a platform to tell them what their product *contains*, what is *known* about it, and how well it is known, will not.

8. Evidence rules and limitations

A knowledge platform earns trust by being clear about what it will not say and where it is incomplete. This section states Terpedia's evidence rules and its known limitations as of September 2026.

8.a. The rules

Terpedia applies the same rules to its data model, its API, its agents, and its own publications.

A database record is not a validated natural product. A compound's presence in COCONUT, TeroKit, or any other source, including Terpedia's own union, means a source classified it as a terpenoid. It does not mean the structure has been independently verified, isolated, or shown to occur in nature. Counts are reported as candidate identities with their identity key.

An occurrence is not a concentration. A `found_in_taxon` statement means a cited work reported detecting the compound in that organism. It says nothing about level, tissue, or commercial viability.

A receptor association is not efficacy. A source-curated compound–protein link, a binding constant, or an assay hit describes a molecular interaction under specified conditions. It is not evidence of a physiological effect, a therapeutic benefit, or human relevance. Terpedia's claims register tracks mechanism evidence and effect support as separate fields for exactly this reason (Terpedia, LLC, 2026c).

A docking score is not affinity. Computational predictions are ranked hypotheses for experimental testing. They are labeled as such and are never merged into the same field as measured interactions.

A graph path is not proof of in-vivo production. Linking a genome, an enzyme, and a product in a pathway graph shows that a route is plausible, not that an organism uses it.

A cultivar name is not a chemistry. Names are attributes of measured records, not sources of chemical information; where Terpedia's own research found that laboratory identity predicted profile shape better than cultivar name, it revised its own classifier accordingly (McShan & Trapp, 2026a).

A claim is a hypothesis until scoped evidence supports it. Promotional effect statements are catalogued as hypotheses with an explicit uncertainty boundary and, where possible, a falsifiable test.

Every response carries its boundary. The `context` block returned by the knowledge API opens with the instruction not to extend claims beyond the cited evidence. Downstream language models are constrained by the data layer, not only by their prompts.

8.b. Known limitations

Terpedia's coverage reports are public, and the following gaps are real.

Resolver coverage is incomplete. The open resolver currently returns no entry for some compounds that are clearly in scope, including thujone and artemisinin, because the resolved-entity layer has been populated from specific source adapters (CannabisDatabase.ca, the terpene registry, EssoilDB, curated core) and has not yet been extended across the full BigQuery census union. The live-examples section shows a miss deliberately. Coverage is versioned and expanding; the SQL layer already holds the structures.

Some labels are unresolved. Occurrence statements sourced from Wikidata sometimes display a bare Q-identifier where the taxon label has not yet been fetched. The identifier is correct and resolvable; the presentation is incomplete.

Source encoding defects propagate. At the time of writing, some CannabisDatabase.ca descriptions carry a character-encoding fault (Greek letters rendered as two-character sequences). This originates in ingestion and is being corrected; it is mentioned here because a platform that promises provenance should also disclose its own defects.

Ingestion quality varies by source. The coverage matrix records, for example, 39 errors in the latest EssoilDB run and a large number of reference and row errors in the SAIR interaction set. Validated tables are marked validated; partial and raw-only sources are marked as such. Consumers should read the status column, not just the source name.

Several adapters are planned, not live. A release-aware UniProt adapter, the remaining PubChem RDF shards, and the Fuseki promotion of MeSH and MONDO are in progress. LOTUS structure serving is validated; the full occurrence metadata promotion awaits a complete upstream archive.

Licensing constrains reuse. CannabisDatabase.ca content is CC BY-NC 4.0 and BRENDA is license-scoped. Commercial users must filter or license accordingly. Terpedia keeps the license on every record so that this is possible, and will not present license-restricted data as if it were open.

Counts are observations, not properties. Every figure in this paper is a dated observation from a specific scope. When a source refreshes, the count is re-run and appended, not overwritten. Readers quoting Terpedia numbers should quote the date and scope with them.

Terpedia is not a medical or legal authority. The platform does not diagnose, recommend treatment, pre-clear claims, or certify compliance. Its agents are configured to say so.

8.c. Why this section is in a marketing document

Buyers of scientific infrastructure have learned to discount vendor claims. The most persuasive thing Terpedia can show a quality director or a regulatory lead is not a large number; it is a platform that already behaves the way their auditors will want it

to. Every limitation above is visible in the product today, and every rule above is enforced in the data rather than promised in a slide.

9. Getting started and roadmap

9.a. Three ways to engage

Explore. Everything described in this paper is reachable today without an account. Search and browse the graph at kb.terpedia.com. Ask the specialist agents and the Tersonae a question at chat.terpedia.com. Look up a molecule's measured presence in a real product at mondays.terpedia.com. Call the open resolver from a terminal:

```
curl -s https://terpedia-knowledge-nanrsdlaoa-uc.a.run.app/v1/cyc/
limonene | head -c 600
```

Or run the live examples in this paper in your browser and change the keywords.

Integrate. Teams that want programmatic access request an API key for the tabular, semantic, and entity endpoints. The knowledge API is documented with request and response examples; the chat backend exposes OpenAI-compatible endpoints so that existing tools, notebooks, and agent frameworks can point at Terpedia with a configuration change; and an MCP tool server exposes knowledge-base operations to agent runtimes that speak the Model Context Protocol. Integration engagements typically begin with a scoped question ("normalize our incoming COAs", "attach evidence to our product catalog") and a two-week proof of value.

Partner. Deeper collaborations include co-branded product catalogs on the MONDAYS model, Terproduct as a certificate-of-analysis system of record for a brand or a laboratory, direct BigQuery and SPARQL access for analytics teams, custom Terports on a product class or question, and joint research on questions where a partner has data and Terpedia has method. Research collaborations are published under the same evidence rules as everything else.

9.b. What an engagement looks like

1. **Scoping (week 0).** Identify the decisions the data must support, the sources already in use, and the regulatory context.
2. **Identity pass (weeks 1–2).** Resolve the partner's compound, ingredient, and product vocabularies to structural identities; report unresolved items rather than forcing matches.
3. **Evidence attachment (weeks 2–4).** Join the resolved identities to occurrence, property, regulatory, assay, and literature records with provenance and evidence type intact.
4. **Surface (weeks 3–6).** Deliver through whichever surface fits the team: API, BigQuery, agent, embedded pages, or Terproduct.
5. **Review and publish.** Optional Terport or joint publication; optional inclusion in the public graph where the partner's data is open.

9.c. Roadmap

The roadmap below reflects the current public plans in Terpedia's repositories. Dates are directional.

Knowledge base (2026 H2). Complete the PubChem RDF shard set; promote MeSH and MONDO into the semantic store; implement the release-aware UniProt adapter;

complete LOTUS occurrence metadata promotion; extend the open resolver across the full census union so that every candidate identity has an entity page.

Product (2026 H2 – 2027 H1). Move terpedia.com to a feed-first experience in which Tersonae surface literature, news, and product updates for the terpenes a user follows; open Tersona chat to every entity in the graph; expand Terproduct's field-scanning and certificate ingestion; publish the Rx (recipes) module for shareable, forkable formulations linked to the encyclopedia and the product catalog.

Research (rolling). Submit the first wave of manuscripts (Census, Claims, Terpotype and its audit, Absinthe, Artemisia); finalize the Classifier and TerpNet methods papers; advance NAuTILUS from design to a validated target-panel docking pipeline with redocking and experimental anchors as gates.

Platform. Continue to publish counting and coverage reports on every source refresh; add per-source data-quality dashboards to the public site; expose license filtering directly in the API.

9.d. Contact

Terpedia, LLC, Monument, Colorado. terpedia.com · dan@terpedia.com · github.com/Terpedia

This white paper is maintained as a living document at its published URL. The Markdown source, the executable examples, and the bibliography are open; the examples are re-executed against the live service on every build, so the numbers on the live-examples page are as current as the last publish date on the site.

10. Appendix A: Source inventory

The table below is condensed from Terpedia's operational coverage matrix as of September 2026. It lists the external and licensed sources in the knowledge base, the serving tier each is promoted to, and its ingestion status. Terpedia-internal derived tables, staging tiers, and infrastructure rows are omitted. Status meanings follow the coverage matrix: **validated** means the raw snapshot and a query-shaped promotion are both working; **partial** means a bounded or derived subset is serving; **raw/document** means bytes are mirrored but no validated serving adapter is enabled yet; **planned** means the adapter is not yet implemented.

Source	Domain	Serving tier	Status (September 2026)
PubChem	Chemical identity, structures, RDF	Fuseki + BigQuery	Partial; repaired descriptor shard loaded, remaining RDF shards staged
PubChem BioAssay	Assay records	Fuseki + BigQuery projections	Partial; XML mirrors complete, RDF promotion pending
ChEBI	Chemical ontology	Fuseki	Validated
Rhea	Biochemical reactions	Fuseki	Validated
UniProt	Proteins	BigQuery + Fuseki (planned)	Planned; adapter not yet implemented
MeSH	Biomedical vocabulary	Fuseki	Raw/manifest validated; Fuseki promotion queued
MONDO	Disease ontology	Fuseki	Raw/manifest validated; Fuseki promotion queued
COCONUT	Open natural products	BigQuery	Validated
LOTUS	Compound–organism occurrences	BigQuery + raw	Structure serving validated; occurrence metadata promotion pending
BRENDA	Enzyme data	BigQuery + raw	Validated; license-scoped

Dr. Duke	Phytochemical and ethnobotanical	BigQuery + document index	Validated
EssoilDB	Essential-oil composition	BigQuery	Validated with quality gaps (39 errors in latest run)
KNApSAcK	Metabolite mass spectra	BigQuery	Validated
TCMID	Traditional Chinese medicine herbs, ingredients, prescriptions	BigQuery + raw	Validated TSV exports
FDA	Regulatory documents and tables	BigQuery / document index	Validated subset
UNII	FDA substance registry	BigQuery	Validated; 159,376 records, 986,598 names
SCOGS	GRAS evaluations	BigQuery	Validated
Cannabis laboratory profiles	Terpene measurements	BigQuery	Validated; 43,018 rows, 409,655 measurements
CannabisDatabase.ca	Compounds, cultivars, proteins, relations	Firestore + BigQuery + raw	Validated; 12,039 records, 54,251 relations (CC BY-NC 4.0)
SAIR	Terpene–protein interactions	BigQuery	Validated with quality gaps (row/reference errors recorded)
SuperNatural II	Natural products	BigQuery	Validated
TeroKit	Terpenome database	BigQuery + Fuseki projections	Validated typed load with classification view
Cellosaurus	Cell lines	BigQuery / Fuseki projection	Validated
NAEB	Native American ethnobotany	BigQuery	Validated
Patents	Patent indexes and documents	BigQuery + document index	Indexes validated; PDFs partial

PubMed (terpene collection)	Literature	Document index / BigQuery	Automated nightly collection
Eden, Ayurvedic, Bach, HPUS, WHO monographs	Pharmacopoeial and traditional-use documents	Document index	Validated subsets
OpenTerps	Community terpene data	Raw	Raw/document
Terport TTL bundles	Terpedia-generated RDF	Fuseki	Validated selected manifest (18 loaded)

Additional acquisition candidates (for example ChEMBL, BindingDB, Open Targets) are tracked in the catalog as candidates and are explicitly not counted as current coverage. The authoritative machine-readable catalog and the full coverage matrix are maintained in the public `kb-source` repository at github.com/Terpedia/kb-source.

10.a. Appendix B: Knowledge API summary

Endpoint	Auth	Purpose
`GET /health`	none	Service health
`GET /v1/cyc`	none	Paged list of resolved entities
`GET /v1/cyc/{keyword}`	none	Resolve a compound, cultivar, or protein name; returns entry, related entities, statements with sources, citations, and evidence context
`POST /v1/tabular/search`	key	Search one source table or the unified source union
`POST /v1/tabular/count`	key	Scoped, classified counts with identity keys
`POST /v1/semantic/search`	key	Semantic search with citations and graph context
`GET /v1/entities/{id}`	key	Entity record by identifier
`GET /v1/entities/{id}/neighbors`	key	Graph neighborhood of an entity

Base URL: `https://terpedia-knowledge-nanrsdlaoa-uc.a.run.app`. Keyed endpoints take the key in the `x-knowledge-key` header. Cross-origin requests are permitted.

References

Adams, T. B., Gavin, C. L., McGowen, M. M., Waddell, W. J., Cohen, S. M., Feron, V. J., Marnett, L. J., Munro, I. C., Portoghese, P. S., Rietjens, I. M. C. M., & Smith, R. L. (2011). The FEMA GRAS assessment of aliphatic and aromatic terpene hydrocarbons used as flavor ingredients. *Food and Chemical Toxicology*, *49*(10), 2471–2494. https://doi.org/10.1016/j.fct.2011.06.011

André, R., Gomes, A. P., Pereira-Leite, C., Marques-da Costa, A., Monteiro Rodrigues, L., Sassano, M., Rijo, P., & Costa, M. do C. (2024). The entourage effect in cannabis medicinal products: a comprehensive review. *Pharmaceuticals*, *17*(11), 1543. https://doi.org/10.3390/ph17111543

Booth, J. K., & Bohlmann, J. (2019). Terpenes in Cannabis sativa – from plant genome to humans. *Plant Science*, *284*, 67–72. https://doi.org/10.1016/j.plantsci.2019.03.022

California Department of Cannabis Control. (2023,). *Guidance: requirements for inhaled cannabis products*. https://cannabis.ca.gov/2023/11/guidance-requirements-for-inhaled-cannabis-products

Christianson, D. W. (2017). Structural and chemical biology of terpenoid cyclases. *Chemical Reviews*, *117*(17), 11570–11648. https://doi.org/10.1021/acs.chemrev.7b00287

Dittmar, D., & Schuttelaar, M. L. A. (2019). Contact sensitization to hydroperoxides of limonene and linalool: results of consecutive patch testing and clinical relevance. *Contact Dermatitis*, *80*(2), 101–109. https://doi.org/10.1111/cod.13137

Gertsch, J., Leonti, M., Raduner, S., Racz, I., Chen, J.-Z., Xie, X.-Q., Altmann, K.-H., Karsak, M., & Zimmer, A. (2008). Beta-caryophyllene is a dietary cannabinoid. *Proceedings of the National Academy of Sciences*, *105*(26), 9099–9104. https://doi.org/10.1073/pnas.0803601105

Grand View Research. (2025,). *Aromatherapy market size and share report, 2026–2033*. https://www.grandviewresearch.com/industry-analysis/aromatherapy-market

Harada, H., Kashiwadani, H., Kanmura, Y., & Kuwaki, T. (2018). Linalool odor-induced anxiolytic effects in mice. *Frontiers in Behavioral Neuroscience*, *12*, 241. https://doi.org/10.3389/fnbeh.2018.00241

IMARC Group. (2025,). *Essential oils market size and industry report 2034*. https://www.imarcgroup.com/essential-oils-market

LaVigne, J. E., Hecksel, R., Keresztes, A., & Streicher, J. M. (2021). Cannabis sativa terpenes are cannabimimetic and selectively enhance cannabinoid activity. *Scientific Reports*, *11*, 8232. https://doi.org/10.1038/s41598-021-87740-8

Market.us. (2025,). *Terpenes market size, share, growth*. https://market.us/report/terpenes-market/

McShan, D. C., & Trapp, S. (2026b,). *A name-independent terpene-profile classification of commercial Cannabis sativa: the Terpedia Terpotype Classifier (TTC-7)*. https://github.com/Terpedia/strain/blob/main/manuscript/terpotype_classification_article.md

McShan, D. C., & Trapp, S. (2026a,). *Provenance audit of a commercial cannabis terpene archive: database identity versus commercial name*. Companion manuscript to the Terpotype classifier, 4 September 2026. https://github.com/Terpedia/strain/blob/main/manuscript/cannabis_archive_audit_article.md

Oregon Liquor and Cannabis Commission. (2020,). *Non-cannabis additives in inhalable cannabinoid products*. https://www.oregon.gov/olcc/Docs/commission_agendas/2020/Non-Cannabis-Addtives-in-Inhalable-Cannabinoid-Products.pdf

Pichersky, E., & Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds?. *New Phytologist*, *220*(3), 692–702. https://doi.org/10.1111/nph.14178

Precedence Research. (2025,). *Terpene market size to hit USD 2.95 billion by 2035*. https://www.precedenceresearch.com/terpene-market

Russo, E. B. (2011). Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. *British Journal of Pharmacology*, *163*(7), 1344–1364. https://doi.org/10.1111/j.1476-5381.2011.01238.x

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J. G., Stephan, R., Page, R., Vondrášek, J., Steinbeck, C., Pauli, G. F., Wolfender, J.-L., Bisson, J., & Allard, P.-M. (2022). The LOTUS initiative for open knowledge management in natural products research. *Elife*, *11*, e70780. https://doi.org/10.7554/eLife.70780

Song, Y., Liu, H., Quax, W. J., Zhang, Z., Chen, Y., Yang, P., Cui, Y., Shi, Q., & Xie, X. (2024). Application of valencene and prospects for its production in engineered microorganisms. *Frontiers in Microbiology*, *15*, 1444099. https://doi.org/10.3389/fmicb.2024.1444099

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., & Steinbeck, C. (2021). COCONUT online: collection of open natural products database. *Journal of Cheminformatics*, *13*, 2. https://doi.org/10.1186/s13321-020-00478-9

Spindle, T. R., Zamarripa, C. A., Russo, E., Pollak, L., Bigelow, G., Ward, A. M., Tompson, B., Sempio, C., Shokati, T., Klawitter, J., Christians, U., & Vandrey, R. (2024). Vaporized D-limonene selectively mitigates the acute anxiogenic effects of delta-9-tetrahydrocannabinol in healthy adults who intermittently use cannabis. *Drug and Alcohol Dependence*, *257*, 111267. https://doi.org/10.1016/j.drugalcdep.2024.111267

Terpedia, LLC. (2026e,). *Artemisia terpene diversity across scales: biosynthetic evolution and bounded antiparasitic evidence*. Frozen scoping review package, 1 September 2026. https://github.com/Terpedia/kb/tree/main/research/artemisia_antiparasitic/review-package-2026-09-01

Terpedia, LLC. (2026a,). *How many terpenes are there? A source-audited census of candidate terpene and terpenoid identities*. Working draft, 4 September 2026. https://github.com/Terpedia/census

Terpedia, LLC. (2026d,). *Is absinthe a different kind of drunk, and is it psychedelic? A critical Terpedia evidence map of ethanol, thujone, and 5-HT2A pharmacology*. Working manuscript, 3 September 2026. https://github.com/Terpedia/absinthe

Terpedia, LLC. (2026c,). *Terpedia claims investigation: promotional terpene effects recast as hypotheses*. https://github.com/Terpedia/claims

Terpedia, LLC. (2026b,). *Terpedia terpene/terpenoid counts*. Entity counting report, retrieved 4 September 2026 from the Terpedia Knowledge API. https://github.com/Terpedia/kb

Zeng, T., Liu, Z., Zhuang, J., Jiang, Y., He, W., Diao, H., Lv, N., Jian, Y., Liang, D., Qiu, Y., Zhang, R., Zhang, F., Tang, X., & Wu, R. (2020). TeroKit: a database-driven web server for terpenome research. *Journal of Chemical Information and Modeling*, *60*(4), 2082–2090. https://doi.org/10.1021/acs.jcim.0c00141