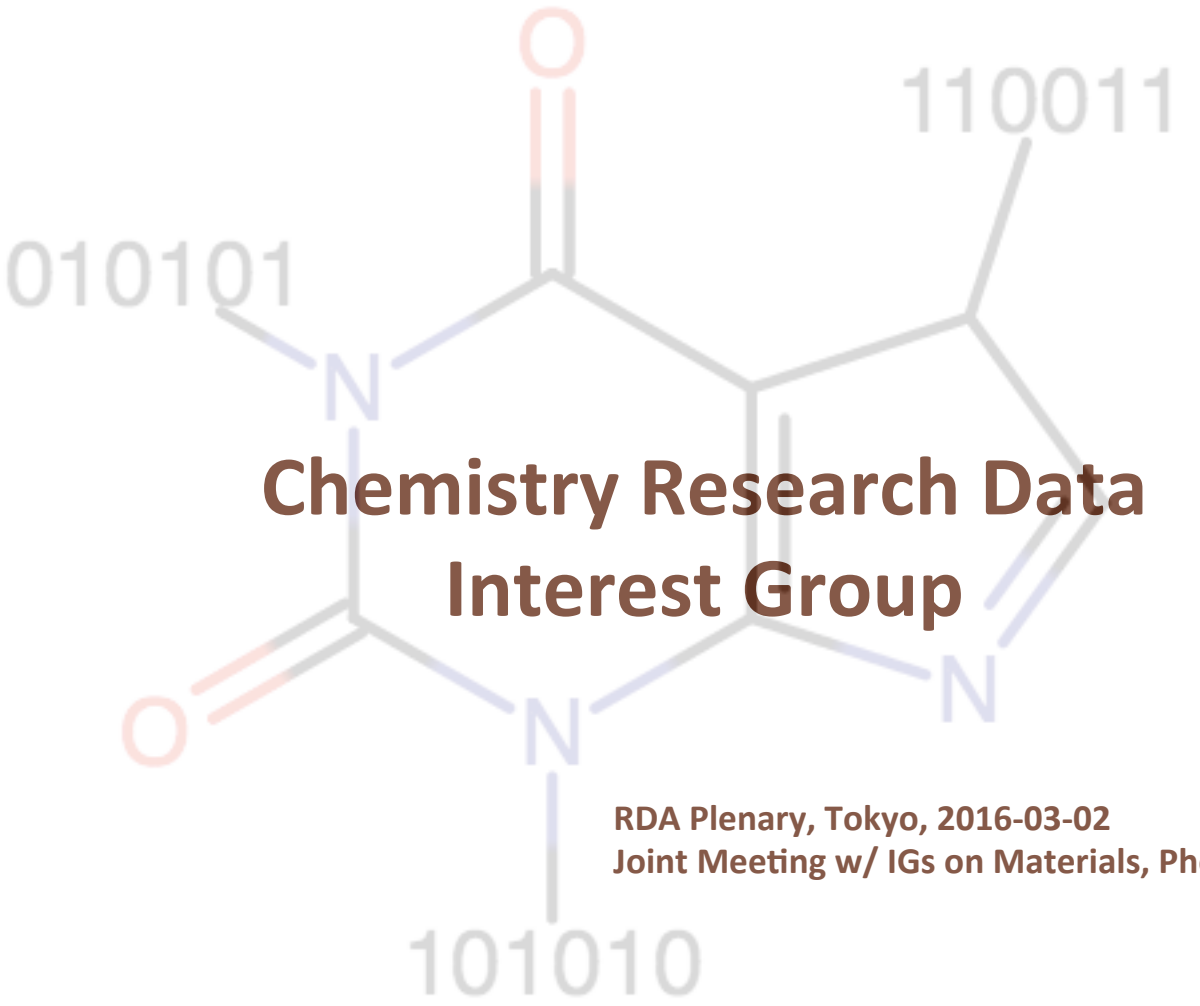

Chemistry Research Data Interest Group

RDA Plenary, Tokyo, 2016-03-02
Joint Meeting w/ IGs on Materials, Photons & Neutrons

https://rd-alliance.org/groups/chemistry-research-data-interest-group.html

DIG Chemistry!

- **ACS Division of Chemical Information**
  - Professional networking in the chemical and information sciences
- **RDA Chemical Research Data Interest Group**
  - BoF at RDA Plenary 6 (Sep 2015)
  - Dinner discussion at Pacifichem (Dec 2015)
  - Endorsed by RDA (Feb 2016)
- **International Union of Pure and Applied Chemistry**
  - Mission: "IUPAC provides leadership, facilitation, and encouragement of chemistry and promotes the norms, values, standards, and ethics of science and the free exchange of scientific information."
  - Committee on Publications and Cheminformatics Data Standards

RDA Interest Group aims

- To be the bridge between the chemistry community and wider research data initiatives via the RDA hub
  - Benefit from relevant data activities (i.e., citation, publishing workflows, etc.)
  - Establish a chemistry presence in RDA for inter-disciplinary needs related to chemical information
- Map the landscape of challenges & opportunities for chemistry digital data to promote and coordinate collaborations
- Germinate Working Groups from our collective expertise and steward enabling projects for filling gaps

How can we benefit from RDA?

- Best practice, guidance and collaboration on several discipline-inclusive challenges:
  - Format interoperability and normalization
  - Domain specific repository development and sustainability
  - Experimental level metadata structure and handling
  - Semantic classification and vocabulary harmonization
  - Data citation, provenance review criteria
  - Automated and community engaged curation
  - Electronic deposition, validation and publication workflows
  - Data management roles and guidance from collection to publication
  - Funding agency expectations internationally
  - Data management plans for sharing and preservation

What can we bring to RDA?

- Expertise working with domain-based data resources:
  - Wide variety of heterogeneous data overlapping with physics, biology, medicine and engineering
  - Discipline practices in pre and post- publication data collection, compilation, curation and critical evaluation
  - Long history of domain-based systematic description
  - Domain sourced standard scientific research terminologies, definitions and data formats
  - Rich diversity of implementations, applications and domain problem solving approaches
  - Network of community and domain based national and international research and professional organizations

Initial IG tasks and deliverables

1. Review of funding agency requirements specific to chemistry research
2. Inventory of current chemistry publisher recommendations (noting specific instrumentation formats or repositories)
3. *Catalogue of standards relevant to chemistry*
4. Compilation of chemistry data management best practices and instruction materials

Brainstorming WG projects

- Establish recommended digital data management workflows and tools for research chemists and data publishing stakeholders
- Machine-readable input/drawing guidelines for chemists and software designers based on chemical structure representation standards, including edge-case examples, technical validation and testing suites, and designations for types of molecules covered, to better enable interpretable chemical structure identification and normalization
- Recommendations for establishing a standard geometric molecular structure representation format that is read/writable among chemical representation software, interoperable between applications and supports semantic inference
- Compile use cases and business scenarios for machine-usable, semantically-enabled digital iterations of experimental terminologies as represented in the IUPAC color books and other domain vocabularies

Chemical Representation/Identification

- Machine readable
- Canonical
- Interoperable
- Supports normalization
- Semantically enabled
- Open standard
- Implementation validation
- Beyond 2D formulation of small molecules (e.g., macromolecules, multi-component systems, geometric structures)

- InChI
- Chemical Markup Language (CML)
- Molfile (2D)
- SMILES (proprietary)
- SDF
- ChEBI ontology
- IUPAC Nomenclature

Chemical Characterization

- Properties
  - Physical, chemical, hazard, incompatibility, toxicity
  - Mixtures? Formulations?
  - Polymeric?
  - Administrative parameters (occupational thresholds)
- Experimental parameters
  - Biomacromolecules

- CIF (crystals)
- JCAMP-DX (spectroscopic data)
- Splash - The Spectral Hash Identifier
- STRENDA (enzymology), MIRAGE (glycomics)
- IUPAC Green Book? (quantities, units, symbols)

Chemical Process Description

- Experimental process
- Measurement parameters
- Sample description/ preparation
- Observation/outcome description
- Data analysis
- Reaction transformation
- Equipment/apparatus
- Laboratory/environmental parameters

- Metadata used in data models (e.g., ore-Chem)
- XML standards (e.g., AnIML, S88)
- Methods ontologies (e.g., ChMO)
- Analytical terminology (e.g, IUPAC Orange Book)
- Incident analysis (e.g., BowTie)

Chemical Roles Classification

- MeSH – Medical Subject Headings
  - NLM controlled vocabulary thesaurus used for indexing articles for PubMed
- ChEBI – Chemical Entities of Biological Interest
  - freely available dictionary of molecular entities focused on 'small' chemical compounds
- GHS – Global Harmonizaton System
  - internationally-harmonized approach to classification and labelling to support national programs for safe use, transport and disposal
- IUPAC Gold Book – Compendium of Chemical Terminology
  - chemical nomenclature, classifications terminology, symbols and units
- Manufactured uses –
  - EPA Chemical Data Reporting industry and consumer use codes

General Sources of Potential Metadata

- International chemical collaborations (e.g., IUPAC)
- Pharma collaborations (e.g., Allotrope, Pistoia, OpenPHACTS)
- National informatics initiatives (e.g., NCBI, EBI)
- National libraries (e.g., NLM, NAL)
- Scientific publishers and professional societies (e.g., RSC, ACS, technical divisions)
- Regulation (e.g., UN GHS, EC, US EPA)