WHITEPAPER

The Means to Use Information

INTELLIGENT AND CONNECTED TAXONOMIES POWER INSIGHTFUL RESEARCH

Information is useful only when it can be found. Reaxys® Medicinal Chemistry delivers the right answers to your query because expert taxonomies organize the millions of data entries in the repository.

ELSEVIER

Information becomes accessible when it is organized according to the current state of knowledge in a discipline.

Taxonomies offer obvious organizational benefits, but their impact on database usability is far-reaching.

Reaxys Medicinal Chemistry is designed to be an immediate source of actionable information to support everyday decisions in drug lead discovery and optimization. In other words, every query in Reaxys Medicinal Chemistry delivers accurate and germane information, be it citations pertinent to a topic, property values of a specific compound, or a matrix of off-target activity for one or more known drugs. Matching database output to a search question requires each data point in the repository to be findable. The expert taxonomies of Reaxys Medicinal Chemistry are at the core of this functionality.

WHAT IS A TAXONOMY?

A taxonomy is a system to classify objects or concepts. We all use such systems to bring order into different aspects of our lives, be it food in the refrigerator, clothes in the closet or files on our laptops. Our systems help us maintain an overview of multiple objects or concepts and, more practically, they help us to quickly find any item we need if it has been appropriately classified. So, for example, you can scan sock colors to match an outfit by opening your sock drawer, but you might not find your bathing suit if you hung it up with your winter coats.

While having an organizational scheme for a handful of items is useful, such classification systems become imperative for compilations of thousands to millions of items, such as in a database. Tracking large data volumes is impossible without organization and a database lacking taxonomy has little value because individual data points cannot be easily found. Without some organizational structure, you have to look through every item until you find data of interest. Add a simple organization, like listing items alphabetically, and you can more efficiently find pertinent data. Superimpose a second classification scheme, like overarching categories, and you reduce your search time further because data appear in shorter alphabetical lists under relevant categories.

The benefit of taxonomies is quite obvious, but their impact on the usability of a database is far-reaching. More than make it easy to find a desired item, an expertly designed taxonomy embedded in a fitting database structure translates a user query into insightful results, unfolds the complete depth and breadth of a data repository, and maintains a database up to date. But, how?

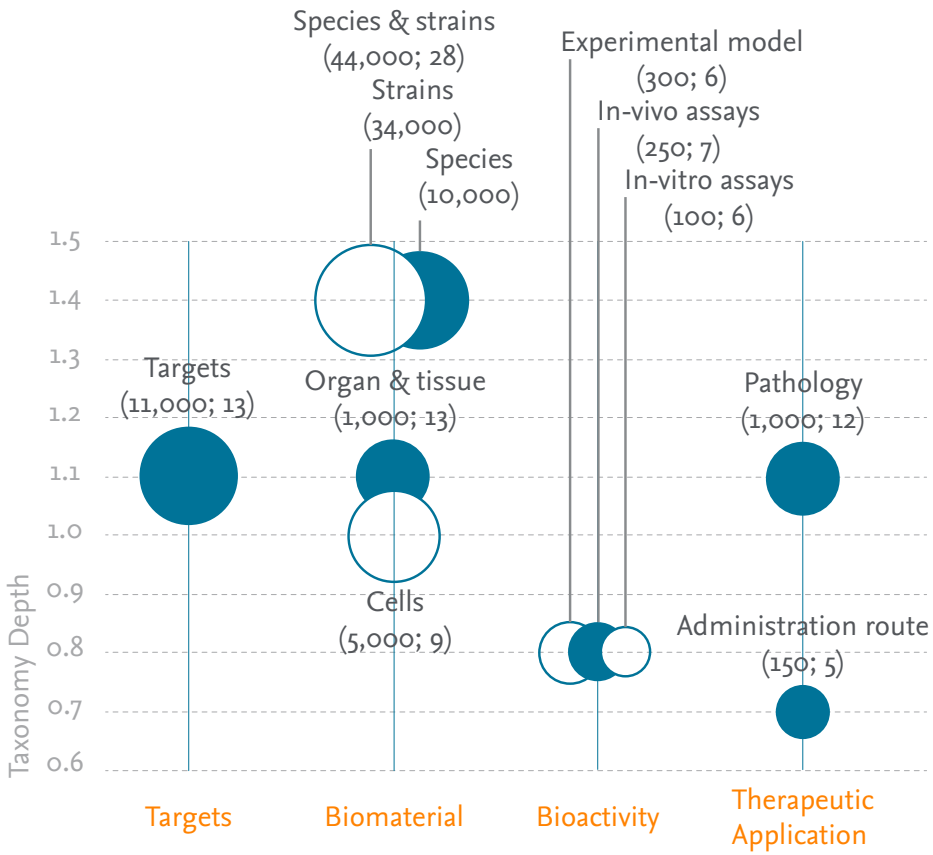


Figure 1. The taxonomies that organize the comprehensive bioactivity data in Reaxys Medicinal Chemistry. Size of circle represents the number of terms and concepts in a taxonomy and location on the y-axis indicates the number of hierarchical levels below the top node of the taxonomy (both on a log scale). These values are listed for each taxonomy in parenthesis (number of terms; number of levels).

GRANULAR DATA AS WELL AS SYSTEMATIC REVIEWS SURFACE FROM DETAILED TAXONOMIES

More than a list of facts, knowledge is understanding how concepts and terms relevant to a topic are interconnected. Knowledge about a particular subject area can be described with a taxonomy as a hierarchical tree that links specific terms to broader, overarching concepts. Perhaps the most well-known taxonomy is the classification of living organisms, with hierarchical levels ranging from kingdoms at the root of the tree, up to species and subspecies at the ends of branches. A critical feature of such a hierarchical classification is that each term included in the taxonomy resides within the context of an overarching term or concept. This relationship of inclusion carries an explicit body of information associated with each item. So, for example, in the classification of living organisms, starfish are included in the kingdom Animalia and the phylum Echinodermata. These categories tell us that starfish are animals with radial symmetry and a calcareous internal skeleton that are only distantly related to fish. A direct consequence of this inclusionary organization is that a taxonomy that describes a knowledge domain in great detail will (a) include a comprehensive inventory of relevant terms and concepts that are (b) arranged into a suitable number of hierarchical levels to represent broader/narrower relationships among taxonomy concepts.

The detailed taxonomies underlying Reaxys Medicinal Chemistry are extensive — both in number of terms included (Figure 1) and in number of hierarchical levels — to accommodate the highly granular data included in the repository. These detailed taxonomies support targeted, user-defined searches that deliver hits imminently relevant to your query. At the same time, connections among terms and concepts laid out in the taxonomies enable a broader search scope for exploratory "probing" of database content. For example, a search for any one concept can be expanded to include all hits associated by hierarchy, ie., all hits classified under the same overarching concept (an explosion search).

Furthermore, a browsable representation of the taxonomies in Reaxys Medicinal Chemistry facilitates exploration of database content. You can start at any term and move down the hierarchy to uncover increasingly narrow and detailed information, or move up the hierarchy to examine broader concepts and venture into adjacent areas of information. Again, the depth of the taxonomies organizing data in Reaxys Medicinal Chemistry is advantageous because each hierarchical level adds to your knowledge about the data classified under a given term or concept. This knowledge enables you to draw conclusions by comparing and contrasting information clustered under one term versus another. For example, you can quickly compare the targets DNA gyrase and Topoisomerase IV, find other similar targets or explore a completely different approach to a research question.

A multi-domain classification of data reflects the multidisciplinary nature of lead discovery and development.

DIRECTLY USABLE RESULTS RESIDE IN MULTIPLE TAXONOMIES

While taxonomy depth accommodates the granularity of the data harbored in Reaxys Medicinal Chemistry, another important feature in making those data directly usable is a multi-domain classification. Information relevant to the discovery and development of chemical leads stems from several knowledge domains: knowledge about potential targets, knowledge about the structure of small molecules, knowledge about the behavior of these molecules in biological materials, knowledge about experimental methodologies and more. Each domain is discrete but multiple domains are conjugated in any research endeavor. Measuring the affinity of ciprofloxacin for DNA gyrase draws on information about the antibiotic, the target and in vitro assays. Consequently, information contained in a research publication or patent must be indexed according to several knowledge domains. A multifaceted classification allows data to be sorted according to parameters that you may use in your queries, be it pharmacological concepts, reaction conditions, pharmacokinetic properties or any other terms relevant to your research question.

Reaxys Medicinal Chemistry classifies excerpted bioactivity data according to eight taxonomies that cover the following knowledge domains:

- **Target**: expert classification of proteins and non-coding RNAs examined and used as therapeutic targets
- **Species/Strain**: extensive taxonomy of study organisms, including all phylogenetic, strain, breed and human population classifications
- **Pathology**: systematic classification of known human diseases
- **Experimental Model**: compilation of induced diseases in animal models
- **In-vivo and In-vitro Assays**: comprehensive description of in vivo and in vitro procedures, including experimental tests and bioassays
- **Organ & Tissue**: organization of relevant anatomical concepts including organs, tissues, body fluids and secretions, and body regions
- **Cells**: detailed categorization of primary and secondary cell cultures as well as cell lines
- **Administration Route**: standardized description of routes of administration

Data are connected across taxonomies via a relational database structure.

INSIGHTS UNFOLD FROM CONNECTED TAXONOMIES

While taxonomies organize data into hierarchical relationships from general, overarching concepts to specific, granular terms of a particular knowledge domain, capturing the conjugation of data in the various domains that inform medicinal chemistry is essential to deliver insightful information that supports daily decisions in your workflow. Highly detailed information about DNA gyrase as a target and ciprofloxacin as a therapeutic compound is not sufficient. Equally relevant is to know the affinity of the antibiotic for the protein, or the relation between a drug, the experimental disease used to test it and the parameters measured or calculated. Because these data are organized according to different taxonomies, database fields in Reaxys Medicinal Chemistry are coupled through a relational database structure. These connections characterize semantic relationships between terms and concepts organized under one taxonomy to those indexed according to the other taxonomies (Figure 2). Thus, for example, the database field with a particular variant of DNA gyrase, classified according to the Target taxonomy, is connected to results from a reported bioassay (indexed according to the in-vivo and in-vitro Assays taxonomy) measuring the affinity of ciprofloxacin for that variant.

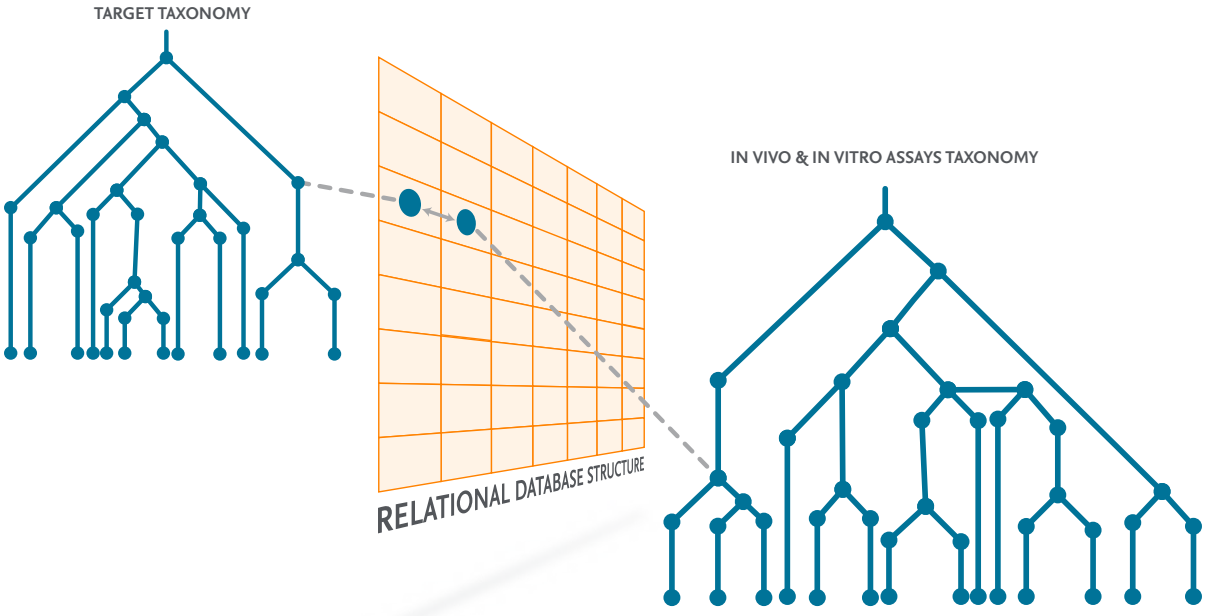


Figure 2. The relational structure of the Reaxys Medicinal Chemistry database provides semantic relationships between the various database fields and their underlying taxonomies.

Database architecture must also account for evolving and expanding terminology.

GERMANE INFORMATION EMERGES FROM REPRESENTATIVE TAXONOMIES

Results from a query in Reaxys Medicinal Chemistry are up to date and current with developments in the discipline. New data are continuously excerpted and accessioned into the database via a standardized production process (see our Whitepaper "Production Innovation to Generate the Best Information"). However, just as knowledge in a research field is continuously updated, the used concepts and terms also change. As a result, ensuring that information contained in Reaxys Medicinal Chemistry remains germane (and findable) requires more than continued accession of new data. The database architecture organizing the data must also account for evolving and expanding terminology. The taxonomies of Reaxys Medicinal Chemistry ensure that every item in the database is linked to the terminology most recommended by the scientific community, regardless of the language used in the source. They also guarantee that connections between concepts and terms reflect the current state of the discipline.

Some taxonomies of Reaxys Medicinal Chemistry are derived from Emtree, the comprehensive life science taxonomy underlying Embase®. Emtree is a multi-domain taxonomy that has matured over decades of life science literature indexing, expert information curation and long-term monitoring of the biomedical sciences. The taxonomies derived for Reaxys Medicinal Chemistry build on that cumulative database architecture, but have been expertly adapted to focus on chemistry and bioactivity of small molecules. As a consequence, these taxonomies represent a comprehensive description of this informational space, encompassing past and present knowledge. New taxonomies developed specifically for Reaxys Medicinal Chemistry build on the terminology and structure of well-known and authoritative information systems. For example, the Target taxonomy was built with terminology from UniProt (www.uniprot.org) and Rfam (rfam.xfam.org), and based on the hierarchical structure of Gene Ontology (geneontology.org) and InterPro (www.ebi.ac.uk/interpro). In this way, data retrieved from Reaxys Medicinal Chemistry is congruent with today's data and knowledge models of medicinal chemistry.

The impact of using representative and comprehensive terminology is two-fold. On one hand, the taxonomies account for differences in use of language and condense results of your search to the most current conceptual understanding of each knowledge domain. On the other hand, the taxonomies include recommended terms that to this date have not been used to organize database entries, but that may become relevant in the future. For example, the Target taxonomy includes 11,000 gene products, of which only roughly 1,000 are known compound targets. This means that when a new gene product is tested as a potential target, the results of experiments can be accessioned into the database without needing to validate either term or position of the term within the taxonomy. Thus, the new information is available sooner.

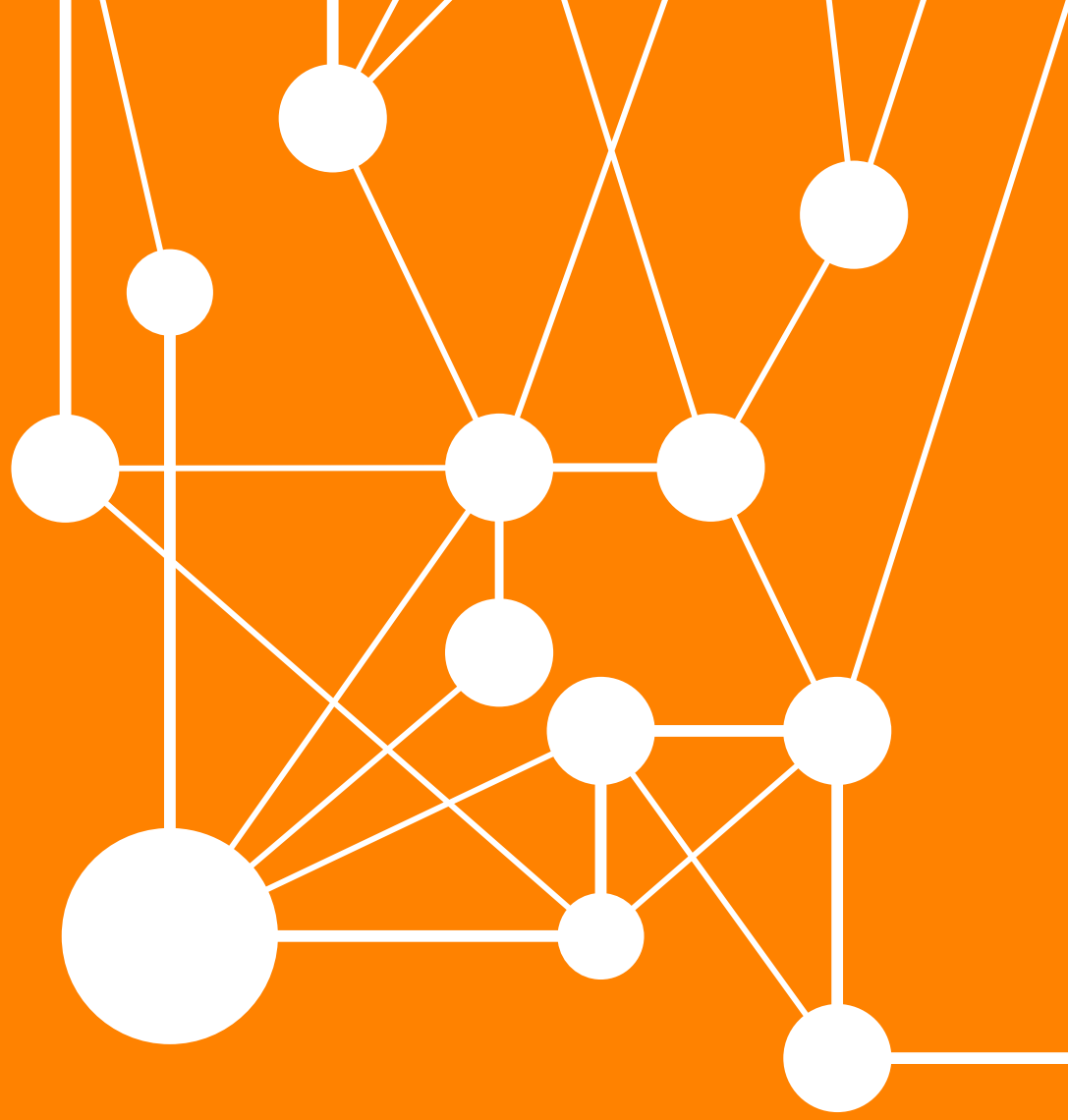
AN UP-TO-DATE INFORMATION TOOL BUILDS ON DYNAMIC TAXONOMIES

Even the most comprehensive, expertly designed taxonomy becomes obsolete over time if it remains static. Any classification system must include a mechanism to adapt to changes in the informational space that it describes. These changes are not restricted to just additional data and altered terminology, but include complete paradigm shifts. Maintaining taxonomy robustness and utility for a discipline that expands and evolves as quickly as medicinal chemistry is challenging. A great advantage of Reaxys Medicinal Chemistry is that the database does not exist in isolation, but rather as a coordinated information system within the Elsevier R&D Solutions for Pharma & Life Sciences portfolio. With taxonomies derived from Emtree, the database architecture of Reaxys Medicinal Chemistry is fed taxonomy updates from two sources. First, the excerptors for Reaxys Medicinal Chemistry experience first-hand any important conceptual shifts in source literature and propose new concepts and terms to incorporate into taxonomies. Second, the production process of Embase also leads to taxonomy updates that, where relevant, are added to Reaxys Medicinal Chemistry. Both update mechanisms initiate a systematic evaluation by a multidisciplinary content specialist team that examines the impact of proposed changes on all taxonomies of Reaxys Medicinal Chemistry, and safeguards the relevance of their terminology, structure and interconnectedness. Currently, the taxonomy hierarchies are updated three times a year.

Hosting Reaxys Medicinal Chemistry within the information environment of all Elsevier R&D Solutions for Pharma & Life Sciences also creates a unique opportunity to link several products in the portfolio and, in this way, grant a user access to an expanded information landscape with a few mouse clicks (or finger taps). Because taxonomies are standardized from one product to another and their updates are coordinated, information from one database can be easily mapped to the structure of another. In the future, your search for information may begin in Reaxys Medicinal Chemistry but then take you to Embase, Reaxys, PathwayStudio®, PharmaPendium® or other domain-specific systems.

BEYOND DATA, THE MEANS TO USE INFORMATION

Ultimately, a database is simply a compilation of data pertaining to a specific topic. The work of compiling these data is undoubtedly a large endeavor in and of itself. The tremendous value latent in that compilation, however, is lost if the database does not include a mechanism to use the data. That is the reason for a taxonomy; and the more sophisticated the database uses, the more intelligent the design of the underlying taxonomies. The content of Reaxys Medicinal Chemistry distinguishes it from other information systems. Highly granular data are accessioned through a detailed excerption process and undergo scrutiny at several points in the production process to safeguard the quality of the compilation. These data are also standardized and normalized so that they can be immediately implemented to inform decisions in the process of discovering and optimizing therapeutic compounds. Yet, what makes that content deliverable in response to your specific query, are the detailed and expert taxonomies that organize the content, support sophisticated retrieval algorithms and keep the information system current.



LEARN MORE

To request information or a product demonstration, please visit elsevier.com/reaxys or email us at reaxys@elsevier.com.

Visit elsevier.com/rd-solutions
or contact your nearest Elsevier office.

ASIA AND AUSTRALIA
Tel: +65 6349 0222
Email: sginfo@elsevier.com

JAPAN
Tel: +81 3 5561 5034
Email: jpinfo@elsevier.com

KOREA AND TAIWAN
Tel: +82 2 6714 3000
Email: krinfo.corp@elsevier.com

EUROPE, MIDDLE EAST AND AFRICA
Tel: +31 20 485 3767
Email: nlinfo@elsevier.com

NORTH AMERICA, CENTRAL AMERICA AND CANADA
Tel: +1 888 615 4500
Email: usinfo@elsevier.com

SOUTH AMERICA
Tel: +55 21 3970 9300
Email: brinfo@elsevier.com

June 2015