Intuitive and integrated browsing of reactions, structures, and citations: The Roche experience

Pharma Research and Early Development Informatics

*Fausto Agnetti[1], Meeuwis van Arkel[3], Michael Bensch[1], Hermann Biller[1], Martin Blapp[1], Gerd Blanke[1], Jennifer Crovatto[2], Ben Cheikh[2], Joerg Degen[1], Bernard Dienon[1], Thomas Doerner[1], Gunther Doernen[1], Frieda Farshchian[1], Werner Gotzeina[1], Peter Hilty[1], Ralf Horstmoeller[1], Thomas Jeker[1], Brian Jones[1], **Michael Kappler**[2], Aslam Momin[2], Antonio Regoli[1], Denis Ribaud[1], Roger Sayle[4], Bernard Starck[1], Daniel Stoffler[1], Klaus Weymann[1], Padmanabha Udupa[2].*

(1) F. Hoffmann-La Roche Ltd., Basel, Switzerland, (2) Hoffmann-La Roche Inc., Nutley, NJ, Unites States, (3) Elsevier Information Systems GmbH, Frankfurt, Germany, (4) NextMove Software Ltd., Cambridge, United Kingdom

1 – Background

Abstract. Roche has integrated propriety reaction information within the Elsevier Reaxys product, which will run on Roche's infrastructure and inside the Roche firewall to provide high performance and security. The incorporation and discoverability of proprietary information along with public information significantly improves productivity. With this development, Roche researchers are able to launch a single search in Reaxys across integrated internal data and experimental data published in journals and patents, with results unified and organized in a context directly relevant to the researcher workflow. Key points of ELN integration, data modeling, and reaction canonicalization will be discussed.

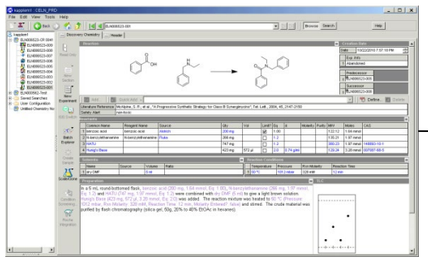


Figure 1: Chemistry Electronic Laboratory Notebook (CELN) database.

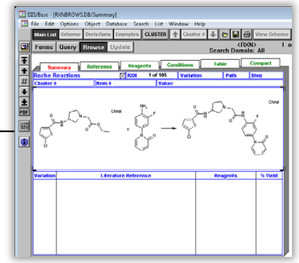


Figure 2: ISIS/Host databases .

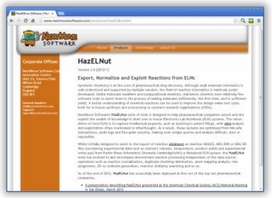


Figure 3: Tool to gather CELN reactions and data.

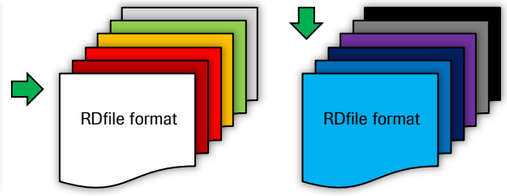
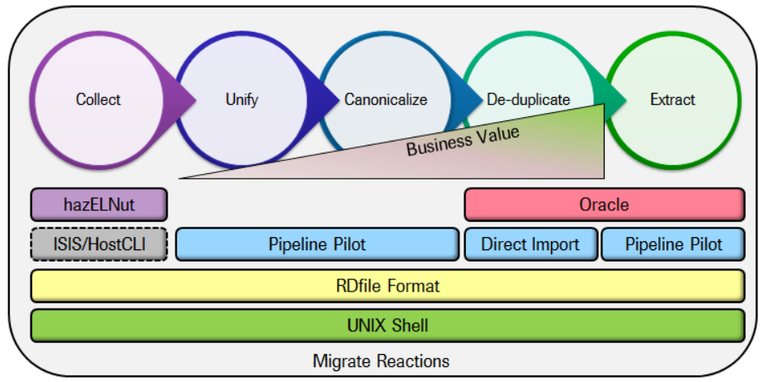


Figure 4: Each database is represented in a different RDfile format (reflected as color).

2 – Concept, Components, and Process



There are five (5) steps to reaction migration. The first step is **COLLECT** reaction data. The second step is **UNIFY** to a single, common format. The third step is **CANONICALIZE** to a unique chemical representation. The next step is **DEDUPLICATE** by grouping reactions into variations and aggregating structure properties. The last step is **EXTRACT** records from the database. The entire process is scripted on Unix and uses the RDfile between steps. Accelrys' ISIS/HostCLI is used to collect one-time from legacy systems and NextMove's HazELNut is used to collect daily from the PerkinElmer ELN. Accelrys' Pipeline Pilot (PP) is used to unify, then canonicalize, and Accelrys' Direct cartridge is used to deduplicate. Accelrys' Integrated Data Source (IDS) via PP is used to deliver data into Elsevier's Reaxys application.

3 – Unified Data Model

The Unified Data Model (UDM) objects are **reaction**, **structure**, and **citation**. A reaction is recognized by the chemical structure of reactants and products only. *Variations* contain experimental data as well as reagents, catalysts, and solvents. Keyword value pairs are used for comments, identifiers, and links,

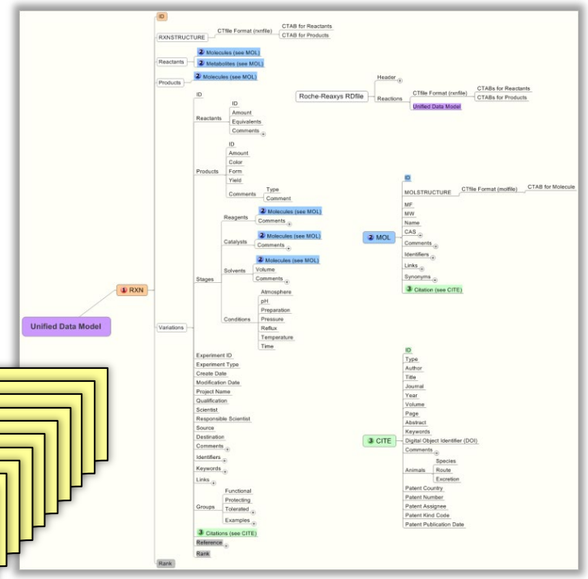


Figure 8: The UDM represents reactions, structures, and citations.

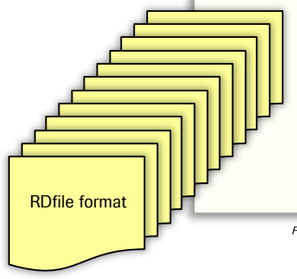


Figure 7: All RDfiles are reformatted to the UDM (reflected as one color).

4 – Canonicalization

To canonicalize means:

- **Normalize.** A sketch is independent of how the sketch is drawn.
  - ü A nitro group can be drawn with either 4 or 5 bonds to nitrogen.
- **Identify.** The role of a component may be reflected as a reaction variation.
  - ü A+B->C in solvent D is grouped with A+B->C in an unspecified solvent.
- **Order.** The sequence of components is independent of how the data is stored.
  - ü A+B->C is equivalent to B+A->C.

Prototype with Pipeline Pilot using standard and extensible components.

1. Get molecules – split reaction into components with a role of reactant or product.
2. **Normalize** graph – perceive groups/salts by SDfile or Cheshire scripts (or both).
3. Represent as *SMILES* – generate a systematic and unique string (canonical).
4. **Identify** roles – separate agents from reactants and by-products from products.
5. **Order** by property – generate a systematic and unique attribute (molecular weight)
6. Construct string – concatenate reactants followed by products (normalized, ordered).
7. Set reaction – recreate reaction from component (reactants and products).

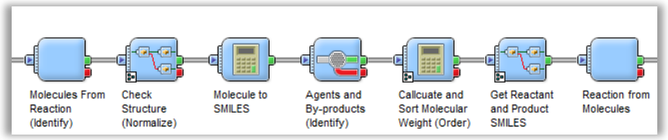


Figure 9: Prototype process in Pipeline Pilot. Actual process uses the Accelrys Direct cartridge.

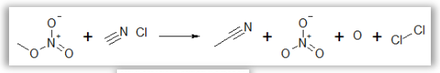


Figure 10: Example reaction #1.

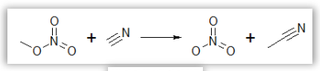


Figure 11: Example reaction #2.

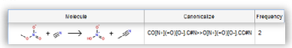


Figure 12: Both reaction #1 and #2 canonicalize to the same representation.

5 – Reaxys for Roche

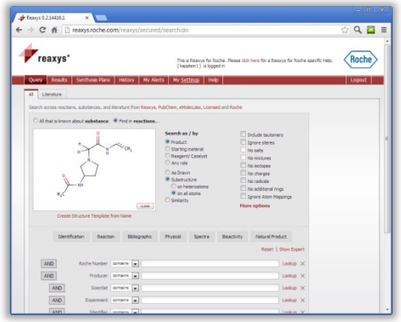


Figure 13: The Query Form supports Roche fields and all common sketchers.

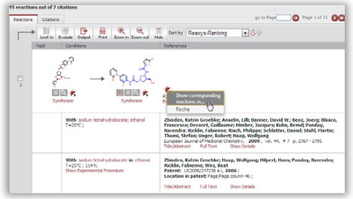


Figure 14: Results are cross-linked by reaction, structure, and citation.

Figure 15: Structures are linked to chemical registration.

Figure 16: Experiments are linked to complete documentation.

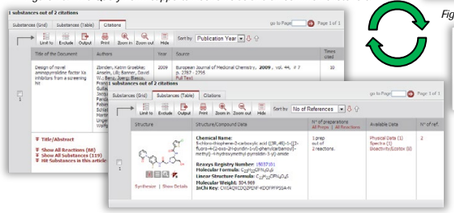


Figure 17: Interact with your history using bread crumbs.

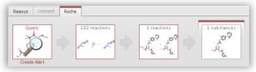


Figure 18: Interact with your history using bread crumbs.

Figure 19: Query by name or number.

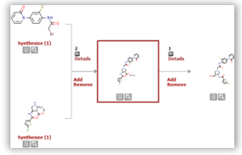


Figure 20: Synthesis planning is integrated and interactive.

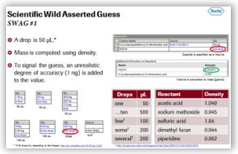
6 – Discussion

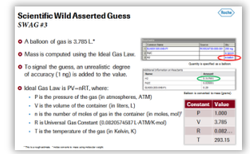
User Feedback – The application has been well-received and significantly improves productivity. Quotes from leaders in chemistry include: "*A really good application*", "*A gift from the heavens*", "*Outstanding*", and "A *colleague praised Reaxys for Roche with words which cannot repeat in written (sic)*".

Data Sets – Millions of reactions and structures were processed from *in-house* and licensed sources.

Performance – Execution time of PP on the epoch data sets was prohibitively long. We used NextMove's PP Accelerator to improve performance 100x.

Quality – Vague measurements, misspelled units, and keyboard translations were resolved.





Automation – An Oracle table was used to "remember" the current collection state. After an "epoch" collection, the incremental collection is specified using parameters '-since LASTTIME –until NOW'.

Future Possibilities – NextMove is developing an "reaction name" tool that will enable query by reaction type, e.g. Claisen, Suzuki, Wittig, etc. Also, integration could be extended further to include *in-house* inventory and commercial availability. Currently, only Reaxys content and structures cross-linked to them show availability. Finally, PerkinElmer's latest ELN appears to support a hyperlink to a notebook page.