Modelling reaction chemistry in database software: the chemical thesaurus

Modelización de reacciones químicas en el software de base de datos: el tesauro químico

MARK R. LEACH

School of Sciences, Chemistry, University of Salford, Salford, Gr. Manchester, M5 4WT, UK
m.r.leach@salford.ac.uk

Abstract

A methodology for modelling reaction chemistry using relational database technology is described. This approach produce an application which allows the user to "click through" from a chemical species to its reactions and on to the associated the reaction mechanism. The systems analysis exercise also required to construct the database illuminates the meta-structure of the underlying science.

Key words: *educational software, reaction mechanism, chemical thesaurus.*

Resumen

Hace una descripción de la metodología sobre modelización de reacciones químicas que usa la tecnología de base de datos. Este enfoque no solamente produce una aplicación que permite al usuario desplazarse desde unas especies químicas a sus reacciones y respectivo mecanismo de reacción y también realizar ejercicios de análisis de sistémico requerido para construir la base de datos sobre metaestructura de la ciencia subyacente.

Palabras clave: *software educativo, mecanismo de reacciones, tesauro químico.*

INTRODUCTION

When the theoretical physicist PAUL DIRAC published his relativistic quantum-mechanical theory for the electron in 1928, he is said to have remarked that his "equation explains most of physics and the whole of chemistry". Since that time theoreticians have modelled chemical species and their interactions by computation at various levels of theory: *ab initio* and semi-empirical, as well as using molecular mechanics and molecular dynamics, in an attempt to fulfill the DIRAC prediction. (Few chemistry calculations are performed at the full DIRAC level.) However, there are problems with the computational quantum chemistry approach (SCERRI, 2000). From a practical point of view, *ab initio* computational chemistry is best suited to gas-phase and non-polar environments and it is far more difficult to model condensed and/or polar phases. While nature may abhor a vacuum, the virtual nature of computational chemistry abhors water. The implication is that —at the present time— the system of chemical species and their reactions cannot be tackled in a general way using the *ab initio* approach and a more encompassing methodology is required.

In 1970 EDGAR CODD invented the modern database with his seminal paper, *The Relational Model of Data* (CODD, 1970), which describes a simple, optimal and elegant data storage and manipulation methodology. The CODD *relational database* uses "key fields" to link tables (relations) together so that tables, parts of tables and groups of tables can be logically manipulated. The power and subtlety of the relational model is best illustrated by example. Consider a relational database, which might be used to run a video store. Initial analysis of the *video store system* suggests that two tables ('flat file' databases) are required: a *customer database table* which holds customer ID, name and address information and a *video database table* which holds ID, title and price information. Further analysis reveals that a third "linking" table, consisting of the customer ID, the video ID and dates of issue and return, is required to monitor the loan of a specific video to a specific customer. (Formally, there is a relationally "impossible" many-to-many relationship between customer table and video table, but this difficulty is removed by the link table which has allowed many-to-one relationships both with the customer table *and* the video table.) The three table construct, figure 1, enables questions to be asked of the type: "how many videos does JOHN DOE have on hire and how many are overdue?". Any number of additional tables can added to the three core tables to give information about film genre and director, as well as tables, which deal with the business accounts.

The relational concept has been extraordinarily successful over the last thirty years and today all major software vendors (Oracle, IBM, Microsoft, etc.) offer relation database management system (RDMS) products. Everything from on-line airline booking to zoo administration uses the CODD model. But the relational approach is more than just a computer programming methodology. The systems analysis process *within the rigour of the relational model* gives extraordinary insight into the system being modelled. In our case that system is *reaction chemistry*.

Video Store Database

Customer Database Table

Cust ID	Customer Name
1	John Doe
2	Joyce Chester
3	Robin Johnson
4	Jenny Johnson
5	Bill Gilfillan
6	Millie Oliver

Link Table

Cust ID	Vid ID	Date of Issue	Date of Return
1	1	2/2/03	5/2/03
1	2	2/2/03	5/2/03
2	4	2/2/03	
4	3	5/2/03	6/2/03
1	5	5/2/03	

Video Database Table

Vid ID	Name of Video
1	The Godfather
2	Apocalypse Now
3	Brazil
4	Some Like It Hot
5	The Crazies

Figure 1.
The outline of a video store database. Each table (relation) consists of records (rows) and fields (columns). Each customer and each video record has a unique ID field. The crucial link table consists of customer ID, video ID and date stamps.

A Reaction Chemistry Relational Database

Chemistry is "the study of matter and its changes" (BRADY, 1993) or the "identification of substances [and] the ways in which they interact" (PEARSALL, 1998). The question is: how can the system of "matter and its changes" or "substances and interactions" —*reaction chemistry space*— can be mapped to a collection of relational database tables?

As any type chemical change can be described by a chemical equation, reaction chemistry space can be defined in terms of chemical equations:

X –> Y or A + B –> C etc.

The chemical equation a powerful metaphor able to describe processes from the composition of equilibrium mixtures to multi-step synthesis. The laws of conservation of mass and conservation of energy can be mapped to chemical equations so they are balanced in terms of stoichiometry, enthalpy and entropy. But this is not a requirement and often equations are not balanced.

Analysis indicates that at least two database tables are required: a chemical species table which holds an appropriate chemical symbol in a picture field and a chemical reactions table which holds information on the group of chemical species which take part in a particular reaction and looks up the corresponding chemical symbols from the species table. A rule emerges from this early analysis which is essential for the referential integrity of the database: a chemical species must be present in the chemical species database table before it can be referred to by the reactions table. Just as in the real world, a chemical reagent must be available in the lab before it can be used in an experiment.

The database structure is more elegant if a link table is placed between the species table and the reaction table (figure 2) because reactions do not have a fixed number of participating species. A reaction may involve just two chemical species: X rearranges to Y, or there may be many if all substrates, reagents, catalysts, solvents, promoters, products and byproducts are taken into account. The link table lists the chemical species associated with a particular reaction with the associated stoichiometry data. As well as looking up symbols from the link table, the reactions table holds information about the reaction conditions (temperature and pressure) and the corresponding equilibrium position expressed as: % yield, K_{eq}, $E°$, pK_a, $t_{1/2}$, $\Delta G°$, etc. A further table to classify reactions by reaction type: nucleophilic substitution, commercially important petrochemical reaction, etc., is also employed.

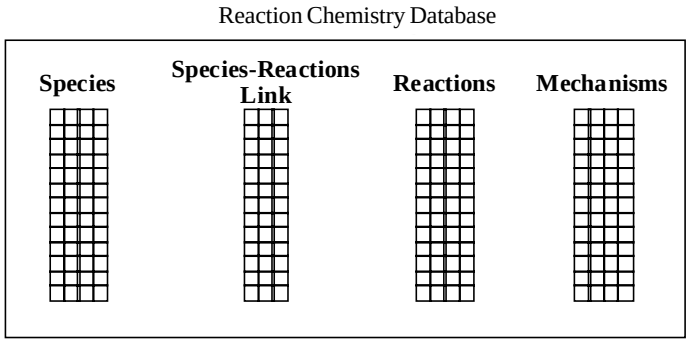

Figure 2.
The main tables of the reaction chemistry database are the Species, Reactions and Reaction Types tables. However, as a particular reaction can have a variable number of participating chemicals species, the design is improved with the addition of a species-reaction link table which links the species' ID number to the reaction ID number.

Additional tables are required:

- Chemical species often have many synonyms: CH_3COCH_3 is known as: acetone, propanone and dimethyl ketone. Therefore a dedicated synonyms table, with relational links to the main chemical species table, is used to hold all name and synonym data.
- A table is added to link generic species, for example aldehyde (generic), with all examples of aldehydes in the database: acetaldehyde, benzaldehyde, etc. This table allows searches to be carried out for all aldehydes, etc.
- Only reagent chemicals, such as acetaldehyde, possess physical properties such as boiling point and density. Therefore physical data is held in a dedicated reagent chemical table.

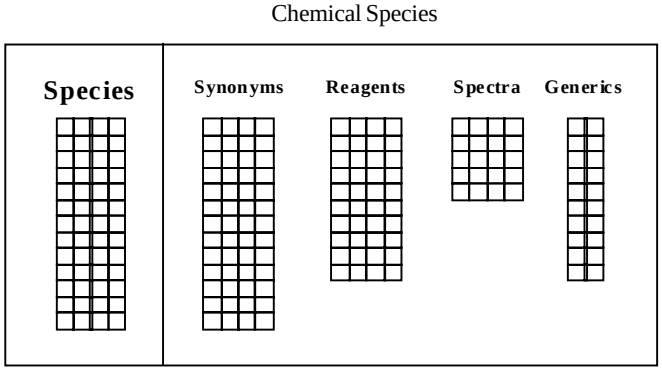

Figure 3.
The chemical species table has a number of subsidiary tables that hold additional data, often specific to a class of chemical species. When the chemical species table is stripped of physical data and names, only four fields remain: species ID, picture, charge and a real/generic flag.

On screen buttons are available which allow the user navigate through the database using relational links so that the software application functions as a hyper-textual chemical thesaurus or encyclopedia, figure 4.

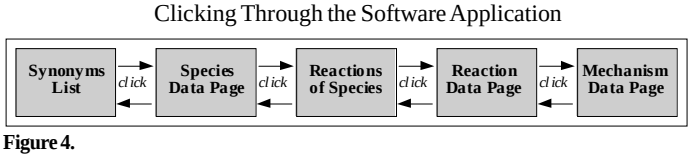

Figure 4.
Users of the software application click on a synonym and go to the species data page which shows the appropriate structural diagram, lists synonyms on a sub-form, gives physical data, etc. The species data page links to all of the reactions in the database in which the selected species participates. Clicking takes the user to related record in the reactions database and on to the mechanisms database. Thus, the user is able to navigate from synonym to species to reaction to mechanism and back from mechanism to reaction to species to synonym.

In its present form, the database (The Chemical Thesaurus 3.1) holds information on 5600 synonyms for 3600 species and 3400 interactions, reactions and other processes. While these numbers are modest compared with the large commercial and academic chemistry databases such as Chemical Abstracts, Beilstein, Gmelin, ISIS, etc., the philosophy of data entry has been to be comprehensive in terms of "simple" chemistry. An attempt has been made to include all of the species and reactions that would be required by a university chemistry major —specialist modules excluded— and to include examples of as many types of species and types of reaction as possible. To date, data entry includes: quarks, fermions, bosons, leptons & selected hadrons; atoms, atomic ions & isotopes; radioactive decay series; simple molecules & molecular ions; main group chemistry; industrial organic chemistry; industrial inorganic chemistry; organic functional groups & reaction chemistry; reaction mechanisms; LEWIS acids, LEWIS bases & LEWIS acid/base complexes; redox agents, radicals, diradicals, photochemistry, pericyclic processes; VSEPR geometries; Brønsted acids & conjugate bases; material types; minerals; flame chemistry; selected natural products; common pharmaceuticals & their classes, etc. The Chemical Thesaurus holds *sample* data on: organic chemistry of real species, synthetic routes, etc., transition metal chemistry, organometallic chemistry and biochemistry. These are truly vast areas of human knowledge and comprehensive coverage is totally outside the scope of the current project. Information on these areas is held in the primary literature as well as commercial and academic databases.

The Chemical Thesaurus also holds all the source data for the author's Patterns In Reaction Chemistry project, including: congeneric series, planars and volumes; an interactive LEWIS acid/base interaction matrix and a set of 12 reaction chemistry tutorials, containing more than 1000 "power point" type slides.

Five chemistry software gadgets have been included. The first two, a clickable periodic table (plus isotopes) and a molecular weight calculator, are self-explanatory. The next three: a redox gadget explore oxidation and reduction reaction chemistry, an aromatic substitution gadget predict the outcome of SEAr and SNAr the reactions of substituted benzenes and a thermochemistry design and explore reactions using the Gibb's equation are more significant. There are areas of reaction chemistry space where it is possible to explain and/or predict reaction chemistry from theory. The data from these regions could be added (hard coded) into the main database— and some of it is—but it is more efficient to build a software gadget to predict the reaction chemistry. The gadgets currently use rather simple levels of theory. As The Chemical Thesaurus develops, more gadgets will be added and the gadgets will use deeper theory, so making the predictions more accurate. For the future it is intended to add more data and to build a layer of "chemical intelligence" on top of the raw reaction chemistry data. This software layer will interrogate the data with an aim of predicting chemical reactivity. The logic will be similar to that employed in Beaker (WERNER, 1993) and CAMEO (SALATIN, 1980) with the difference that the lookup tables will be more comprehensive and will not be restricted to organic chemistry.

REACTION CHEMISTRY: THE *META*-VIEW

Reaction chemistry database development within the rigour of the relational model does more than create a computer application; the analysis process lays bare the meta-structure of the underlying science. One way of representing this meta-view takes is as a map, figure 5, with six regions: species, reactions, reaction types, theory, analytical methodology and the literature. The map has the property that each of the six regions is linked to each other region. For example, analytical methodology is used to determine substrate, reagent and product structure and purity, reaction equilibrium position and the reaction mechanism. The meta-structure map is formally simpler than the database table structure because coddian normal form and link tables are not required.

Species. Anything which participates in a chemical reaction is a chemical species, including: real species, reagent chemicals, generic species and super-generic species:

- All species with mass are *real*. The set of real species includes all: atoms, ions, molecules, molecular ions, radicals, excited state species, etc.
- The set of reagent chemicals, or *chemicals in bottles*, includes: elements, pure compounds, homogeneous mixtures (solutions) and heterogeneous mixtures. Reagent chemicals are always charge neutral and non-transient. Only reagent chemicals can possess properties such as melting point, density or toxicity. The set of reagent chemicals is a sub-set the real species set. For example, sodium chloride, NaCl, is both a real and a reagent whereas the chloride ion, Cl^-, is real but it is not a reagent.

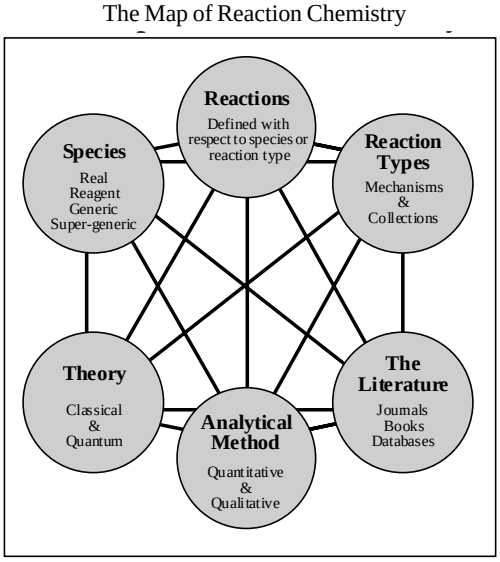

Figure 5.
The main tables of the reaction chemistry database, species, reactions and reaction types appear in the map of reaction chemistry along with theory, analytical methodology and the literature. The six regions of the map are interconnected with each other.

- Generic species are hypothetical entities with a representative structure, for example, dialkyl ketone (generic). In The Chemical Thesaurus, species are defined with respect to any number of appropriate generic species, thus, propanone is an example of a dialkyl ketone (generic). Large multifunctional organic molecules are defined in terms of their generic sub-structures.
- Super-generic species are hypothetical entities with a representative reactivity. Examples of super-generic species include radical, nucleophile and Brønsted acid. Analysis of species in the data set show that many species are Lewis acids, Lewis bases, or Lewis acid/base complexes. The chemistry of Lewis acid/base species is discussed elsewhere (LEACH 1999). Real species and generic species can be classified with respect to one or more super-generics. For example, the proton, H^+, is classified as an electrophile, a hard LEWIS acid and as a single electron transfer oxidising agent.
- Even specialist apparatus, such as a Dean and Stark trap (which is used for removing water), can be considered to be "a species".

Reactions. A chemical interaction or reaction is deemed to occurs when chemical species transform themselves into other, chemically distinct, species. Chemical interactions and reactions can be classified by the species, which take part as substrates, reagents, catalysts, solvents, products or byproducts, etc., or by reaction type (below). Thus, and somewhat surprisingly, the set of chemical reactions is seldom considered *per se*. The set of chemical reactions is potentially so huge that, without the benefit of database technology, the information is difficult to explore. The Organic Syntheses Reaction Guide (LIOTTA, 1991) is one of the few texts, which lists chemical reactions: some 3200 given, classified by reaction type. However, close examination reveals extensive duplication.

Reaction Types. The Chemical Thesaurus groups reactions by *mechanism* and by *collection.*

- The definition of a *mechanism* is clear: "In its most detailed form a reaction mechanism describes, as a function of time, the relative positions of all microscopic particles whose motion is necessary for the reaction to occur" (MOORE, 1981). The term *mechanism* includes: electronic transition, phase transformation, resonance structure interconversion, electron transfer, tautomerisation, ionisation, etc., as well as the classic reaction mechanisms, such as second order nucleophilic substitution and the atom-to-atom mappings associated with name reactions such as the Claisen condensation.
- A *collection* is any particular grouping of reactions: the reactions involved in the synthesis of chloramphenicol, carbon-carbon bond forming reactions, the reactions in a Ph.D. thesis, and so on. While it is formally true that the mechanisms are a sub-set of the collections set, it is convenient to promote mechanisms to an equal footing with collections because of their importance to the history and understanding of reaction chemistry.

Theory is used to model and explain the various phenomena of reaction chemistry in terms of a deeper physical understanding. However, reaction chemistry falls victim to the massive *schism* which exists at the very heart of modern physics, *viz.* the difficulty of reconciling classical mechanics with quantum theory. Quantum theory deals with the very small and the very fast, a world where behavior is described in terms of probable transitions between various states. Classical theory describes the familiar newtonian macro world with its continuous cause-and-effect relationships. Our problem is that reaction chemistry straddles the classical-quantum boundary and takes on board both world-views. However, the classical view is pervasive in reaction chemistry. Every time a line is drawn between two atoms to represent a chemical bond, H-H for example, classical theory of 1916 vintage is invoked (LEWIS, 1916). The very idea of a reaction *mechanism* is classical, even though the species which take part in the mechanism are, undeniably, quantum objects.

Analytical Methodology uses established chemical and physical phenomena to probe real chemical species and their reactions. Analytical chemistry is a comparative, empirical science and all inferences are statistical. Analysis can be qualitative (*what is it?*) or quantitative (*how much of it is there?*). With the careful study of chemical species before, during and after a reaction it is possible to draw inferences about the intermediates, transition states and the reaction mechanism.

The Literature consists of the more than 100 journals which publish papers of interest to chemists and the 20% of all patents that concern chemical compounds or methods for their synthesis. This *primary* literature is comprehensively abstracted by organisations and individuals to produce the *secondary* literature: review articles, monographs, databases, etc. Textbooks constitute the tertiary literature. However, this structure is rapidly changing as the information technology revolution gathers pace.

CONCLUSIONS

The approach to modelling reaction chemistry employed in The Chemical Thesaurus is agnostic to the quantum-classical dichotomy: species are simply records in a database and groupings of species interact with each other. Examples from quark chemistry through to biochemistry are included.

The map shown in figure 5, with its foundations in relational database technology, delineates the essential components of reaction chemistry. While there is something very satisfying about the good "fit" between reaction chemistry and a relational database structure, there are pedagogic implications. There are no real distinctions between main group, organic, inorganic, organometallic and bio chemistries. Different types of chemistry do exist, but they are sub-divisions *within* reaction chemistry.

The final word is part of a review which appeared in *The Journal of The American Chemical Society*:

"The Chemical Thesaurus is a reaction chemistry information system that extends traditional references by providing hyperlinks between related information. This program goes a long way toward meeting its ambitious goal of creating a non-linear reference for reaction information. With its built-in connections, organising themes, and multiple ways to sort and view data, The Chemical Thesaurus is much greater than the sum of the data in its database. The program does an excellent job of removing the artificial barriers between different subdisciplinary areas of chemistry by presenting a unified vision of inorganic and organic reaction chemistry." (COUSINS, 2001)

BIBLIOGRAPHY

BRADY, J.E. and HOLUM, J.R., *Chemistry: The Study of Matter and Its Changes*, John Wiley & Sons, Inc., New York, 1993.

CODD, E.F., *A Relational Model of Data for Large Shared Data Banks.Communications*, Association for Computing Machinery, 13, pp. 377-387, 1970.

COUSINS, K.R., *Review: The Chemical Thesaurus 2*, JACS, 123, 35, pp 8645-6, 2001.

LEACH, M.R., *Lewis Acid/Base Reaction Chemistry. meta-synthesis*, Brighton, 1999.

LEWIS G.N., "The atom and the molecule", *Journal of the American Chemical Society* 38, pp. 762-785, 1916.

LIOTTA, D.C. and Volmer, M., *Organic Syntheseses. Reaction Guide*, 1-854. John Wiley & Sons, New York, 1991.

MOORE, W.J. and PEARSON, R.G., *Kinetics and Mechanism*. J. Wiley & Sons, New York, 1981.

PEARSALL, J. (ed.), *The Oxford English Dictionary*, Clarendon Press, Oxford, 1989.

SALATIN, T.D. and JØRGENSEN, W.L., "Computer-assisted mechanistic evaluation of organic reactions", 1, Overview, *Journal of Organic Chemistry* 45. pp. 2043-2051, 1980.

SCERRI, E.R., *The Failure of Reduction*, Science & Education, 9, pp. 405-425, 2000.

WERNER, J.; BROCKWELL, J.; TOWNSEND, S. and TEA, N., *Beaker*, Brooks/Cole, Pacific Grove, 1990.

Received: 10.05.2003, **accepted:** 23.09.2003