THE PERFECT BALANCE
CREATING COMPREHENSIVE YET RELEVANT INFORMATION SYSTEMS

Dr. Jürgen Swienty-Busch
Director of Product Management for Chemistry
BALANCING COMPREHENSIVENESS AND RELEVANCE

Comprehensiveness and relevance can be difficult to balance in an information system designed to support chemists in all industry and academic environments. The information needs of an engineer at a chemical manufacturer differ from those of a medicinal chemist in the pharmaceutical industry or a doctoral student in an academic inorganic chemistry laboratory. More than differences in the type of information supporting various research questions, each specialty operates with domain-specific data models to assist in processing available information; so while a synthetic chemist may create a chemistry space in terms of alterations to a given chemical structure, a medicinal chemist examines the same space in terms of effects caused by a given chemical structure.

Achieving the right balance between comprehensive and relevant results for every chemist is also a moving target because chemistry is a dynamic knowledge space. Each development toward that end is part of a continuous effort to match the output of our information solutions to the demands of the user, regardless of specialty. This match requires creative solutions that, on one hand, grant access to the full spectrum of information encompassed by our databases and, on the other hand, enable the nuanced data processing that is meaningful to each scientific domain. The following are examples of such solutions that make Reaxys and Reaxys Medicinal Chemistry powerful information and analytical tools to support chemistry and drug discovery.
Chemistry data, even within a single publication, are highly heterogeneous, ranging from text descriptions in a methods section to diagrams of chemical reactions, illustrations of chemical structures and tables summarizing physicochemical properties. Retrieving these different data forms typically requires different search strategies. Text searches pull up hits containing queried words and synonyms, but cannot be used to retrieve drawn chemical structures. Structural searches pull up hits linked to queried chemical structures if they are indexed as connection tables.

Most chemistry searches, however, begin with a phrase or string of words representing a chemical name, topic or concept. Such a search should grant access to a comprehensive list of hits related to that topic or concept, regardless of whether the search terms appear specifically in the text, title or abstract of an article or patent. Enabling this level of functionality is not trivial: the system must be able to (1) interpret correctly any string of words used to name a concept and (2) identify database entries that are relevant to the queried concept. The system must have a comprehensive conceptual description of the informational space it covers and database entries must be mapped to the terms in that description.

The Reaxys team has been working on precisely such a description to allow the system to understand concept searches. The result of these efforts is a sophisticated taxonomy called ReaxysTree. Building on the mature data structure evolved from a long history of abstracting and indexing chemical information, ReaxysTree currently arranges approximately 15,500 concepts enriched with 43,000 synonyms into multiple hierarchies. Each hierarchy describes chemical terms and topics as they relate to one another within the context of a specific chemical subject (e.g., reactions, physicochemical properties). The terms in ReaxysTree are used to index the database entries and thus, ReaxysTree functions as a chemistry dictionary helping to interpret the context of a search phrase and transform concept searches into technical queries that retrieve relevant data.

Reaxys leverages the relationships described by ReaxysTree to afford searches that reveal the full richness of its content. On one hand, Reaxys uses ReaxysTree to assign context to a phrase typed into the free-text search box AskReaxys and has the capability to request more input from the user if there is ambiguity. This ensures that the hits retrieved for the search are concept-relevant, regardless of the specific phrasing used. On the other hand, ReaxysTree serves as a clickable framework for literature searches based on browsing a hierarchy of the topics covered by Reaxys. With a few clicks, users can select topic categories relevant to their areas of interest to uncover the full content of Reaxys, compiled from over 16,000 publications. By aligning the entries in Reaxys with a description of the relationships between chemical terms and concepts, ReaxysTree facilitates matching search results to user intention. Furthermore, there is work underway to align also factual database entries with ReaxysTree so that all forms of data in Reaxys are readily accessible.
RELEVANCE: NORMALIZING LARGE DATASETS AND MAKING THEM COMPARABLE

Most disciplines have multiple research domains that converge in some aspects of their information needs, but are distinct in the way the information is used. Medicinal chemistry is a good example. Bioactive compound identification and optimization demands information not only on structure and properties of compounds, but also on their bioactivity. The latter data are less relevant to other chemists and thus, Reaxys Medicinal Chemistry was designed to hold data on the bioactivity of compounds, while their structure, reactions and physicochemical properties remain in the domain of Reaxys.

The content of Reaxys Medicinal Chemistry has been significantly enhanced over the last twelve months with the integration of high-value, third party data, resulting in nearly 20 million bioactivity data entries on approximately 4.5 million bioactive substances. Extensive data normalization was required to generate integrated data units that are meaningful and relevant in the context of medicinal chemistry and bioactive compound identification.

As an example, consider bioactivity data for a compound-target pair sourced from different databases. How can you ensure that all data points representing structure-activity relationships refer to the same target and chemical compound? While chemical compounds typically can be identified by comparing the connection tables between atoms of their chemical structures, targets are simple text names and the meaning depends on the source system; some define the triplet “target-subunit-species” as one target, others only refer to the duplet “target-subunit” and keep the species separate if they discuss a certain target-activity relationship. This difference in definitions makes it difficult to compare and analyze data and is the reason for normalizing data to build the coherent data space at the core of Reaxys Medicinal Chemistry (Figure 1).

With this extensive normalization, data can be presented in a model commonly used for drug discovery: bioactivity data in Reaxys Medicinal Chemistry can be compared at a glance in an intuitive compound-target matrix or heatmap. Comparing bioactivity data across assay types and from different source publications requires a meaningful common denominator. In the context of medicinal compound assessment, concentration-dependent activity of a given compound towards its target is highly informative, however, activity is reported using a variety of measures, ranging from half maximal inhibitory (IC50) or effective concentration (EC50) to effective dose (ED50), dissociation constant (Ki) or drug potency parameters. To enable a quick comparison of compound activity, Reaxys Medicinal Chemistry computes a unifying value, pX, based on these commonly used activity parameters. Displayed in the heatmap view of Reaxys Medicinal Chemistry, the activity of compound-target pairs can be quantitatively compared to identify the most interesting compound-target interactions. As in the example with ReaxysTree, the organization of information in Reaxys Medicinal Chemistry enables data access that matches user intention.
TYING IT ALL TOGETHER

High data volume and heterogeneity are hallmarks of a rapidly growing knowledge space and the magnified workload to process these data is the reason why information management systems were developed. Information management systems must now go beyond data retrieval and take over processing steps to extract meaning from these data. Enabling flexibility in the way a search is performed (e.g., correctly interpreting search strategies) and presenting data in a way that fits to the research question asked (e.g., heatmap comparison of compound-target activities) are developments in Reaxys and Reaxys Medicinal Chemistry that exemplify the ongoing work at Elsevier to meet demands for comprehensive and relevant information. The next step we are taking is to allow users to seamlessly access information from two or more of our Life Science Solutions platforms. We envision navigating with a few clicks from search results in one platform to relevant data in another system, and this cross-platform integration is already a reality between Reaxys and Reaxys Medicinal Chemistry. In this way, users have a comprehensive view of an interdisciplinary knowledge space containing data organized to retain domain-specific relevance: the perfect balance.

**Figure 1.** Normalizing imported data to fit taxonomic requirements of Reaxys Medicinal Chemistry.

Discover how Reaxys and Reaxys Medicinal Chemistry support chemistry research and drug discovery.
CONTACT INFORMATION

Please visit www.elsevier.com/reaxys

AMERICAS
(8am – 8pm CET – St. Louis)
Tel: US toll-free: +1 888 615 4500
Tel: Non toll-free: +1 314 523 4900
Email: usinfo@elsevier.com
Email Brazil: brinfo@elsevier.com

ASIA AND PACIFIC
(9.00am – 6pm SST – Singapore office)
Tel: +65 6349 0222
Email: sginfo@elsevier.com

EUROPE AND ALL OTHER REGIONS
(9.30am – 5.30pm JST – Tokyo office)
Tel: +31 20 485 3767
Email: nlinfo@elsevier.com

JAPAN
(9.30am – 5.30pm JST – Tokyo office)
Tel: +81 3 5561 5035
Email: jpinfoelsevier.com
www.elsevier.com/jp