A modular approach to developing interdisciplinary, interoperable standards for geochemical data based on the Semantic Sensor Network (SSN) Ontology.

Lesley Wyborn, Simon Cox, Kerstin Lehnert, Jens Klump

# Summary (100-250 words)

Interoperability across multiple scientific disciplines has generally had limited success, as most standards are specific to a single community or discipline. However, many components of scientific data are common to multiple disciplines, and if these can be identified and leveraged then valuable foundations are laid for cross-discipline interoperability. A recent joint W3C and OGC standard - the Semantic Sensor Network (SSN) ontology - specifies the semantics of sensors, observation, sampling, and actuation, and supports a modular approach to the specification of the details of these elements of a scientific information system. This separation of concerns then allows the relevant experts to define community-endorsed standards for individual disciplines, with protocols and vocabularies in each of the modular components, within a common overall framework. Where the same component is used in many disciplines (e.g., geographical location, units of measure, colour, elements of the periodic table, definition of magnetic properties) common (universal) reference standards, protocols, can be established. This modular approach will ultimately facilitate interdisciplinary science and lay the foundations for digital scientific data to be born connected to multiple disciplines. An example of this approach is shown through the subdiscipline of geochemistry.

# Abstract (600-1000 words)

Interoperability of data within a single scientific domain has been the focus of both technical data managers and scientific discipline specialists for many decades. For example, geochemistry is widely used in the field of Earth and environmental sciences, and several efforts have been undertaken to standardise of geochemical data for publication within the geoscience discipline (e.g., Deines, 2003; Staudigel et al., 2003; Potts, 2012; Goldstein et al., 2014). However, as they are designed for standardising data in publications, they do not support use in data processing and analysis, and are also limited in their capacity for interoperability with other disciplines. There has also been some development of standards to allow data from multiple geochemical databases to be brought into a coherent entity (e.g., EarthChem XML, IEDA 2018), but currently these are only local efforts and have not yet been endorsed as agreed international standards.

Interoperability across multiple scientific disciplines has generally had limited success, mainly because the standards within a discipline are too specific to the discipline that developed it. However, many components of scientific data are common to multiple disciplines and if these can be identified and leveraged, then valuable foundations are being laid for cross-discipline interoperability. Observations and measurements are the basis for all empirical science. An observation can be understood as an act designed to determine values of properties through application of some procedure at a particular time and place, the result of an observation is strictly an estimate of the true value, conditioned by procedure and circumstances. These concepts were crystallised by Cox in a conceptual model and encodings for observations and measurements (“O&M”) (Cox, 2011, 2013; Cox and Taylor, 2015; ISO/TC 211, 2011) elaborating a pattern developed originally for medical data by Fowler (1997). O&M defines a discipline-neutral vocabulary for an observation and its properties and associated concepts. O&M was developed in parallel with models and encodings for sensor data (Botts and Robin, 2013) and strongly influenced the design of the initial SSN ontology (Compton et al., 2012).

In 2017 these evolved into a standardised version of the Semantic Sensor Network (SSN) ontology, jointly issued by Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C) (Haller et al., in press). The 2017 standard improves alignment of SSN with O&M, extends the scope of SSN to cover sampling and also actuation, and also extracts the core elements into a simplified module suitable for tagging web-pages. Of particular interest for scientific applications, the model for sampling recognizes that most observations are made on extracts or subsets of the ultimate feature of interest, and that description of the sampling process, and the nature of the relationships between samples, and with the real-world feature are critical in characterizing scientific data. Where the sample is taken from a real world object (i.e., is not synthetic), the details of the sample and the feature that it is a sample of (e.g., its location) are kept separate. The properties that may be observed and/or measured are provided by a domain model for features of interest and samples of them (Figure 1).

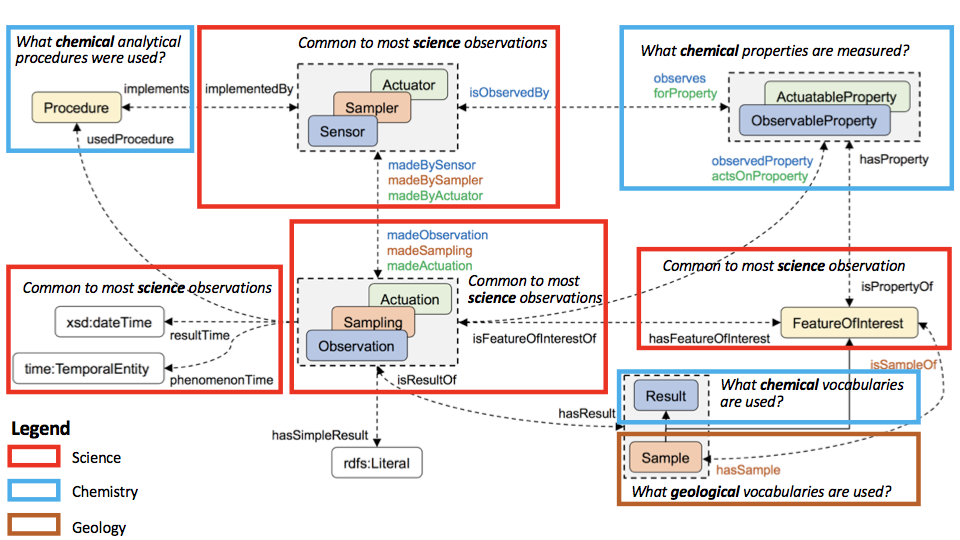


Figure 1. The core structure of the SSN Ontology showing the common patterns used (taken from Figure 1 of Haller et al. (in press)). For describing a geochemical analytical activity, modules common to most science are in red, those that have their origins in chemistry are in blue and those sourced in the geosciences are in brown.

Discipline experts provide the detailed definitions of feature types, procedures, and the sensors or samplers that implement them that are used in their community, within an overall common framework which allows these to be related using a discipline-neutral language. And where the same component is used in many disciplines (e.g., geographical location, units of measure, colour, elements of the periodic table, definition of magnetic properties) common ‘universal’ reference standards, protocols, etc. can be set up, published through standard web technologies (e.g. as linked data) and used across disciplines. This will ultimately facilitate interdisciplinary science.

However, many standards, vocabularies, protocols, etc. have already been developed in isolation within a single scientific discipline. For example, methods of recording the precise geographic location of the sample can be so variable between scientific disciplines that it is hard to collectively access, aggregate and process all data that has been collected from the same sample locality. Likewise, differing ways of providing unique identifiers to samples makes it hard to compare diverse analytical results on the same sample from different disciplines.

For the more applied Earth and environmental sciences this is a common problem, particularly in the subdisciplines of geochemistry and geophysics, which are fundamentally based on the application of the sciences of chemistry and physics respectively to geological materials. But in these applied subdisciplines, vocabularies, ontologies, protocols and procedures have often been developed independent of the core parent science disciplines.

For geochemistry, the actual geological material that is analysed for chemical properties can be described using the controlled and governed vocabularies developed by the International Union of Geological Sciences (e.g., CGI IUGS, 2016) or alternatively the Observation Data Model 2 (ODM2; Horsburgh, 2016). However, to describe measured values of the elements of the periodic table and the procedures used, geoscientists can either develop their own (e.g., EarthChem XML, Interdisciplinary Earth Data Alliance (IEDA), 2018), or else seek for equivalents developed by the chemistry community (Figure 1). A Google search for ontologies and vocabularies for the periodic table returned numerous results, and it is not possible for most geochemists to determine which are authoritative or have been endorsed.

New initiatives within International Science Unions and CODATA (e.g., CODATA, 2016) are working towards coordinating the International Science Unions to identify and endorse the more authoritative standards (including vocabularies and ontologies) within each leading scientific discipline. At the same time, initiatives within the OGC, W3C and the Research Data Alliance are providing frameworks for the development of standards that enable translation of information across disciplinary boundaries within the framework of the core SSN ontology as illustrated in Figure 1. Combined, these initiatives will start to enable interdisciplinary science and ensure that modern digital data capture is born ‘connected’ to many disciplines.

# **References**

Botts, M., and Robin, A., 2013. Sensor Model Language (SensorML). OGC, Open Geospatial Consortium, Wayland, Massachusetts. <http://www.opengeospatial.org/standards/sensorml>. Accessed on 24 June 2018.

Commission for the Management and Application of Geoscience Information of the International Union of Geological Sciences (CGI IUGS) 2016. CGI Vocabularies Register. <http://resource.geosciml.org/def/voc/>. Accessed on 24 June 2018.

CODATA, 2016. Task Group for Coordinating Data Standards amongst Scientific Unions. <http://www.codata.org/task-groups/coordinating-data-standards>. Accessed on 24 June 2018.

Compton, M., Barnaghi, P., Bermudez, L., GarcíaCastro, R., Corcho, O., Cox, S.J.D., Graybeal, J., Hauswirth, M.,Henson, C., Herzog, A., Huang, V., Janowicz, K., Kelsey, W.D., Phuoc, D.L., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K., Passant, A., Sheth, A. and Taylor, K.. 2012. The SSN ontology of the W3C semantic sensor network incubator group. Web Semantics: Science, Services and Agents World Wide Web, 17, 25–32. <https://doi.org/10.1016/j.websem.2012.05.003>. Accessed on 24 June 2018.

Cox, S.J.D., 2011. Observations and Measurements - XML Implementation. (S.J.D. Cox, Editor), OGC Implementation Standard. Wayland, Massachusetts. Open Geospatial Consortium. <http://portal.opengeospatial.org/files/41510>. Accessed on 25 June 2018.

Cox, S.J.D., 2013. Topic 20 - Geographic Information - Observations and Measurements (same as ISO 19156:2011). OGC Abstract Specification, 10–004r3, 54. <https://doi.org/10.13140/2.1.1142.3042>. Accessed on 25 June 2018.

Cox, S. J. D., and Taylor, P., 2015. OGC Observations and Measurements – JSON implementation. OGC Discussion Paper. Wayland, Massachusetts. Open Geospatial Consortium. <https://portal.opengeospatial.org/files/64910>. Accessed on 25 June 2018.

Deines, P., Goldstein, S. L., Oelkers, E. H., Rudnick, R. L., and Walter, L. M., 2003. Standards for publication of isotope ratio and chemical data in Chemical Geology. Chemical Geology, 202(1–2), 1–4. <https://doi.org/10.1016/j.chemgeo.2003.08.003>. Accessed on 24 June 2018.

Fowler, M., 1997. Analysis Patterns: Reusable Object Models, Addison-Wesley, Available at: <http://martinfowler.com/books/ap.html>. Accessed on 24 June 2018.

Goldstein, S.L., Hofmann, A.W., and Lehnert, K.A., 2014. Requirements for the Publication of Geochemical Data. Interdisciplinary Earth Data Alliance (IEDA). <http://dx.doi.org/10.1594/IEDA/100426>. Accessed 24 June 2018.

Haller, A., Janowicz, K., Cox, S.J.D., Lefrançois, M., Taylor, K., Le Phuoc, D., Lieberman, J., García-Castro, R., Atkinson, R., and Stadler, C., in press. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. Semantic Web Journal. <http://www.semantic-web-journal.net/system/files/swj1878.pdf>. Accessed on 24 June 2018.

Horsburgh, J.S., Aufdenkampe, J.S., Mayorga, E., Lehnert, K.A., Hsu, L., Song, L., Spackman Jones, A., Damiano, S.G., Tarboton, D.G., Valentine, D., Zaslavsky, I., Whitenack, T., 2016. Observations Data Model 2: a community information model for spatially discrete Earth observations. <https://doi.org/10.1016/j.envsoft.2016.01.010>. Accessed on 24 June 2018.

Interdisciplinary Earth Data Alliance (IEDA), 2018. EarthChem WFS service. <http://ecp.iedadata.org/webservices>. Accessed on 24 June 2018.

ISO/TC 211, 2011. ISO 19156:2011 Geographic Information – Observations and Measurements. (S.J.D. Cox, Editor). Geneva: International Organization for Standardization. <https://www.iso.org/standard/32574.html>. Accessed on 25 June 2018.

Potts, P. J., 2012. A Proposal for the Publication of Geochemical Data in the Scientific Literature. Geostandards and Geoanalytical Research, 36(3), 225–230. <https://doi.org/10.1111/j.1751-908X.2011.00121.x> Accessed on 24 June, 2018.

Staudigel, H., Helly, J., Koppers, A., Shaw, H., McDonough, W. F., Hofmann, A. W., Langmuir, C.H., Lehnert, K.A., Sarbas, B., Derry, L.A., and Zindler, A., 2003. Electronic data publication in geochemistry. Geochemistry, Geophysics, Geosystems - G (Super 3), 4(3), 17. <https://doi.org/10.1029/2002GC000314> Accessed on 24 June, 2018.

# Notes

## Abstract Guidelines from SciDataCon.

Deadline for abstract submission: 31 May 2018

For SciDataCon we request that proposals should be submitted for research papers, **practice papers**, lightning talks and panel contributions.

In some cases, session organisers will be inviting proposals, but please note that you may submit a proposal to any accepted session if you think your proposal is a good fit with that session topic.

**Abstracts must be submitted to an accepted session or to the General Submission session.** Select the appropriate option to submit your abstract. Abstracts submitted to specific sessions will be reviewed for their individual quality and according to their appropriateness for that session. Abstracts submitted to the General Submission option will be reviewed for individual quality and, if accepted, will be assigned to an accepted session or grouped into thematic sessions.

Accepted abstracts will be available from the conference website and will form a persistent collection, so please consider the following recommendations carefully:

* Please submit your proposal of a length appropriate to the intervention. For example, we recommend that the proposal for a research paper should be 800-1600 words and for a **practice paper 600-1000 words**. For lightning talks or panel interventions 300-600 words is sufficient. These are guides and please use your judgement. We strongly discourage proposals of more than 1600 words.
* **Submission must include a summary**. The summary should be brief - 100-250 words, and certainly no more than 250 words.

## 

## Session Description

#### **Session Title: Data interoperability in chemistry, biology, and crystallography: Enabling multidisciplinary solutions to societal challenges**

##### **Session Organisers: Ian Bruno, Jeremy Frey, John Helliwell, Leah McEwen**

#### **Session Description:**

Interoperability of data – what are the scientific success stories? What challenges had to be faced and how were they overcome? What is necessary to ensure that data from one discipline can be appropriately evaluated and applied by researchers in other disciplines? What are the current opportunities for meaningfully making data more interoperable and how can the scientific unions and cross-sector initiatives help?

This session, proposed jointly by the International Union of Crystallography (IUCr) and the International Union of Pure and Applied Chemistry (IUPAC), aims to explore these questions around interoperability with particular reference to the disciplines of chemistry, crystallography and biology and with UN Sustainable Development Goals around health, disease and energy storage in mind. Our aim is to foster awareness of data activities within these disciplines and identify opportunities for improving interoperability across domains. We hope to give voice to data initiatives that have their roots in industry as well as in academia and to feed into broader discussions around inter-disciplinary interoperability, including CODATA’s task group on Coordinating Data Standards amongst Scientific Unions.

We are seeking Practice Papers from anyone with an active desire to interoperate with chemical, biological and crystallographic data in order to address research questions relevant to societal challenges. We particularly welcome papers that address the following:

* Success stories where interoperability has helped advance science
* Experiences making data from one domain interoperable to another
* Case studies of data being made interoperable across many different systems
* Pain points encountered when attempting to use data across disciplines
* Role of data standards in sustaining data interoperability
* Overviews of the standards and tools available today – and those in development
* Pre-competitive industry initiatives tackling issues related to data interoperability
* Critical evaluation of data and enabling data literacy across disciplines

We encourage submissions to have a focus on use and reuse of data across disciplines and highlight scientific case studies that demonstrate the benefits and/or challenges of using data from chemistry, biology and crystallography in cross-disciplinary research. It is anticipated that presentations will lead into a panel discussion aimed at capturing current barriers to data interoperability.

Our overriding goal for the session is to engage the broad scientific community in identifying priorities for improving the interoperability of data from chemistry, crystallography and structural biology and the contributions that these disciplines can make to wider initiatives. We expect that the outcomes of this session will inform IUCr and IUPAC and other community initiatives seeking practical solutions to domain-specific challenges and feed in to CODATA, RDA and other activities addressing the area of data interoperability more generally.

Relevant Conference Themes

* The Digital Frontiers of Global Science:
  + digital frontiers, data and inter-disciplinary research challenges;
* A Global and Inclusive Data Revolution
  + data, research and Sustainable Development Goals;
* Research and Data:
  + disciplinary and interdisciplinary case studies of data issues, barriers and solutions;
* Data Science and Data Analysis:
  + interoperability standards in data and metadata
* Data Stewardship:
  + developing institutional, national and international data repositories and services;
* Policy and Practice of Data in Research:
  + data policy development and harmonization;
* Open Data, Innovation, Industry and Development:
  + collaborations between the research sector and commercial organisations

# Suggestion for paper:

Base it on Simon’s paper on [The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation](http://www.semantic-web-journal.net/system/files/swj1878.pdf)

The premise is that most of the puzzle pieces we require to build data interoperability standards for geochemistry already exist.

If we do it right, we can bring in ODM2/GeoSciML, UoM, IGSN etc and then say that the missing piece of the puzzle for the geochemists is an authoritative ontology/vocabulary for the periodic table and other fundamental chemical standards for isotopes and organics.

It might help us sell IGSN to a new audience,

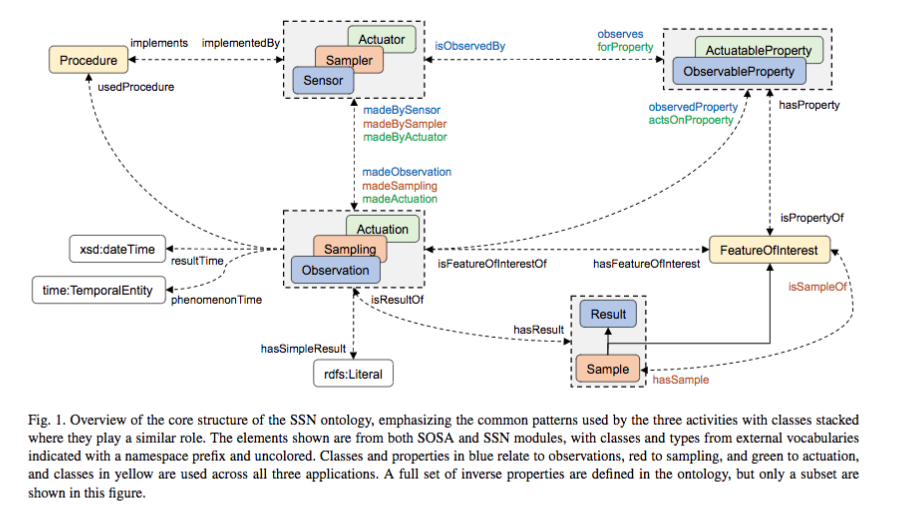
At each step we could show how identifiers are brought in.

Key points:

1. how universal the process of scientific sampling is.
2. Starting at the point of sampling O&M/SSN/SOSA we assign the unique identifier for the sample and the feature of interest it is taken from
   1. run through sampling procedures
3. We have existing standards for ‘geoscience materials’ -
   1. review those - GeoSciML, ODM
4. Through we can describe the feature that is sampled, including parent child subsampling in the laboratory.
5. Through O&M/SSN/SOSA we can describe how, by whom and when the sample is analysed
6. We can also describe the ‘sensor’ used to make the measurement and the procedures - we could bring in software citation in that we capture the software used to reduce the analytical data.
7. Using stuff from IUPAC we can describe the properties we are observing as well as results and associated units of measure
8. Finally we publish the results in a paper as Data with a DOI
9. And the paper has a DOI.
10. We could argue that this generic pattern suits analytical chemistry of any type provided someone has the equivalent of GeoSciML or ODM or suitable vocabs,

The accuracy of GPS measurements is now down to centimeters. Australia is sitting on one of the fastest moving plates on earth, and positions measured in the past are degrading. New technologies and techniques now enable more precise and accurate measurements of where our scientific observations are made, whilst multidisciplinary sensors can record suites of phenomena at the same location. However, integrating these new observations with existing legacy data from similar locations can be difficult. Too many standards, vocabularies and protocols have been developed in isolation within a single discipline, whilst methods of recording ‘where’ can be so variable between disciplines that it is hard to collectively access, aggregate and process data to create cohesive, calibrated, national datasets that can be used in transdisciplinary analysis.

Base it around Fig. 1 from <http://www.semantic-web-journal.net/system/files/swj1878.pdf>



Deines, P., Goldstein, S. L., Oelkers, E. H., Rudnick, R. L., & Walter, L. M. (2003). Standards for publication of isotope ratio and chemical data in Chemical Geology. Chemical Geology, 202(1–2), 1–4. https://doi.org/10.1016/j.chemgeo.2003.08.003

Goldstein, S. L., Hofmann, A., Lehnert, K. A., Andreae, M., Arndt, N., Baker, J., et al. (2009). Requirements for the publication of geochemical data. Integrated Earth Data Applications (IEDA), 10. Retrieved from <http://www.earthchem.org/sites/geoinfogeochem.org/files/Policy_GeochemDataPubl_v1.1_0.pdf> (page not found)

Staudigel, H., Helly, J., Koppers, A., Shaw, H., McDonough, W. F., Hofmann, A. W., et al. (2003). Electronic data publication in geochemistry. Geochemistry, Geophysics, Geosystems - G (Super 3), 4(3), 17. https://doi.org/10.1029/2002GC000314

Also: Potts, P. J. (2012). A Proposal for the Publication of Geochemical Data in the Scientific Literature. Geostandards and Geoanalytical Research, 36(3), 225–230. https://doi.org/10.1111/j.1751-908X.2011.00121.x

# 

# 