Other Related IUPAC RDA Updates

David Martinsen
Subcommittee on Cheminformatics Data Standards (SCDS)
RDA Chemistry Research Data Interest Group (CRDIG)
InChI Workshop @ NIH, August 16, 2017

Agenda

- What is SCDS?
- What is CRDIG?
- SCDS Activity
- CRDIG Activity

Supporting Digital IUPAC

“a consistent global framework for Human AND Machine-readable (and “understandable”) chemical information in collaboration with other science communities, industry, and governments”

How best to disseminate and deploy IUPAC chemical data standards and related assets to support this digital framework?

CPCDS

Committee on Publications and Cheminformatics Data Standards
(approved by IUPAC Council August 2015)

(i) To advise the President, Executive Committee, other Standing Committees, Divisions, and Commissions on all aspects of the design and implementation of publications and **data-sharing,** including computerized databases of all sorts, and to promote the compatibility of the electronic transmission, **storage, and management of digital content through the development of standards for the creation of a consistent and interoperable global framework for human and machine-readable chemical information**.

SCDS

Subcommittee on Cheminformatics Data Standards
(approved by CPCDS July 2016)

The Subcommittee on Cheminformatics Data Standards (SCDS) will monitor the cheminformatics data standards needs of the chemical community **with the objective of coordinating the collective expertise of relevant IUPAC Divisions and Committees as well as collaborating with external global organizations, such as NIST, CODATA, the Research Data Alliance and others in order to prioritize and efficiently meet those needs through IUPAC projects and those of related organizations.** SCDS will promote this work of IUPAC through symposia, workshops, presentations at relevant conferences, and articles in *Chemistry International* and other appropriate publications. To ensure IUPAC-wide communication, SCDS will circulate its reports to all IUPAC Divisions and Committees.

SCDS Members

Dr. Gregory M. Banik, USA
Bio-Rad Laboratories, Inc.

Dr. Ian Bruno, GBR
Cambridge Crystallographic Data Centre

Dr. Stuart Chalk, USA
University of North Florida

Dr. Antony Davies, GBR
AkzoNobel Chemicals bv

Dr. Chris Fellows, AUS
Royal Australian Chemical Institute

Dr. Jeremy Frey, GBR
University of Southampton

Dr. Robert Lancashire, JAM
University of the West Indies (retired)

Ms. Leah McEwen, USA, co-chair
Cornell University

Dr. David Martinsen, USA, co-chair
David Martinsen Consulting

Dr. Claudio dos Santos, BRA
Federal University of Ouro Preto

Dr. Ron D. Weir, CAN
Royal Military College of Canada

Dr. Andrey Yerin, RUS
ACD/Labs

CRDIG

The Chemistry Research Data Interest Group, is intended to foster diverse professional exchange on issues particular to data originating from the field of chemistry. Chemistry, as one of the central sciences, has fundamental impact on the fields of health, pharmaceuticals, materials, energy and many other applied sciences. There is a wealth of chemical data in various heterogeneous formats, distributed across a myriad of systems with endless potential for reuse in chemistry research and many related domains. However, many social, technical and administrative factors have limited the opportunities for open sharing and interoperable exchange.

The RDA mission is to build the social and technical bridges that enable open sharing of data. Organizing a forum for professional exchange directed at addressing opportunities and challenges for chemistry data management within the RDA framework will support international participation across a broad range of stakeholders and foster connections with data types and user scenarios in many disciplines. Bringing in IUPAC (International Union of Pure & Applied Chemistry) as co-sponsor of the group would clearly bridge the activities of this group between those of RDA and the responsible standards body for chemistry.

CRDIG Co-Chairs

Dr. Ian Bruno, GBR
Cambridge Crystallographic Data Centre

Dr. Stuart Chalk, USA
University of North Florida

Mr. Richard Kidd, GBR
Royal Society of Chemistry

Ms. Leah McEwen, USA
Cornell University

Dr. David Martinsen, USA
David Martinsen Consulting

SCDS Activity

- **Assemble directory of existing** chemical digital data and information **standards**, within IUPAC and other scientific standards organizations (initially targeting RDA/CRDIG, NIST, and CODATA);
- **Review the data communication needs** of the chemistry community and compare to existing standards on a regular rotation to identify necessary updates and gaps;
- **Promote working groups** around these areas, including identifying topical experts willing and able to contribute, through collaboration and coordination with related efforts in other venues (i.e., clearinghouse)
- *https://iupac.org/body/036*

SCDS Activity

- IUPAC Gold Book Production Model
- IUPAC Color Book Updates
- IUPAC Solubility Data Series
- IUPAC Critically Evaluated Polymer Properties
- NMR Formats, Detection Limits
- NMR Repository Recommendations

SCDS Activity

- Symposium on Big Data, Research Data, and Chemistry at the IUPAC World Congress, São Paulo, July 13, 2017

- Special Issue of Chemistry International

A Cheminformatics Handbook

- A mechanism for dissemination of the SCDS deliverables
- Community discussions scheduled so far:
  - SCDS meeting at the IUPAC General Assembly, Brazil, July, *COMPLETED*
  - Breakout session at InChI Users Meeting, Washington DC, August
  - Chemistry Data IG session at the RDA 10th Plenary, Montreal, September
  - Chemistry data session at CODATA 2017, St. Petersburg, October

2016 Chemistry Data Events

RDA 7th Plenary Meeting, Tokyo, March 2016

- *Inaugural session of RDA CRDIG, present at several Tokyo area universities*

ACS Spring 2016 Meeting, San Diego

- *CINF Chemistry Data Summit*

RDA-US/IUPAC Joint Workshop, North Carolina, March 2017

- *Prioritizing Digital Data Challenges in Chemistry*

ACS Fall 2016 Meeting, Philadelphia

- *CINF symposium on "Chemistry Data for the People"*

RDA 8th Plenary Meeting, Denver, September 2016

- *International Data Week, CPCDS presentation at International Unions session*
- *CRDIG session on "Chemical Terminology: the IUPAC Gold Book"*

2017 Chemistry Data Events

EMBL-EBI Industry Programme Workshop, March 2017

- *Expanding IUPAC Standards for Chemical Information*

ACS Spring 2017 Meeting, San Francisco – RDA CRDIG Open Meeting

- *Advancing Chemical Information Standards*

RDA 9th Plenary Meeting, Barcelona, April 2017

- *Contributions to sessions on interoperability across disciplines*

Beilstein Symposium – Open Science and the Chemistry Lab of the Future

- *22 – 24 May 2017, Rüdesheim, Germany*

RSC-CICAG meeting on Structure Representation, Liverpool, 22 June 2017

IUPAC World Congress, Sao Paulo, July 2017

- *Special Symposia: Research Data, Big Data, and Chemistry*
- *Chemistry International Special Issue:* https://www.degruyter.com/view/j/ci.2017.39.issue-3/issue-files/ci.2017.39.issue-3.xml

InChI/IUPAC Workshop, NIH, Maryland, August 16-18, 2017

- *in conjunction with the InChI Trust*

ACS Fall 2017 Meeting, Washington DC

- *Joint symposium on "Open Structures": CSA Trust, ACS (CINF), RDA (CRDIG), IUPAC (CPCDS)*

RDA 10th Plenary Meeting, Montreal, September 2017

- *Joint session on "Domain Repositories – Best Practices": RDA (CRDIG, Biodiversity IG), IUPAC (CPCDS)*

CODATA 2017, St. Petersburg, October 2017

- *Proposed session on "Developing & Adopting Digital Standards for FAIR Exchange of Chemical Data"*

Existing IUPAC Digital Resources

- **JCAMP-DX** – spectra data file format (SCDS, several extensions, in need of update)
- **InChI** – chemical identifier (InChI Trust, several extensions)
  - **RInChI** – reaction identifier
- **ThermoML** – thermo-property data markup (NIST, current project revision)
- **Gold Book** – compendium of IUPAC terminology (SCDS, current project revision)
- In principle: 2013 Blue Book, Nomenclature for Organic Compounds
  - Hierarchical criteria for preferred IUPAC name (PIN) allows for more systematic encoding of rule-sets in computer algorithms

External Chemical Digital Data Standards

- **AnIML** – analytical data markup (ASTM, pre-release)
- **Allotrope Data Format** – laboratory data framework (Allotrope, pre-release)
- **CIF** – Crystallographic Information Framework (IUCr, regularly reviewed)
- **PDB** – Protein Data Bank format (RCSB, regular updates)
- **HELM** – Hierarchical Editing Language for Macromolecules (Pistoia Alliance, current update project)
- **CTAB** files (MOLfile, SDFile, RXNfile, etc) – digital chemical representation files (Biovia, documentation available)
- **SMILES** – linear chemical representation (Daylight, documentation available; OpenSMILES effort)
- **CHEMINF** – chemical entity information ontology (open, some updates)
- **ChEBI** – chemical entity ontology (EBI, regularly updated)
- **CML** – chemical markup (Cambridge, status?)
- **STRENDA** – standards for reporting enzymology data (Beilstein, active)
- **MIRAGE** – minimum information required for a glycomics experiment (Beilstein, active)
- several other less formal, less complete efforts

CRDIG Activity

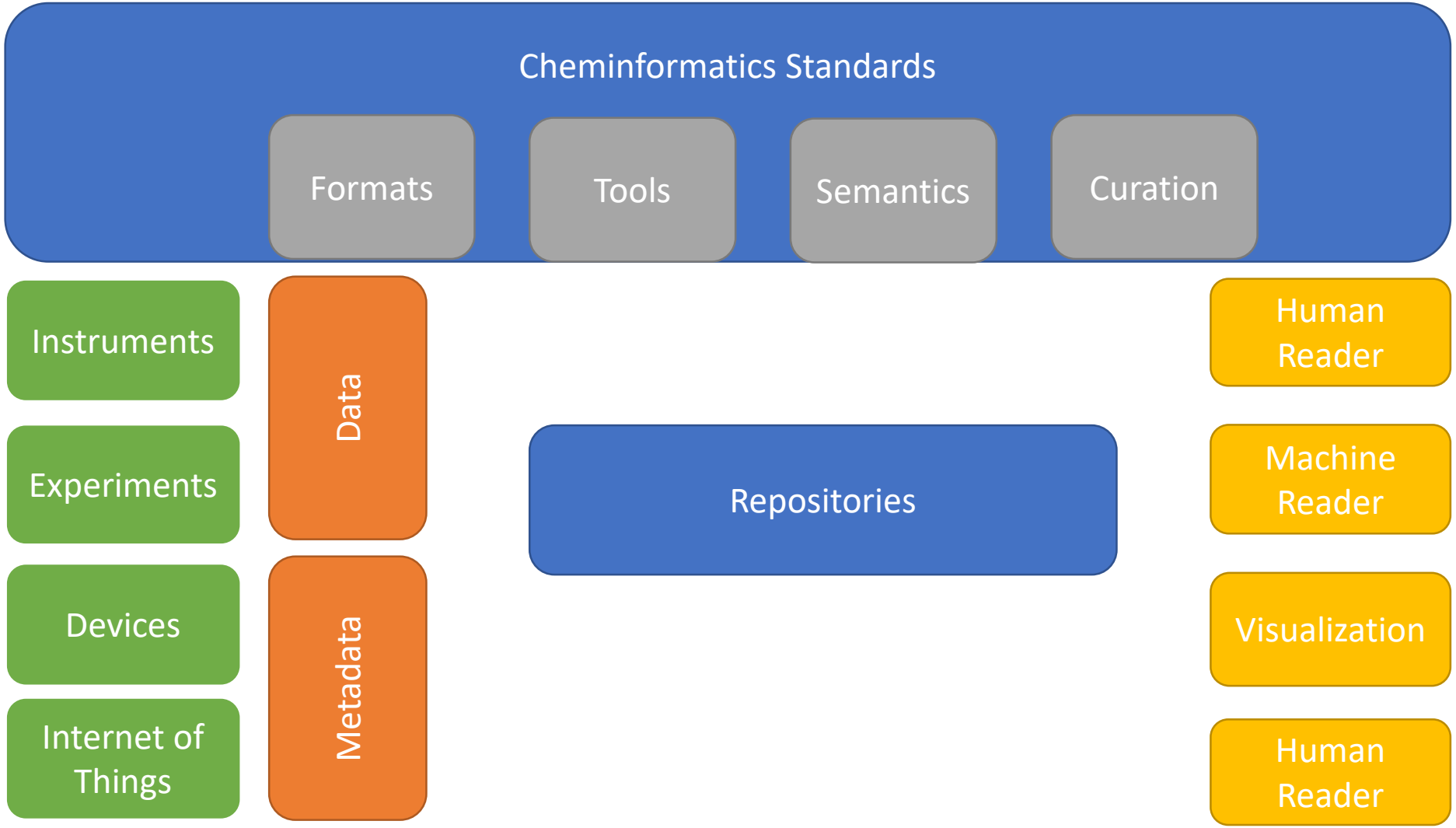
- DC VoCamp 2016: University of Maryland, November 29-December 1, 2016
  - Leah McEwen, Cornell University: RDF vocabulary for Chemical Safety & Chemical Terminology
  - Stuart Chalk, University of North Florida: the Gold Book as a step to publication as Linked Data
- Plenary 9: Barcelona, April 5-7, 2017
  - Ian Bruno, CCDC
  - David Martinsen, David Martinsen Consulting
  - Interaction with other Working and Interest Groups
    - Materials Data Infrastructure Interoperability; International Materials Resource Registries; Research data needs of the Photon and Neutron Science community; Metadata
    - Data policy standardisation and implementation
    - Disciplinary Interoperability Framework (DIF)

CRDIG Activity

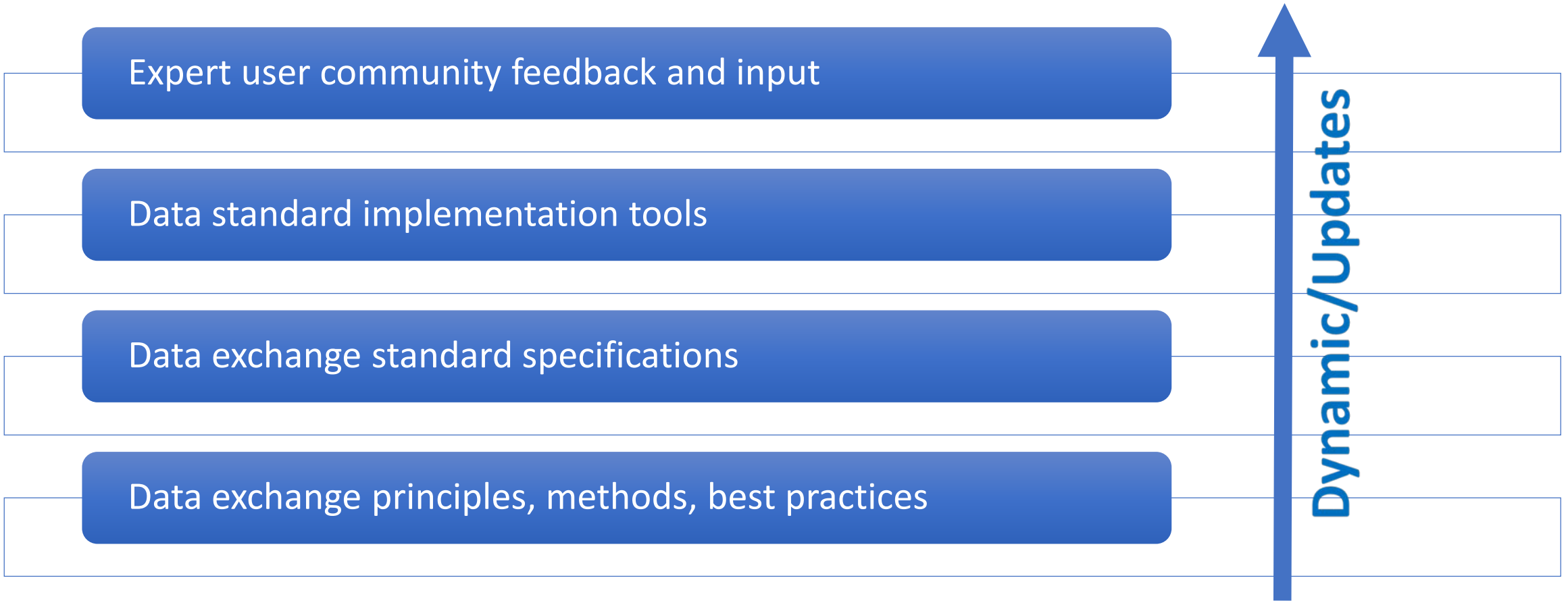
Plenary 10: Montreal, September 19-21, 2017

- Present on projects in progress
- Seek input on data challenges in these contexts
- Discuss application of best practices to ensure interoperability
- Identify interest in further collaboration on these or other projects
- Meeting agenda:
    - Update on Chemistry IG community engagement
    - Presentation on current Chemistry IG projects
        - International Chemical Identifier
        - Gold Book of Chemical Terminology
        - machine accessible data standards handbook
        - domain repository for spectra
    - Open discussion on common challenges and best practices
- Joint meeting: IG Chemistry Research Data, IG Biodiversity Data Integration, IG Domain Repositories

Scope and Vision of the Subcommittee



4 Layers of Dynamic Content 4 Dissemination and Community Engagement



Use Case: Crystallography