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
Antony Williams and Valery Tkachenko, RSC Cheminformatics

schalk@unf.edu

2014 Fall ACS Meeting
Overview

- Initial Idea
- Motivation
- Why a Platform?
- Pieces of the Puzzle
- Existing Resources
- What are the Most Important Metadata?
- Minimum Information About a Chemical Analysis?
- Different Perspectives on a Chemical Analysis
- Example: Metadata Categories in XML
- Example: Metadata Categories in JSON-LD
- Planned Activities
- Conclusion
Initial Idea

* Develop a set of data standards for representation/annotation of chemical analysis information

* Are there important characteristics (metadata) about analysis methodologies that, if captured, would add value to a resource?

* Must be easy to implement

* Must be useful across multiple disciplines
RSC Data Repository

User interface tier (examples)

Paid 3rd party integrations (various platforms – SharePoint, Google, etc)

Electronic Laboratory Notebook

Analytical Laboratory application

Chemical Inventory application

User interface components tier

Compounds Widgets

Reactions Widgets

Spectra Widgets

Materials Widgets

Documents Widgets

Data access tier

Compounds API

Reactions API

Spectra API

Materials API

Documents API

Data tier

Compounds

Reactions

Spectra

Materials

Documents
Motivation

* Access to knowledge in existing literature
* Annotation of research in future publications
* Annotation of unpublished/self published (but potentially useful) work
* Annotation of data captured in ELN’s
* Data ingest into the RSC’s Data Repository

* Complements/enhances existing activities

* The haystack is so big – we need to make it easy to make the needle show up
Why a Platform?

- Develop it to be as broadly applicable as possible
- Chemical analysis is a not tangible like a spectrum
- Users have domain specific needs
- Users has a favorite/required format to store information
  - SQL Relational Database
  - Excel Spreadsheet
  - XML, YAML
  - JSON or JSON-LD
- ChAMP should define the types of metadata and general organization of the information, not the format it is stored in (this is like MIAME [1])

First Thoughts

* Covers metadata for a chemical analysis methodology not raw analytical instrument data

* Two main sections?
  * Fundamental method development
  * Method application

* How big should the platform scope be?
* What information is most important?
* How do we get community involvement/buy-in?
Pieces of the Puzzle

- Ontology of chemical analysis terms
- Taxonomy of chemical analysis metadata
- Controlled vocabularies for specific metadata items
- Definitions of required metadata (in context)
- Naming and design rules
Existing Resources

* Ontologies
  * Chemical Methods Ontology (CMO) [2]
  * SemanticScience CHEMINF Ontology [3]
  * ChEBI [4]
  * “Ontology on Property” by René Dybkær [5]
  * Ontobee (ontology search) [6]

Existing Resources

* Controlled Vocabularies/Taxonomies
  * MESH [6]
  * LCSH [7]
  * CAS Subject Headings [8]
  * IUPAC Orange Book [9]
  * IUPAC Gold Book [10]
* ... do they address how to organize the metadata?

Existing Resources

* Other
  * JCAMP-DX [11]
  * Analytical Information Markup Language (AnIML) [12]
  * Units Markup Language (UnitsML) [13]
  * NASA Quantities, Units, Dimensions and Data Types [14]
  * Electronic Laboratory Notebook Manifest (elnItemManifest) [15]

What are the Most Important Metadata?

* Depends on who you talk to...

* Platform should describe (as completely as possible) the types of metadata important in analysis...

* ... but leave the description of what’s important to the users

* Standards for different industries, with different requirements, could be developed based on the platform
Minimum Information About a Chemical Analysis?

* MIAChA (my-ache-a?)

* Can the community agree on a minimum set of metadata items needed to annotate an analysis?

* Must be for a more specific area of analysis
  * MIASA – Spectrochemical Analysis
  * MIACA – Chromatographic Analysis
  * MIAEA – Electrochemical Analysis
  * MIATA – Thermal Analysis
Users and publishers have different needs/wants when they view/present information

Defining perspectives (views) would extract out of a record only what a certain type of chemist would expect to see

Could be defined broadly or narrowly

This could include aggregation and/or calculation of new metrics derived from the basic metadata
Example: Metadata Categories in XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<chemicalAnalysis id="http://example.com/analysis_007">
  <description>
    <title/> <focus/> <citation/> <doi/> <analysisType/> <applicationArea/>
  </description>
  <analytes>
    <analyte id="http://www.chemspider.com/..." type="sci:CHEMINFO_000000">
      <inchikey/> <name/>
    </analyte>
  </analytes>
  <matrices>
    <matrix id="http://champ.org/ont/champ:m0000001" type="champ:MAT_000001">
      <name/> <state/> <form/>
    </matrix>
  </matrices>
  <instruments>
    <instrument id="http://champ.org/ont/champ:i0000001" type="champ:INS_000001">
      <name/> <description/> <settings/>
    </instrument>
  </instruments>
  <validation>
    <srmAnalysis/> <recoveryStudy/> <methodComparison/> <interferences/>
  </validation>
  <samplePrep>
    <collection/> <stabilization/> <storage/> <workup/>
  </samplePrep>
</chemicalAnalysis>
```
Example: Metadata Categories in JSON-LD

```json
{
  "@context": "http://champ.org/chemicalanalysis.jsonld",
  "@id": "http://example.com/analysis_007",
  "description": {
    "title": ..., "focus": ..., "citation": ..., "doi": ...
    "analysisType": ..., "applicationArea": ...
  }
  "analytes": [{"@id": ..., "@type": ..., "inchikey": ..., "name": ... }, ... ],
  "matrices": [{"@id": ..., "@type": ..., "name": ..., "state": ..., "form": ... }, ... ]
  "instruments": [{"@id": ..., "@type": ..., "name": ..., "description": ..., "settings": ... }, ... ],
  "metrics": { "detection limit": ..., "linear dynamic range": ... },
  "validation": { "reference material": ..., "recovery study": ..., "method comparison": ..., "interferences": ... },
  "samplePrep": { "collection": ..., "stabilization": ..., "storage": ..., "workup": ... }
}
```
Immediate Plans

* Get the word out
* Put up a website to provide focal point for project
* Get on social media and promote, encourage participation
* Survey the community
* Do an analysis of existing literature for metadata
* Using resources develop an initial alpha (first pass) version of the platform
* Provide mechanism for crowdsourced feedback
* Publish examples of the use in different scenarios
Longer Term Plan

* Version 1 of platform
* Controlled vocabularies
* Example documents
* Example applications
Conclusion

- This approach to metadata identification will provide value to existing resources
- It will enhance basic searching
- It will allow semantic searching
- It will provide efficient annotation of large amounts of curated data that is not from traditional publishing
Conclusion

* It also fits well with the mission of the Research Data Alliance (RDA) [16]
  
  * **RDA Vision:** Researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.
  
  * **RDA Mission:** The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data.

Questions?

* [schalk@unf.edu](mailto:schalk@unf.edu)
* Phone: 904-620-5311
* Skype: stuartchalk
* LinkedIn/Slidehare: [https://www.linkedin.com/in/stuchalk](https://www.linkedin.com/in/stuchalk)
* ORCID: [http://orcid.org/0000-0002-0703-7776](http://orcid.org/0000-0002-0703-7776)