Provenance in Chemistry

Stuart Chalk

Questions

1. What is the typical provenance and data transformation information that your domain needs to capture? Or What is data provenance in your domain?
   a. Varies with subdiscipline, but in general has typically been very high level info - paper, author, general instrument type, apparatus used, chemical/mixture (CASRN or IUPAC Name/InChI SMILES), samples/ing
   b. Analytical is probably the most detailed -> spectra -> JCAMP-DX (lab, inst type, software version, etc.)
2. Is there a practice around provenance information? If so, how is it captured and shared?
   a. Practice has been to include general information in the materials/methods section of research papers
   b. Less of a focus on instrumental details in paper more on chemicals (manufacturer, cat #, lot #, CoA)
3. How widespread is it? How much of the domain has a shared or best practice? What is the demand in your domain?
   a. All chemists are taught how to write a laboratory notebook, research story, include information about instruments, equipment, chemicals, instrument settings, important expt. conditions, and observations
4. What semantics and tools, software are used to do this in your domain?
   a. We are far behind on semantics. We are working toward this with the GB but we have published definitions of concepts from PAC in over the last 50 years… (legal)
   b. In the industrial arena there has been a lot of work on LIMS and ELNs to manage data especially relative to legal requirements, but these are expensive and have not made it into many academic research labs
   c. There are open standards JCAMP-DX (FAIR Spectral Data Project) that allow open use…
   d. ThermoML for thermophysical data

Questions

1. What is the role of fine grained process (transformations, normalisation etc) vs the contextual, provenance information?
    a. Fine grain processes (data processing) are very important in processing raw data (spectra) from instruments and there is now more focus on this...
2. What is the level of granularity about provenance? What do users want in terms of provenance information, and at what level of detail?
    a. Top level contextual info - conditions, chemical/system, how conditions where controlled, SOP
    This has typically been at level where the work can be ‘reproduced’ but only i terms of the expt. Not the data...
3. What demand is there for provenance and process information?
    a. Its growing but has not been a focus on data in Chemistry the way we are thinking about this now…
    b. ThermoML for thermophysical data
4. How do researchers / users use provenance and process information?
    a. One example -> critical evaluated data (solubility) from IUPAC

Chemical Data

- Physical/Chemical Property Data - single data points, series of data points
- Reaction/Reactivity Data - reaction yields, products, conditions, kinetics
- Spectral Data - UV/Vis, IR, MS, NMR (raw/processed, JCAMP-DX)
- Chromatographic Data - 2D/3D GC, LC, IC, SFC
- Data from Hyphenated Techniques - Chromatographic + Spectral
- Crystallography Data - raw? Processed file (CIF)
- Surface Analysis Techniques - 3D SEM, AFM, etc.
- Chemical Structures - .molfile, .sdf
- Chemical Identifiers - names, InChI/InChIKey, SMILES, CASRN
- Other Identifiers - RInChI (reactions), MInChI (mixtures)

Chemical Metadata

- Conditions of an experiment
    - Timing
    - Error in measured values
- Chemical substance/mixture studied
    - Supplier, storage conditions, precautions of use, catalog #, lot #, CofA
- Organism studied
    - Strain, source, storage conditions, precautions of use
- Apparatus used
    - Manufacturer, model, any special features
- Instrumentation used
    - Calibration - standards, standard reference materials, line of best fit, repeatability
    - Settings, installed software and version, any post-processing involved
- Sample
    - Sample type, location, collection, processing, preservation, storage
- Procedure
    - Detailed instructions, standard operating procedure, standard method
- Safety information
    - Hazardous chemicals, safety data, specific instructions for handling/storage/ventilation

Chemical Standards

- IUPAC
  - Naming conventions (Blue Book, Red Book)
  - JCAMP-DX - text based spectral data format (current project for FAIR spectra)
  - ThermoML - XML format to represent thermophysica data (new revision in development)
  - InChI (the International Chemical Identifier) - line notation of chemical structure
    - InChiKey (hash of InChI), Reaction InChI, Mixture InChI, many others in development
  - Critically evaluated data - solubility, reaction kinetics of polymers, others
  - The Gold Book - online definitions of chemical concepts (currently under expansion) (terms published (PDF) in IUPAC Recommendations on terminology)
- Community
  - .molfile, .sdf file - text based formats for representing molecular structure
  - Simplified Molecular-Input Line-Entry System (SMILES) - line notation for chemical structure
  - The Analytical Markup Language (AnIML) - XML specification for spectral files

Chemical Data Repositories

- Zinc 15 () - ~ 1 billion purchasable compounds, calculated data
- PubChem (US NLM) - 110 chemical substances, highly linked, provenance
- ChemSpider (RSC UK) -
- Cambridge Structural Database (CCDC UK) - 1.1 million crystal structures
- Common Chemistry (ACS US) - 500K regulatory chemicals with identifiers
- ChEMBL (EBI UK) -
- DrugBank () -
- Chemistry Webbook (NIST US) -
- Wikidata () -

Tools/Services Needed to Support Chemical Data Provenance

- Chemical substance resolver - unique reference points for all substances
- Chemical identifier resolver - many names/identifiers one compound
- Large corpus of critically evaluated data - as reference
- Service to allow accurate calculation of molar mass of compounds
- Digital certificates of analysis for reference and calculations
- Instrument identifiers, standard methods of capturing instrument settings
- Online reference points (unique ids) for physical constant values
- Standard open formats for instrument spectra (in progress)...
- ...and standard spectral processing format
- Online chemical concept vocabulary/ontology for semantic applications
- Referenceable chemical safety information