InChI as a Publishing Application

Dr. Bernd Berger
Graeme Whitley
Background

• In additional to journals, books, major reference works in chemistry, Wiley publishes a series of spectroscopy databases
• Chiefly in mass spectrometry, C- H- X-NMR, and FT-IR
• Mainstays in QA, forensics, toxicology labs
• Real world data –complex
The Wiley Registry 8th

• A Considerable Challenge
  – Identify replicates, duplicates
  – Validate internal data integrity
  – New compounds, novel compounds
  – Add structures to the library
  – Consistency checks
  – Metabolites, precursors
Identify Duplicates & Replicates

- Wiley Registry = 399,385 spectra or 160 billion comparisons
- Wiley/NIST = 399,385 x 195,000, or 76 billion comparisons
- Translates to 4 months of processing
- A mistake in the script would require starting over from the beginning
Adding New Structures

• Validating names, formula, mass, spectra, CAS RN, Reg ID, other identifiers
• Several months processing time drawing Mol files
• Several months processing time in QA
• Needed to improve library and to aid in downstream processing
Chemistry Publishing Challenges

• Alphabet soup of formats
  – Formats from CambridgeSoft, InfoChem, BioRad, NIST, ACD, Agilent, Waters, PerkinElmer, MDL/Symyx, Daylight, IUPAC
  – Proprietary registration systems
    • CAS RN, PubChem ID, canonical Smiles

• Naming conventions
  – CAS, IUPAC, local names, trade names, synonyms, etc.
Chemistry Publishing Challenges

• Software vendor’s perception of publishers
  – Upgrade Treadmill
  – The license that costs more than it would to develop yourself

• Disruptive behavior by hardware/software companies

• Customer expectations/requirements
Why Not Rely on CAS RNs?

• Not a registry of compounds, per se
• Records can be and are superseded
• Contain no useful information – requires linked data
• Proprietary – only partial coverage
• Retrospective – does not help with novel compounds
• Incomplete – only covers journals and patents
Why Not Rely on Smiles or Mol Files?

• Smiles
  – Proprietary, many versions, just not on our radar

• Mol
  – Matching not as fast as we would like, software not geared for high-throughput
Why Not Rely on Pubchem IDs

• Same issue as with CAS Registry
• In addition, no sense at the time that there was enough quality control.
Where We Were

• Several months in processing to generate and validate structures
• Issues in comparing spectra to identify and differentiate duplicates from replicates
• Many months in spectral comparisons
• No possibility of making the launch
To: Individuals Interested in the IUPAC Chemical Identifier.

A new test version of the IUPAC-NIST Chemical Identifier (INChI) is now available. It replaces the previous test version issued last November. All features planned for inclusion in the final release have now been implemented and the final format for Identifier has been proposed (it is simply a text string). Also, the Identifier called INChI (formerly IChI) to acknowledge the development work at NIST. The test program accepts input in the form of MOLFiles (or SDfiles) and CML files. An Application Program Interface (API) for communicating with external programs is under development (see later) . . . .

We very much welcome comments concerning the INChI and will be glad to assist you in its testing or implementation.

Steve Stein  
Dmitrii Tchekhovskoi  
Steve Heller  
NIST  

Alan McNaught  
Royal Society of Chemistry
Off to the Races

- Initial Reaction at Wiley
  - Check this out
  - What the heck are Steve and Gary up to?
  - Possibly of use
  - Could be used to solve the new compounds problem
  - Could be a lot faster than the software from traditional vendors
Problem Solving

• Unique compound registry
  – Enforcing Data Integrity
  – Multiple field matching
  – Break up big problems into smaller problems

• Generate the Mols, generate the InChIs

• Unique InChI registry
  – Match to spectrum records
  – Now matching for replicates can begin
  – . . . Or can it.
Problems Arise

• We still needed our own compound registry
  – But . . . InChI’s are not necessarily unique
    • Depend on structural representation
  – Don’t really do anything more than Mol files
  – Why use an untested technology?

• We needed speed
  – Long strings made computation inefficient
  – Memory size problems
  – But . . . structure-centric programs were not amenable to the processing we needed to do
The InChI Key Precursor

- Matching required exact matches
  - Strings were too long
    - Matching was slow
    - Limitations in matching long strings
    - Operating system limitations
- InChIs on their own were a Dead End
- How to shorten an InChI
  - Compressed
  - Unique
- Use An Old Approach From Web Publishing
  - MD5 Hash
    - Compressed
    - Unique
Avoiding Non-Unique MD5 String

• Rely on Other Database Hooks
  – Names
  – Molecular weight
  – Chemical formula
  – Mol
  – InChI
  – Spectrum
  – Other Identifiers
    • CAS RN, Wiley RegID, etc.
InChI Is Part of the Process
Where Are We Headed?

• Continue to build up main compound file
• Identifying and normalizing data in full text
• Applications relating full text, enhanced searching, enhanced linking
Where We Are Not Headed

• We will not be building things for the sake of building them
• Creating links for the sake of making links
The Net Result

• Break up large problem
  – Net reduction in direct comparisons by around 95%

• Look to other disciplines for solutions
  – MD5 compression allowed us to compare compounds rapidly
Why Does This Matter?

• Borders and customs
• Homeland security
• Drug research
• Toxicology and Forensics
Thank You

Visit Wiley-Blackwell at the ACS Booth 1501