The Chemical Analysis Metadata Platform (ChAMP): Thoughts and Ideas on the Semantic Identification of Analytical Metrics

Stuart J. Chalk, Department of Chemistry, University of North Florida, Jacksonville, FL 32224, USA
Antony Williams and Valery Tkachenko, RSC Cheminformatics, Wake Forest, NC 27587 USA

In this work, we propose the development of standards for organization, representation, and annotation of analytical science information, based around the definition of a 'Chemical Analysis' metadata platform (ChAMP). As a consequence of this work, an approach to metadata extraction for analytical science will be delineated and applied to the RSC journal paper archive. Once optimized and documented these standards will serve as the basis for authors to provide analysis metadata upon submission of papers for publication. As part of this effort, a working group on this area will be proposed as part of the Research Data Alliance [1] (RDA) to encourage community involvement, and promote this activity as a fundamental piece of the standard architecture underneath data exchange networks for 'Big Science'.

Thoughts

Chemical analysis is required in all areas of chemical research even if it is not the main focus of the work. Verification of chemical structures, evaluation of product purity and determination of levels of chemicals before and after processing require the collection of analytical data. Thus, extraction and organization of analytical science information from research papers is a vitally important data mining activity in support of the 'Big Data' revolution in chemistry, and science in general. The RSC is presently participating in a number of projects requiring the direct management of analytical data, specifically the Chemical Database Service data repository [2]. The RSC Cheminformatics team has already developed chemical compound management capabilities with rudimentary integration to spectral data (via ChemSpider). The overarching architecture for managing chemistry related data involves the segregation of various containers into compounds, reactions, analytical data, crystals, data tables and so on. Data standards/frameworks, such as ChAMP, are essential in constructing the data repository and it is important they be ratified by the appropriate organizations where possible. Such standards are also important for the growing area of electronic laboratory notebooks such as LabTrove/ ChemTrove.

In a broad sense, papers describing new chemical analysis methods are relatively well defined in terms of scope and extent. Generally, the important facets of these publications are

- ❑ General scope
  - Analyte(s), sample type (matrix), analysis technique, application area
- ❑ Analysis method metrics
  - Qualitative
    - ➢ Spectral quality, resolution, discriminating power
  - Quantitative
    - ➢ Limit of detection, limit of quantitation, limit of linearity, sensitivity
    - ➢ Calibration type, calibration equation, correlation coefficient
- ❑ Interferences
  - Chemical, spectral
- ❑ Accuracy evaluation
  - SRM analysis, recovery study, method comparison
- ❑ Sample analysis
- ❑ Long-term reliability
- ❑ Sample preparation

For publications focused on analytical method development identification of this information should be well identified. However, extraction of this information from papers where chemical analysis is used, but is not the focus of the work, will be more difficult.

Extracting this information from existing articles will involve processing text, tables, and images

- ❑ Text
  - Descriptive information, equations, spectral information (e.g. NMR peaks)
- ❑ Tables
  - Interference study, sample analysis, recovery study data
  - Instrument parameters
  - Performance metrics – analysis rate, resolution (chromatographic)
- ❑ Images
  - Spectra, chromatograms, sample preparation workflows

Extracting all this information will require processing files multiple times with different software, however it will help clarify the important metadata that is needed to accurately describe the chemical analysis process and subsequently give clear guidance to authors of new publications of data they should provide with their work.

Ideas

ChAMP is envisioned as a framework for organizing metadata about chemical analysis rather than a specification. This perspective means it can be applied to any/all data formats that exist now (e.g. XML, JSON-LD), and in the future.

There are number of parts to this project. The first is identifying appropriate software for processing, text, tables, and images in research articles. The second is identification of data types for each of the metadata items which includes using/updating existing controlled vocabularies for appropriate fields. Third is the development of a general structure for the representation of the metadata, and the final part is linking the data to appropriate ontologies.

As part of the development process we plan to use existing standards/ technologies where appropriate. These include

- ❑ JCAMP-DX [3]
- ❑ Analytical Information Markup Language (AnIML) [4]
- ❑ Units Markup Language (UnitsML) [5]
- ❑ Experiment Markup Language (ExptML) [6]
- ❑ IUPAC Orange Book (Analytical terminology) [7]
- ❑ Controlled Vocabularies: MeSH [8], LCSH [9], CAS Subject Headings [10], Analytical Abstracts vocabularies
- ❑ Existing Ontologies: ChEBI [11], CMO [12], Semanticscience [13]

Where no standards exist we will model new ones after existing related standards.

Planned activities currently include the following

- ❑ Identification/creation of controlled vocabularies for matrix, instrument, metrics and keywords/terms
- ❑ Development of an ontology for analytical science
- ❑ Development of software, utilizing AMI2, for identification of analytes, matrices, etc. from research articles for subsequent conversion into ChAMP
- ❑ Data mining of the RSC Archive and Flow Analysis Database (includes non English articles) for evaluation of the accuracy of the developed software
- ❑ Metadata/RDF crosswalk from ELN 'elnItemManifest' [14] to ChAMP
- ❑ Integration of automatic generation of ChAMP markup from ChemTrove and the Eureka Research Workbench ELNs
- ❑ Evaluation of cross-integration of ChAMP and AnIML
- ❑ Formation of a chemical analysis working group under the RDA
- ❑ Evaluation of an International Chemical Analysis Identifier (InChAI)

The top level of the specification, encoded as XML, might look like this

```
<analysis xml:lang="en-GB">
   <title>{one}</title>
   <description>{one}</description>
   <focus>{one to many}</focus>
   <analyte>{one to many}</analyte>
   <matrix>{one to many}</matrix>
   <instrument>{one to many}</instrument>
   <instrumentcondition>{one to many}</instrumentcondition>
   <sampleprep>{one}</sampleprep>
   <metric>{one to many}</metric>
   <keyword>{one to many}</keyword>
   <comment>{one}</comment>
</analysis>
```

Thoughts, comments, suggestions? If you are interested in this project email schalk@unf.edu.

References

[1] RDA – http://rd-alliance.org
[2] RSC CDS – http://cds.rsc.org/
[3] JCAMP-DX – http://www.jcamp-dx.org/
[4] AnIML – http://animl.sourceforge.net/
[5] UnitsML – http://unitsml.nist.gov/
[6] ExptML – http://exptml.sourceforge.net/
[7] IUPAC Orange Book – http://iupac.org/publications/analytical_compendium
[8] MeSH – http://www.ncbi.nlm.nih.gov/mesh
[9] LCSH – http://id.loc.gov/authorities/subjects.html
[10] CAS – http://cas.org
[11] ChEBI – http://www.ebi.ac.uk/chebi/
[12] CMO – http://www.rsc.org/ontologies/CMO/
[13] Semantic Science – https://code.google.com/p/semanticscience/
[14] elnItemManifest – http://www.jcheminf.com/content/5/1/52

Presented at the Dial-a-Molecule Annual Meeting – Birmingham UK, July 10th 2014