The Chemical Analysis Metadata Platform (ChAMP)

(http://champ-project.org)

Stuart J. Chalk, Department of Chemistry, University of North Florida

Antony Williams and Valery Tkachenko, RSC Cheminformatics

schalk@unf.edu

NLM March 2015

Overview

- Initial Idea
- Motivation
- Why a Platform?
- Pieces of the Puzzle
- Existing Resources
- What are the Most Important Metadata?
- Ontology Development
- Example Application
- Future Developments
- Conclusion

Initial Idea

- Develop a set of metadata items for representation/ annotation of chemical analysis information

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

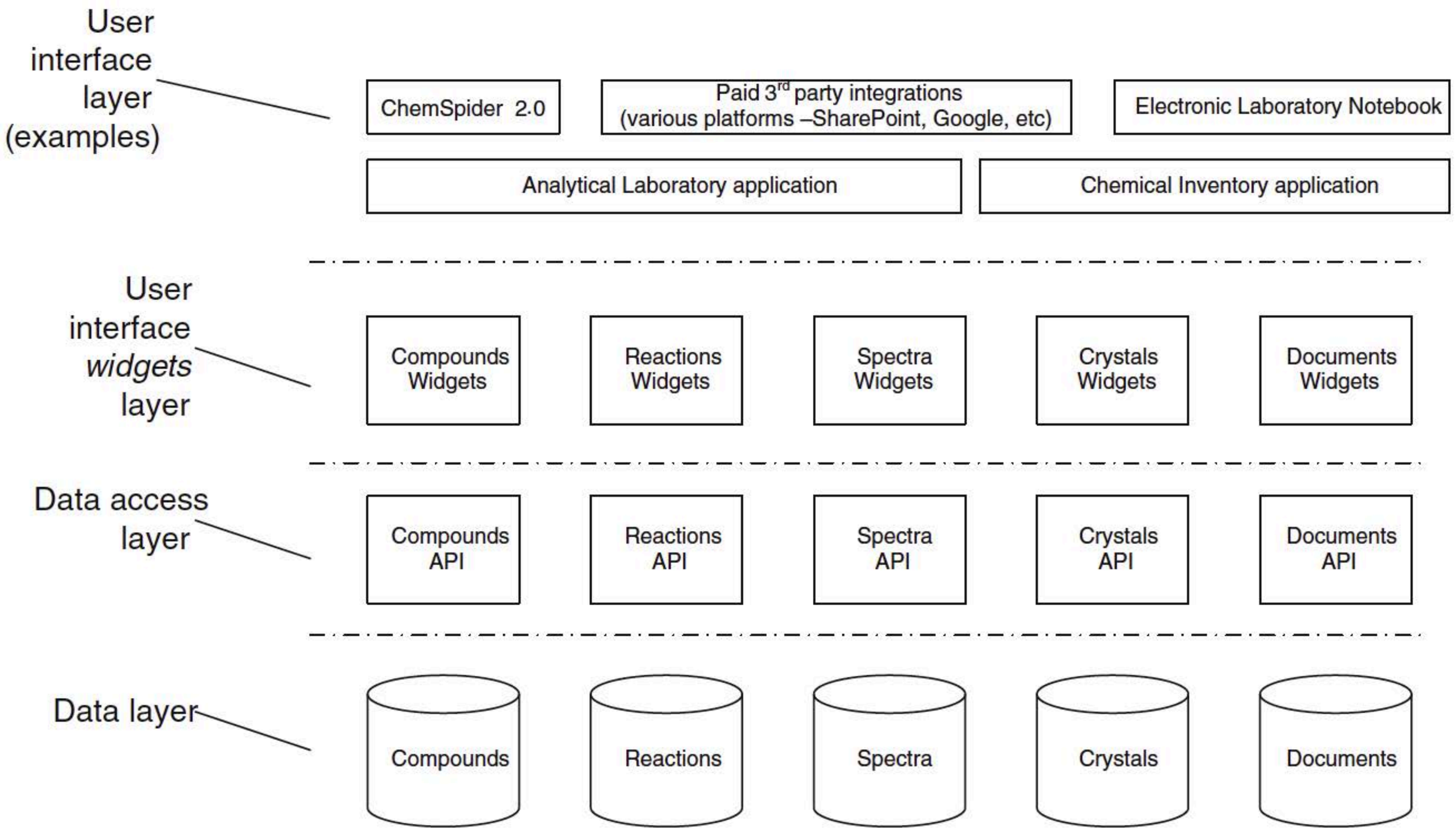
- Must be easy to implement

- Must be useful across multiple disciplines

Motivation

- How to facilitate aggregation/searching of CA information?
  - Knowledge in existing literature
  - Annotation of research in future publications
  - Annotation of unpublished/self published work
  - Annotation of data captured in ELN's
- Need tool to annotate data in digital repositories
  - Provide users with uniform (but flexible) mechanism to categorize data they contribute
  - Help researchers articulate data management plans in grants
- Complement/extend existing activities

RSC Data Repository



Motivation

- Look at the posts for analytical method help on Linked-In
  - *'I need an ICP-MS application note about direct determination of sulfur and phosphate in microwave digested plant material and soil without using external oxygen as a reaction gas.' (ICP-OES and ICP-MS)*
  - *'I want to validate a method of detecting As in glass vials with the aid of atomic absorption and air-acetylene flame'. (Analytical Method Validation)*
  - *'Does anyone know another method for determining total iron and copper in water other than calorimeter and wet chemistry?' (Analytical Chemistry)*
  - *'Anyone with knowledge in electrochemical detection of Homovanillic acid in urine samples?' (Analytical Chemistry)*

Why a Platform (Toolkit)?

- Develop it to be as broadly applicable as possible
- Chemical analysis is a not tangible like a spectrum
- Users have domain specific needs/goals
- Users has a favorite/required format to store information
  - SQL Relational Database, No-SQL, Excel Spreadsheet
  - XML, YAML, JSON or JSON-LD
- Allows use in different ways – facilitates usage
  - Build a new data standard using ChAMP
  - Annotate an existing data standard
- 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])

[1] http://www.mged.org/Workgroups/MIAME/miame.html

First Thoughts

- Covers metadata for a chemical analysis methodology but not raw analytical instrument data
- Use existing technology/standards where-ever possible
- Nothing is required – some things highly recommended
- Can use all of specification, some parts, or only one piece
- Useful for both method development and application
- Platform scope should be as wide as possible

- What information is most important?
- How do we get community involvement/buy-in?

Pieces of the Puzzle

- Description of important CA metadata
- Ontology of chemical analysis terms
  - Broad terms initially
  - Development of technique specific terms/concepts later*
- 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]
  - Chemical Entities of Biological Interest (ChEBI) [4]
  - Basic Formal Ontology [5]

[2] http://www.rsc.org/ontologies/CMO/
[3] https://code.google.com/p/semanticscience/
[4] http://www.ebi.ac.uk/chebi/
[5] http://ifomis.uni-saarland.de/bfo/

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?

[6] http://www.ncbi.nlm.nih.gov/mesh
[7] http://id.loc.gov/authorities/subjects.html
[8] http://cas.org
[9] http://iupac.org/publications/analytical_compendium
[10] http://goldbook.iupac.org/

Existing Resources

- Existing Standards
  - 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]

[11] JCAMP-DX – http://www.jcamp-dx.org/
[12] AnIML – http://animl.sourceforge.net/
[13] UnitsML – http://unitsml.nist.gov/
[14] QUDT– http://qudt.org/
[15] elnItemManifest –http://www.jcheminf.com/content/5/1/52

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

Categories of Metadata

- Description
- Infrastructure
- SamplePrep
- Analyte(s)
- Sample(s)
- Instrument(s)
- Quality
- Material(s)*
- Concept(s)*

Although the metadata is organized under these main areas implementers are free use only what they need and organize the metadata they need in any way.

Reminder: ChAMP is focused on metadata about a chemical analysis, not about the instrument data that is generated when doing a chemical analysis (although they are of course related).

The ‘Description’ Category

- **title**: the descriptive title of the method (string)
- **creator**: who is the primary author responsible for this method (string) (ORCID or name)
- **description**: textual description of the method (string)
- **analytical focus**: what is the main reason for development of the method? (string) (e.g. improvement of the detection limit)
- **application area**: broad area where the method will be most useful/used (string/enum) (e.g. 'pharmaceutical', 'environmental', 'petrochemical',...)
- **analysis type**: what is the type of analysis done in the method (string/enum) (e.g. either 'quantitative', 'qualitative', 'property’)
- **analysis format**: what is the format of the analysis (string/enum) (e.g. 'wet chemical', 'instrumental', 'sensor', 'remote')
- **analysis usage**: in what context is the method to be used – general or specific (string/enum) (e.g. 'clinical trial', 'QC', 'QA', 'general')
- **analysis locale**: the environment that the method has been developed for (string/enum) (e.g. 'laboratory', 'field', 'industrial plant', 'atmosphere’)
- **citation**: literature citation (string)

The ‘Infrastructure’ Category

- **contact:** a specific individual that can be contacted about the analysis (string)
- **person:** an individual that has participated in part in the development/production/publication of the chemical analysis (string)
- **organization:** a company/institution/organization that was part of the development/production/publication of the chemical analysis (string)
- **funding agency:** a public or private group that was a source of funding relative to the chemical analysis (string)
- **role:** the part that a contact plays in the development/production/publication of the chemical analysis (string/enum)
- **address:** physical address identifying the location of a contact (string)
- **phone:** telephone number for communicating with a contact (string)
- **email:** electronic mailing address for communicating with a contact (string)
- **location:** a place where one or more activities was performed in the development/production/publication (e.g. building/lab) of the chemical analysis (string)

The ‘SamplePrep’ Category

- **sample name:** the name given to a sample (string)
- **subsample id:** the unique identifier(s) of a subsample(s) (string)
- **subsample amount:** the mass/volume of a subsample(s) processed for analysis (float/decimal)
- **subsample unit:** the unit of the quantity of a subsample amount (string)
- **procedure:** a textual description of the procedural steps used to convert the raw sample into a sample ready for analysis (string) -- OR -- **step(s):** each procedural step separately recorded with some indication of the order the steps were taken (string) (e.g. 1-10, A thru M)
- **storage conditions:** how/where a sample is stored after laboratory processing, prior to analysis
- **chain of custody:** whether a chain-of-custody was maintained and/or the chain of custody record
- **interferences:** information about interference(s) that can be an issue in the sample preparation
- **safety:** any safety issues relative the sample preparation (string)
- **waste:** information about waste generated from the preparation procedure (string)
- **keywords:** any important terms that characterize the sample preparation process (string)
- **reference:** a formatted citation to a published version of the procedure used (string)

The 'Analyte' Category

- **substance**: a discrete chemical species identified by its InChI key/string (string)
- **substance class**: named group of chemical substances identified as a specific class by structure, use, size, or action determined in chemical analysis procedures (string) (e.g. PCB's, amino acids, PAH's, pharmaceuticals, heavy metals, enzymes, etc.)
- **functional group**: chemical test for an organic functional group (string)
- **biological property**: a property specific to a biological process (string) (e.g. biological activity, QSAR)
- **chemical property**: any property to due with a chemical reaction (string) (e.g. heat of combustion, enthalpy of formation, toxicity, rate of reaction, etc.)
- **physical property**: measurement of a bulk (material) property, or characteristic substance property (string) (e.g. solubility, RI, electrical conductivity, etc.)
- **analyzed form**: a descriptive term to indicate the state of the analyte as it was measured (oxidation state should be indicated in the InChI) (string) (e.g. dissolved, labile, total, volatile, extractable, free, residual, etc.)

The 'Sample' Category (1/2)

- **identifier:** the unique identifier of the sample (string)
- **amount:** the mass or volume of the sample received or collected (float/decimal)
- **amount unit:** the unit of the quantity of the sample amount (string/enum/vocab)
- **aggregation:** if the sample was obtained by collecting it at multiple locations and combining then it should be described here (string)
- **matrix:** description of the type of sample material (from a controlled vocabulary) (string)
- **physical state:** the phase of the sample (e.g. solid, liquid, gas, slurry, etc.) (enum)
- **homogeneity:** the homogeneity of the sample at collection (e.g. homogeneous, heterogeneous, emulsion) (enum)
- **field stabilization:** a description of the any processes used to stabilize the sample in the field
- **field additives:** list of substances added to the sample to stabilize the concentration of the analyte(s) to be determined (string)
- **storage container:** container that the collected is stored/placed in for transport/storage (include material and container type)
- **storage conditions:** how/where the sample is stored after collection, prior to lab processing and/or analysis

The ‘Sample’ Category (2/2)

- **sampling event:** was the sample collected as part of a specific trip/exploration/voyage? (string)
- **sampling location:** a description of or GPS coordinates for where the sample was obtained (string)
- **sampling depth:** the depth below sea level the sample was collected (string)
- **sampling depth unit:** the unit for the sampling depth (string/enum/vocab)
- **sampling altitude:** the altitude above sea level the sample was collected (string)
- **sampling altitude unit:** the unit for the sampling altitude (string/enum/vocab)
- **sampling conditions:** what were the environmental conditions (weather) where the sample was collected (string)
- **sampling protocol:** how the sample was collected (string)
- **sampling equipment:** the apparatus used to collect the sample (string)

The ‘Technique’ Category

- **instrument:** the general type of instrument being used to the analysis (vocabulary)
- **apparatus:** non-instrumental equipment used to do an analysis (string) (e.g. 50 mL burette, sintered glass crucible, etc.)
- **manufacturer:** the name of the manufacturer of the instrument being used (string)
- **model number:** the manufacturers model number used to identify the instrument (string)
- **serial number:** the serial number of the instrument (string)
- **software name:** name of the software used to run the instrument (string)
- **software version:** version of the software used to run the instrument (string)
- **operating system:** the operating system used to run the instrument software (string)
- **accessories:** a list of any accessories installed onto the main instrument (string) (e.g. autosampler, fraction collector, etc.)
- **configuration:** a textual description of the (physical) configuration of the instrument, used to highlight any unique/interesting aspects of the system (string)
- **settings:** A textual list of the values used for important instrumental parameters (string)

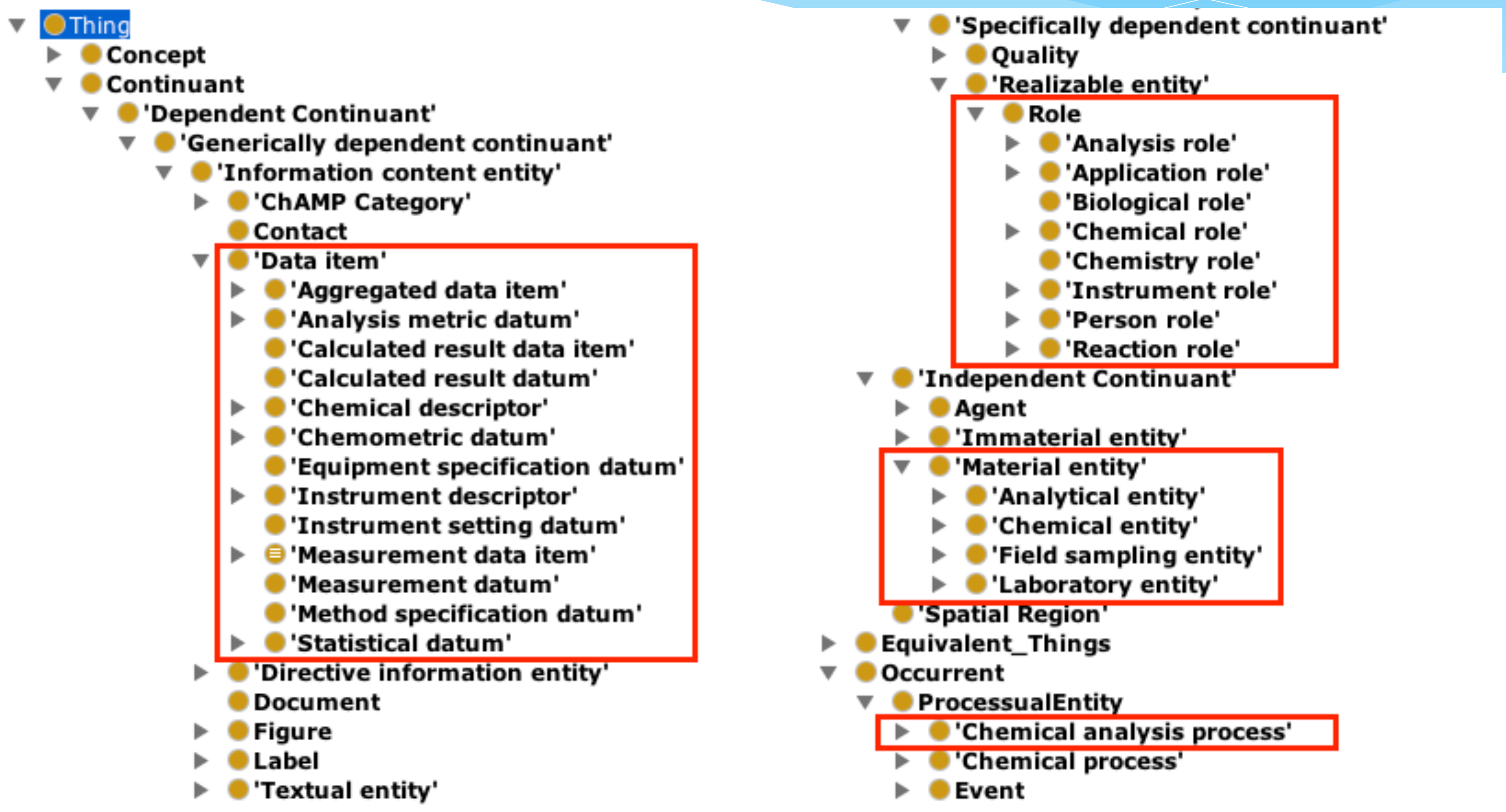
The ‘Quality’ Category

- Metrics
  - Coefficient of Determination (R^2)
  - Confidence Interval
  - Detection Limit
  - Limit of Quantitation
- Chemometrics
  - F-test
  - Paired t-test
  - One-way ANOVA
  - Non-parametric Test
- Validation
  - Quality Control (QC)
  - Statistical Process Control
  - SRM/CRM Analysis
  - Sample Spike Recovery
- Example Data
  - Chromatogram
  - Peak Table
  - Spectrum
  - Calibration Curve

Chemical Analysis Ontology

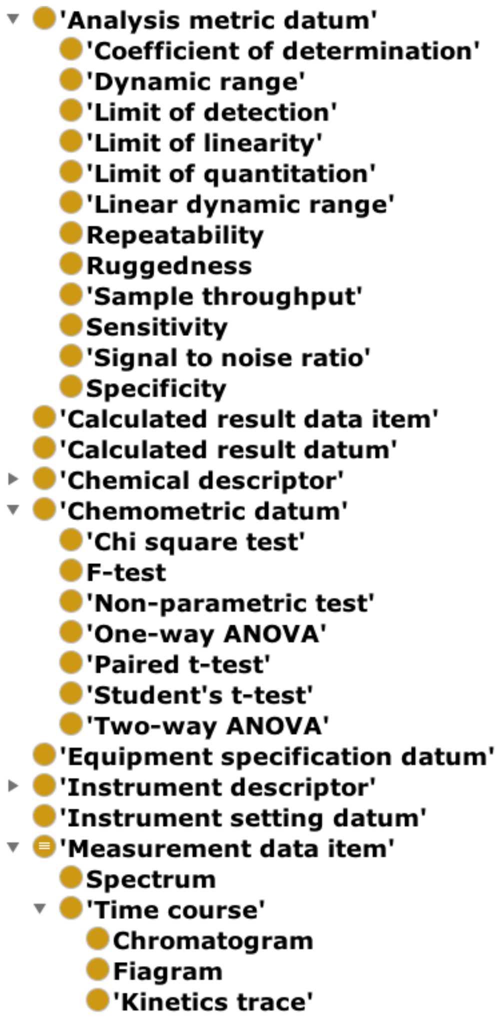
- An ontology to represent the concepts in the discipline of chemical analysis AND the metadata and data structures important to the area
- Borrows heavily from
  - Chemical Methods Ontology
  - Chemical Information Ontology
  - Chemical Entities of Biological Interest Ontology
  - Basic Formal Ontology
  - Unit of Measure Ontology

Chemical Analysis Ontology

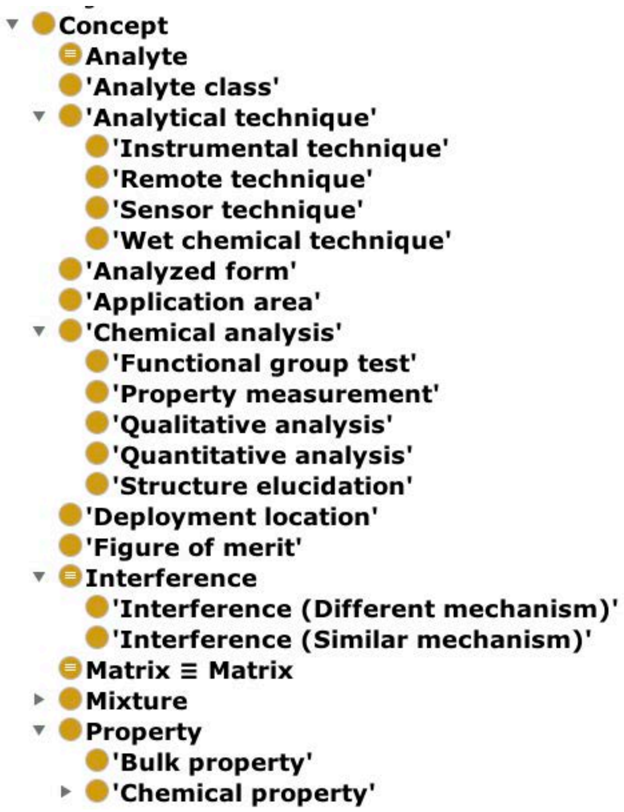


- Thing
  - Concept
  - Continuant
    - 'Dependent Continuant'
      - 'Generically dependent continuant'
        - 'Information content entity'
          - 'ChAMP Category'
          - Contact
          - 'Data item'
            - 'Aggregated data item'
            - 'Analysis metric datum'
            - 'Calculated result data item'
            - 'Calculated result datum'
            - 'Chemical descriptor'
            - 'Chemometric datum'
            - 'Equipment specification datum'
            - 'Instrument descriptor'
            - 'Instrument setting datum'
            - 'Measurement data item'
            - 'Measurement datum'
            - 'Method specification datum'
            - 'Statistical datum'
          - 'Directive information entity'
          - Document
          - Figure
          - Label
          - 'Textual entity'
      - 'Specifically dependent continuant'
        - Quality
        - 'Realizable entity'
          - Role
            - 'Analysis role'
            - 'Application role'
            - 'Biological role'
            - 'Chemical role'
            - 'Chemistry role'
            - 'Instrument role'
            - 'Person role'
            - 'Reaction role'
    - 'Independent Continuant'
      - Agent
      - 'Immaterial entity'
      - 'Material entity'
        - 'Analytical entity'
        - 'Chemical entity'
        - 'Field sampling entity'
        - 'Laboratory entity'
      - 'Spatial Region'
  - Equivalent_Things
  - Occurrent
    - ProcessualEntity
      - 'Chemical analysis process'
      - 'Chemical process'
      - Event

Chemical Analysis Ontology

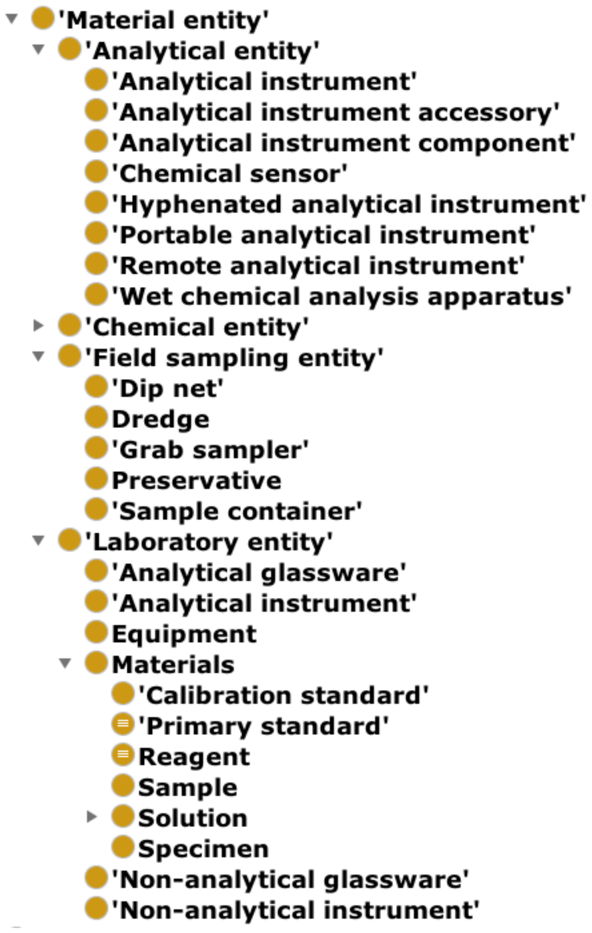


- 'Analysis metric datum'
  - 'Coefficient of determination'
  - 'Dynamic range'
  - 'Limit of detection'
  - 'Limit of linearity'
  - 'Limit of quantitation'
  - 'Linear dynamic range'
  - Repeatability
  - Ruggedness
  - 'Sample throughput'
  - Sensitivity
  - 'Signal to noise ratio'
  - Specificity
- 'Calculated result data item'
- 'Calculated result datum'
- 'Chemical descriptor'
- 'Chemometric datum'
  - 'Chi square test'
  - F-test
  - 'Non-parametric test'
  - 'One-way ANOVA'
  - 'Paired t-test'
  - 'Student's t-test'
  - 'Two-way ANOVA'
- 'Equipment specification datum'
- 'Instrument descriptor'
- 'Instrument setting datum'
- 'Measurement data item'
  - Spectrum
  - 'Time course'
    - Chromatogram
    - Fiagram
    - 'Kinetics trace'

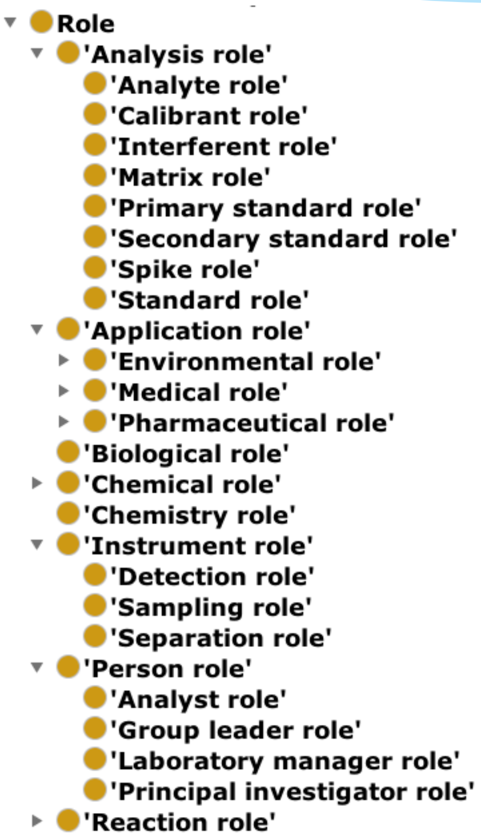


- Concept
  - Analyte
  - 'Analyte class'
  - 'Analytical technique'
    - 'Instrumental technique'
    - 'Remote technique'
    - 'Sensor technique'
    - 'Wet chemical technique'
  - 'Analyzed form'
  - 'Application area'
  - 'Chemical analysis'
    - 'Functional group test'
    - 'Property measurement'
    - 'Qualitative analysis'
    - 'Quantitative analysis'
    - 'Structure elucidation'
  - 'Deployment location'
  - 'Figure of merit'
  - Interference
    - 'Interference (Different mechanism)'
    - 'Interference (Similar mechanism)'
  - Matrix ≡ Matrix
  - Mixture
  - Property
    - 'Bulk property'
    - 'Chemical property'

Chemical Analysis Ontology



- 'Material entity'
  - 'Analytical entity'
    - 'Analytical instrument'
    - 'Analytical instrument accessory'
    - 'Analytical instrument component'
    - 'Chemical sensor'
    - 'Hyphenated analytical instrument'
    - 'Portable analytical instrument'
    - 'Remote analytical instrument'
    - 'Wet chemical analysis apparatus'
  - 'Chemical entity'
  - 'Field sampling entity'
    - 'Dip net'
    - Dredge
    - 'Grab sampler'
    - Preservative
    - 'Sample container'
  - 'Laboratory entity'
    - 'Analytical glassware'
    - 'Analytical instrument'
    - Equipment
    - Materials
      - 'Calibration standard'
      - 'Primary standard'
      - Reagent
      - Sample
      - Solution
      - Specimen
    - 'Non-analytical glassware'
    - 'Non-analytical instrument'



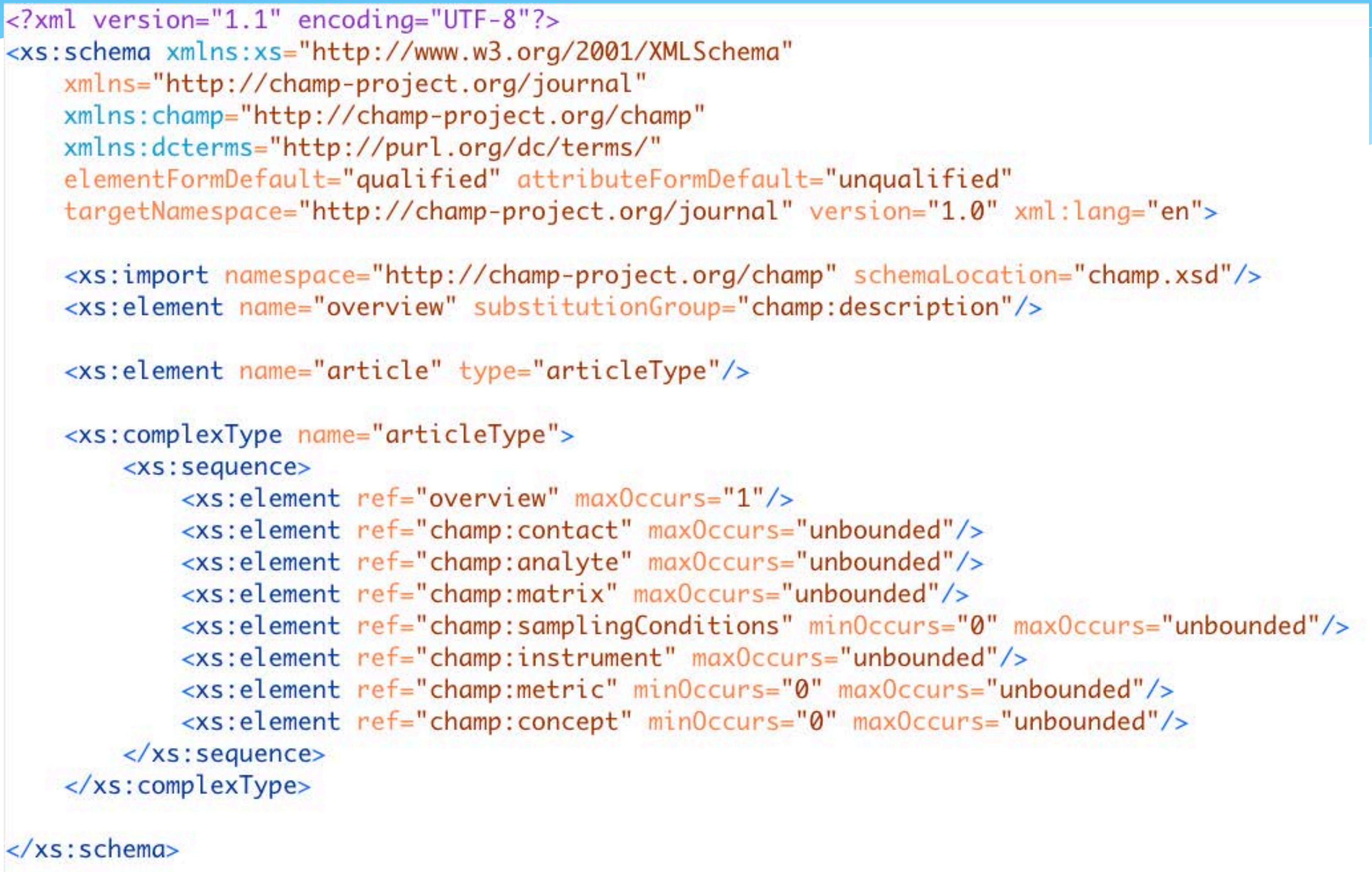
- Role
  - 'Analysis role'
    - 'Analyte role'
    - 'Calibrant role'
    - 'Interferent role'
    - 'Matrix role'
    - 'Primary standard role'
    - 'Secondary standard role'
    - 'Spike role'
    - 'Standard role'
  - 'Application role'
    - 'Environmental role'
    - 'Medical role'
    - 'Pharmaceutical role'
  - 'Biological role'
  - 'Chemical role'
  - 'Chemistry role'
  - 'Instrument role'
    - 'Detection role'
    - 'Sampling role'
    - 'Separation role'
  - 'Person role'
    - 'Analyst role'
    - 'Group leader role'
    - 'Laboratory manager role'
    - 'Principal investigator role'
  - 'Reaction role'

Example Application

- Summary information for a journal article
- Implementing ChAMP in XML

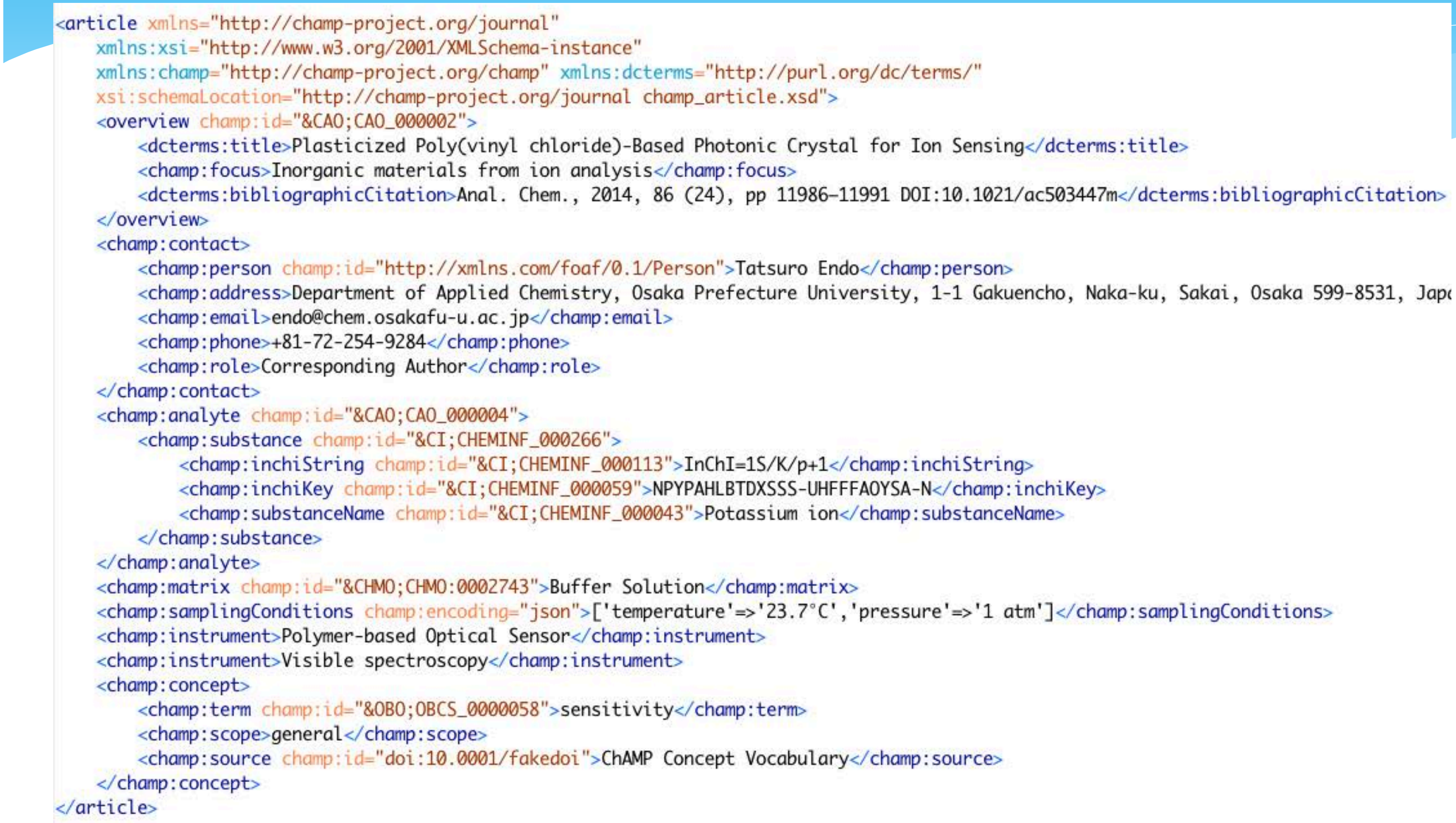
- ChAMP XML Schema
- Journal Article Metadata Specification Schema
- Instance file (XML file for one journal article)

Journal Article Metadata Schema



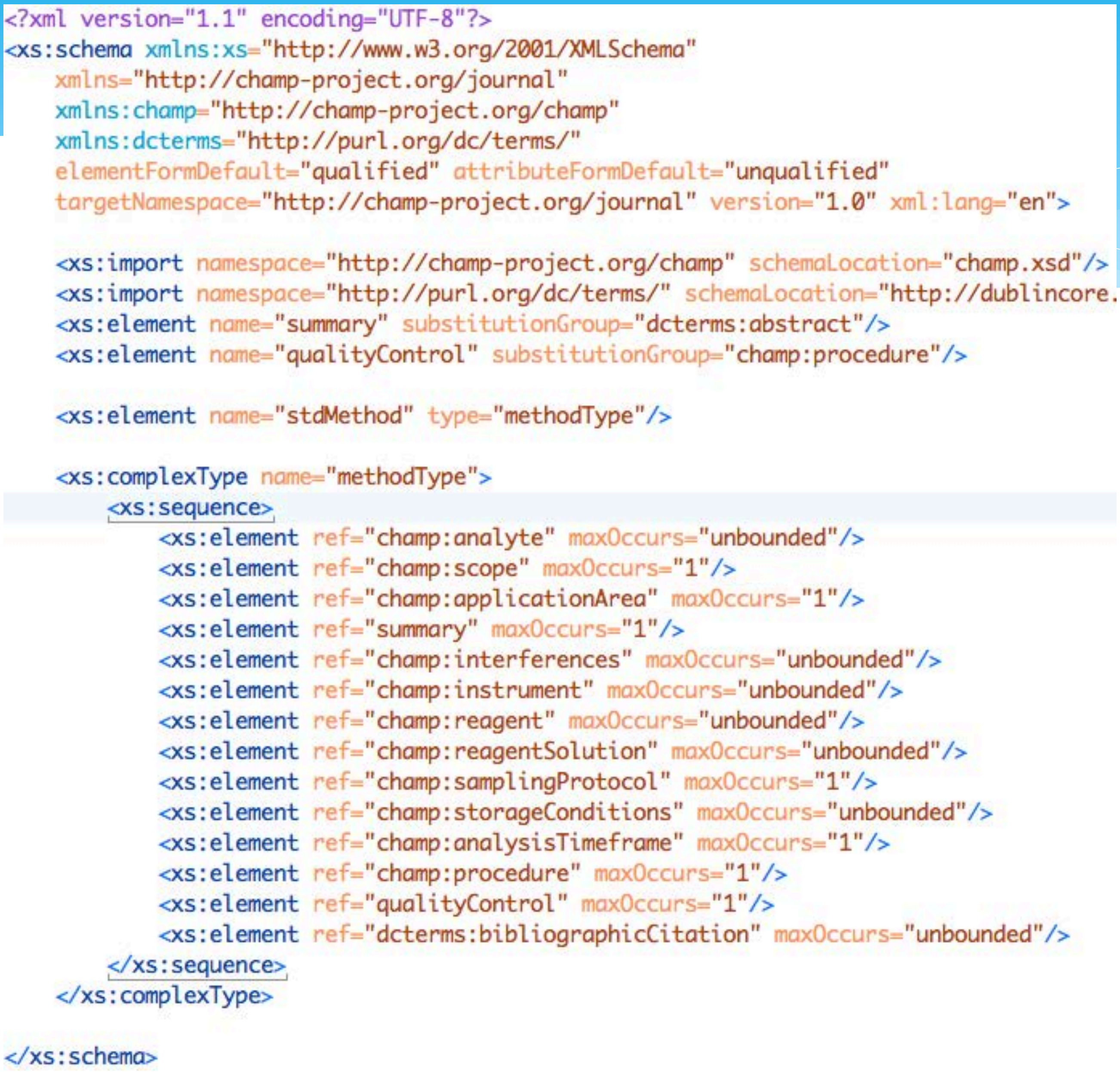
```
<?xml version="1.1" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns="http://champ-project.org/journal"
    xmlns:champ="http://champ-project.org/champ"
    xmlns:dcterms="http://purl.org/dc/terms/"
    elementFormDefault="qualified" attributeFormDefault="unqualified"
    targetNamespace="http://champ-project.org/journal" version="1.0" xml:lang="en">

    <xs:import namespace="http://champ-project.org/champ" schemaLocation="champ.xsd"/>
    <xs:element name="overview" substitutionGroup="champ:description"/>

    <xs:element name="article" type="articleType"/>

    <xs:complexType name="articleType">
        <xs:sequence>
            <xs:element ref="overview" maxOccurs="1"/>
            <xs:element ref="champ:contact" maxOccurs="unbounded"/>
            <xs:element ref="champ:analyte" maxOccurs="unbounded"/>
            <xs:element ref="champ:matrix" maxOccurs="unbounded"/>
            <xs:element ref="champ:samplingConditions" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:instrument" maxOccurs="unbounded"/>
            <xs:element ref="champ:metric" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:concept" minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>

</xs:schema>
```

Journal Article Metadata



```
<article xmlns="http://champ-project.org/journal"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:champ="http://champ-project.org/champ" xmlns:dcterms="http://purl.org/dc/terms/"
    xsi:schemaLocation="http://champ-project.org/journal champ_article.xsd">
    <overview champ:id="&CAO;CAO_000002">
        <dcterms:title>Plasticized Poly(vinyl chloride)-Based Photonic Crystal for Ion Sensing</dcterms:title>
        <champ:focus>Inorganic materials from ion analysis</champ:focus>
        <dcterms:bibliographicCitation>Anal. Chem., 2014, 86 (24), pp 11986–11991 DOI:10.1021/ac503447m</dcterms:bibliographicCitation>
    </overview>
    <champ:contact>
        <champ:person champ:id="http://xmlns.com/foaf/0.1/Person">Tatsuro Endo</champ:person>
        <champ:address>Department of Applied Chemistry, Osaka Prefecture University, 1-1 Gakuencho, Naka-ku, Sakai, Osaka 599-8531, Jap
        <champ:email>endo@chem.osakafu-u.ac.jp</champ:email>
        <champ:phone>+81-72-254-9284</champ:phone>
        <champ:role>Corresponding Author</champ:role>
    </champ:contact>
    <champ:analyte champ:id="&CAO;CAO_000004">
        <champ:substance champ:id="&CI;CHEMINF_000266">
            <champ:inchiString champ:id="&CI;CHEMINF_000113">InChI=1S/K/p+1</champ:inchiString>
            <champ:inchiKey champ:id="&CI;CHEMINF_000059">NPYPAHLBTDXSSS-UHFFFAOYSA-N</champ:inchiKey>
            <champ:substanceName champ:id="&CI;CHEMINF_000043">Potassium ion</champ:substanceName>
        </champ:substance>
    </champ:analyte>
    <champ:matrix champ:id="&CHMO;CHMO:0002743">Buffer Solution</champ:matrix>
    <champ:samplingConditions champ:encoding="json">['temperature'=>'23.7°C','pressure'=>'1 atm']</champ:samplingConditions>
    <champ:instrument>Polymer-based Optical Sensor</champ:instrument>
    <champ:instrument>Visible spectroscopy</champ:instrument>
    <champ:concept>
        <champ:term champ:id="&OBO;OBCS_0000058">sensitivity</champ:term>
        <champ:scope>general</champ:scope>
        <champ:source champ:id="doi:10.0001/fakedoi">ChAMP Concept Vocabulary</champ:source>
    </champ:concept>
</article>
```

Standard Method Metadata Schema



```
<?xml version="1.1" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns="http://champ-project.org/journal"
    xmlns:champ="http://champ-project.org/champ"
    xmlns:dcterms="http://purl.org/dc/terms/"
    elementFormDefault="qualified" attributeFormDefault="unqualified"
    targetNamespace="http://champ-project.org/journal" version="1.0" xml:lang="en">

    <xs:import namespace="http://champ-project.org/champ" schemaLocation="champ.xsd"/>
    <xs:import namespace="http://purl.org/dc/terms/" schemaLocation="http://dublincore.
    <xs:element name="summary" substitutionGroup="dcterms:abstract"/>
    <xs:element name="qualityControl" substitutionGroup="champ:procedure"/>

    <xs:element name="stdMethod" type="methodType"/>

    <xs:complexType name="methodType">
        <xs:sequence>
            <xs:element ref="champ:analyte" maxOccurs="unbounded"/>
            <xs:element ref="champ:scope" maxOccurs="1"/>
            <xs:element ref="champ:applicationArea" maxOccurs="1"/>
            <xs:element ref="summary" maxOccurs="1"/>
            <xs:element ref="champ:interferences" maxOccurs="unbounded"/>
            <xs:element ref="champ:instrument" maxOccurs="unbounded"/>
            <xs:element ref="champ:reagent" maxOccurs="unbounded"/>
            <xs:element ref="champ:reagentSolution" maxOccurs="unbounded"/>
            <xs:element ref="champ:samplingProtocol" maxOccurs="1"/>
            <xs:element ref="champ:storageConditions" maxOccurs="unbounded"/>
            <xs:element ref="champ:analysisTimeframe" maxOccurs="1"/>
            <xs:element ref="champ:procedure" maxOccurs="1"/>
            <xs:element ref="qualityControl" maxOccurs="1"/>
            <xs:element ref="dcterms:bibliographicCitation" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>

</xs:schema>
```

Future Developments

- Publish version 1 of platform (with best practices)
- General Concept Vocabulary for Chemical Analysis
- Concept Vocabularies for Specific Techniques
  - Repurpose any existing vocabularies (with permission)
  - Convert/integrate IUPAC ‘terminology’ publications
- Provide example documents in different formats
- Additional example applications
  - Partner with groups in different areas

Conclusion

- The ‘platform’ approach will make it easier for scientists to
  - Develop new standards for representing chemical analysis information
  - Integrate semantic annotation into exiting standards
- It will enhance basic searching (through standardization and vocabularies)
- It will allow semantic searching
- It will provide efficient annotation of large amounts of curated data that is not from traditional publishing
- Fits with the mission of the Research Data Alliance [16]

[16] http://rd-alliance.org