Proceedings of The National Conference
On Undergraduate Research (NCUR) 2017
University of Memphis
Memphis, Tennessee
April 6-8, 2017

Harmonization of Sensor Measurements to Support Health Research

Nicole Burnett
Departments of Biomedical Informatics and Chemical Engineering
University of Utah
421 Wakara Way
Salt Lake City, Utah 84112 USA

Faculty Advisor: Dr. Ramkiran Gouripeddi

Abstract

The Salt Lake Valley has three permanent Environmental Protection Agency (EPA) certified air quality monitoring stations that intake air samples and produce results of air quality pollutants in the proximity of the monitor station. Because these monitors only represent a small area of the 500-square-mile Salt Lake Valley, there are spatial gaps when using these air qualities monitoring data for epidemiological studies. In addition to the monitoring stations, researchers as well as Salt Lake Valley residents are recording air quality measurements from their individual sensors. Systematic utilization of these vast amounts of air quality data to support translational exposomic studies necessitates the development of a conceptual data model that harmonizes and stores air quality measurements from different sensors. A literature review using PubMed with the search criterion "Pediatric Asthma Sensor Studies' was performed by the study team. A list of metadata elements were manually extracted from literature and a first draft of sensor metadata specification was developed. Sample data from different sources was collected and used to evaluate the initial specification. Existing fields found in the data, but not present in the specification were added to the model. Air quality experts then reviewed the specification and modifications were made based on their inputs. The final product is a specification that harmonizes and stores vast amounts of air quality data from different sensors. This model is being used in data integration platforms such as OpenFurther to support the study of effects of the environment (exposome) on health and well-being.

Keywords: Air Quality, Sensor, Data Harmonization

1. Introduction

Sensors, especially personal and mobile sensors provide methods for measuring environmental exposures of individuals and populations. However, sensors use different methods and technologies for measuring different environmental species and output their measurements in different formats. In addition, sensors have differences in their performances and uncertainties associated with their measurements. It is therefore necessary to describe sensors in a generalized, harmonized and sharable manner to support their proper use in translational exposomic research.

The *Sensor Common Metadata Specification* (SCMS) provides generalized, harmonized and sharable representation of sensor metadata. It is designed to support the conduct of research utilizing personalized and environmental sensors. The scope of the specification ranges from nano-sensors to satellites. It supports measurements of physical, chemical, and biological species. In addition, the specification describes sensors that instantaneously (or with a transient storage) measure these species or those that collect physical samples for later analysis. Finally, it supports sensors that may be deployed in various environments, including personal (i.e. implanted & mobile), immediate (i.e. indoor), and general environment (i.e. external environmental protection agency monitors).

A *sensor* (a device that measures a specific species) in this document is used interchangeably with *instrument*. The reason being an instrument may be comprised of one or more measuring sensors, where a hardware device (sensor or instrument) may contain other devices in a hierarchical manner. This terminology attempts to reduce the confusion that a sensor may contain other sensors. The specification developed here describe metadata of the individual sensors and instruments composed of these individual sensors.

2. Background

Researchers at the University of Utah and around the globe are embarking on various biomedical studies to understand these associations between air quality and potential health effects[1–5]. For example, the National Institute of Health through the Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) program[6] is performing exposomic studies of pediatric asthma and other chronic diseases.

A comprehensive understanding of the effects of the modern environment, including air pollution on health requires generation of a complete picture of environmental exposures, clinical, biological and socio-behavioral factors. Such an exposome require integration of data from wearable and stationary sensors, environmental monitors, physiology, medication use and other clinical data.

Using vast amounts of data acquired from various sensor types for different research study analyses and visualizations require their semantic harmonization and integration. The SCMS provide the means for such semantic harmonization and integration and for use in a data integration platform.

Table 1: Research questions supported by these specifications.

1. *Mobile Instrument Models that can measure PM2.5.*
2. *Mobile Instrument Models that have been deployed to measure PM2.5.*
3. *Serial Number of all Instruments deployed supporting REST Data Transport Protocols and capturing output of PM2.5 (Or PM10, or Ozone).*
4. *Instrument Models that were manufactured by the AirMetrics.*
5. *Deployed Instruments owned University of Utah and currently measuring Ozone.*
6. *Organizations the collected personal exposures of PM2.5 in indoor and outdoor environments.*
7. *Calibration procedures used for MiniVol when deployed in an area with tall buildings.*
8. *Reference detection limit of MiniVol to assess quality of data quality in a study.*
9. *Number of sensors deployed by University of Utah in Salt Lake County that are less than 100 meters of I-15 in April 2016, and give the geolocation of each sensor and species measured by each sensor.*

The scope of the specification is to support a diverse set of exposomic research questions and studies (Table *1***Error! Reference source not found.**) including observational, epidemiological and prospective studies (Figure *1*). The SCMS is being developed to serve the following purposes:

1. Establish a library of instruments: Investigators can use this library to select appropriate instruments for different studies and acquire information necessary to contact the organizations owning or manufacturing these with instruments.
2. Describe and document deployments of sensors: Store a sensor’s environmental and deployment attributes that are useful when using the measurements for analysis.
3. Assess quality of data collected by different instruments within its deployment environments: Use descriptive metadata to compare sensors and check if measurements are as expected.
4. Support harmonization and integration of data collected from various sensors.
5. Provide a guide for structuring and storing sensor output data. For example, some sensors might output a latitude and longitude value while others might exclude a latitude and longitude measurement. The labeling of content of the output can vary as well. For example, a sensor could represent the latitude field by “lat” where another sensor outputting its latitude could call it “latitude.”

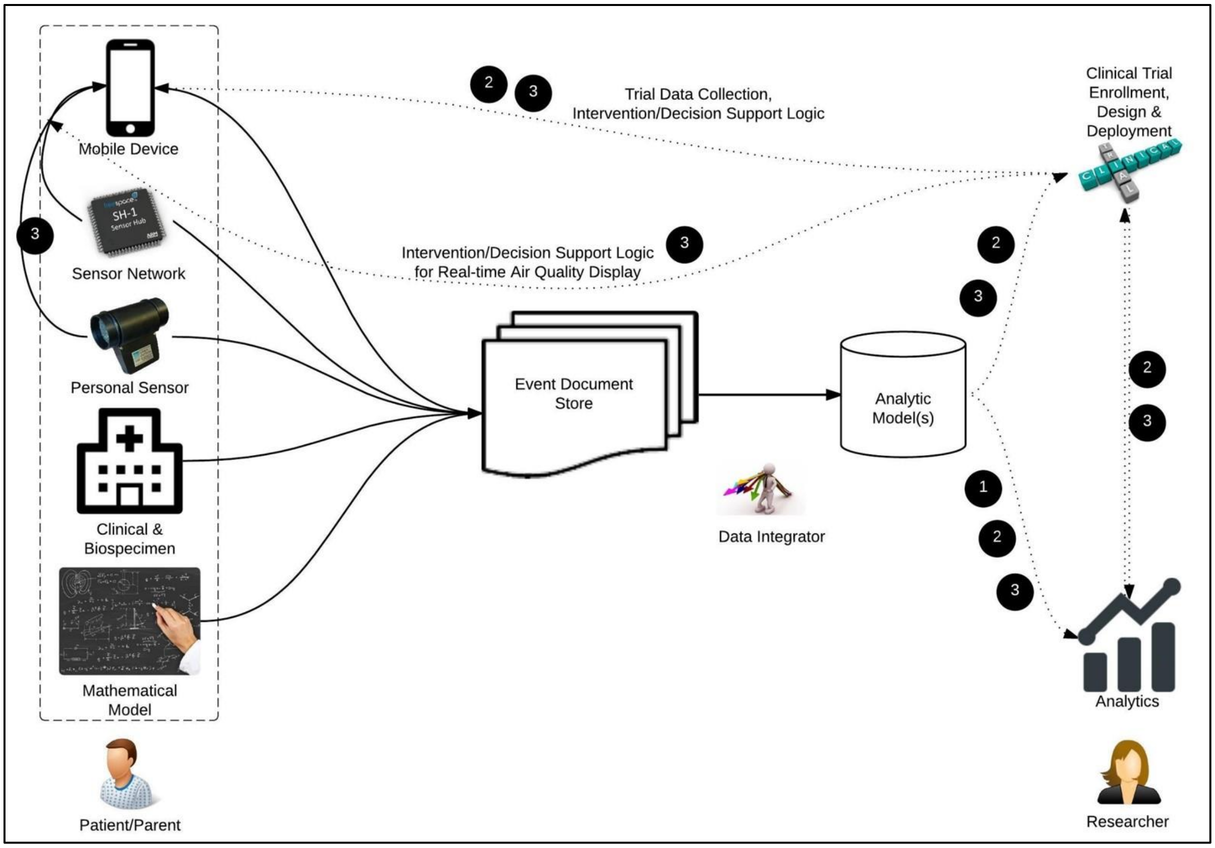

Figure 1: Diverse use-cases supported by the specifications.

3. Methods

The following four-step approach was used in developing the Sensor Common Metadata Specification.

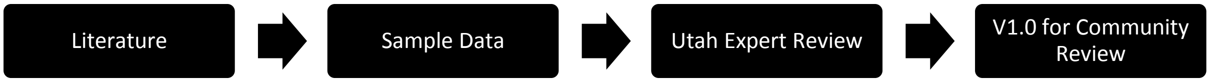

Figure 2: Four step specifications development process

3.1 Literature Review:

A literature review using PubMed with the search criterion “Pediatric Asthma Sensor Studies” was performed by the study team. This returned 231 journal articles from August 1985 - December 2015, of which 40 full texts were read. Sensor types found in this literature corpus included stationary sensors (e.g. EPA), personal sensors (mobile), and indoor sensors. A list of metadata elements were manually extracted from this literature corpus and an initial specification representing sensor metadata was established.

3.2 Preliminary Mapping With Sample Data:

To further refine the specification, the study team performed a preliminary mapping of sample data with the conceptual model. Sample data included data from:

1. Environmental Protection Agency
2. Utah Department of Air Quality
3. MesoWest[7]
4. West Valley Study
5. Asthma Triggers (Dr. Rima Habre)
6. Wood Burning (Dr. Kerry Kelly)
7. Purple Air[8]
8. Modeled Air Quality Data[9]
9. Measured Air Quality Data with Altitude[10]
10. Hierarchical Bayesian Modeled Air Quality Data[11].

Existing fields found in the data, but not present in the specification, were added to the specification.

Environmental Protection Agency
Utah Department of Air Quality (UDAQ)
MesoWest http://synopticlabs.org/
UDAQ - West Valley Study
Dr. Rima Habre – Asthma Triggers
Dr. Kerry Kelly - Wood burning
Mr. Adrian Dybwad – PurpleAir http://www.purpleair.org/
Dr. Jeffrey Yanosky: Modeled air quality data (1999 to 2007), 6 km grid
Dr. Geoff Silcox: Measured air quality data with altitude
EPA's Hierarchical Bayesian Modeled Air Quality Data

Figure 3: Sample data used for developing the specifications.

3.3 Utah Expert Review:

The specification was reviewed with air quality experts in Utah. Experts included:

11. Dr. Kerry E. Kelly, Assistant Professor, Chemical Engineering, University of Utah
12. Dr. John D. Horel, Professor, Atmospheric Sciences, University of Utah
13. Dr. Scott C. Collingwood, Research Assistant Professor, Pediatrics, University of Utah
14. Mr. Adrian Dybwad, Purple Air
15. Dr. Neal Patwari, Associate Professor, Electrical Engineering, University of Utah.

In addition, the specifications was reviewed with the PRISMS group including investigators from the PRISMS Data Model Working group, the PRISMS Data Coordinating Center at University of Southern California, and a second PRISMS Informatics Center at University of California, Los Angeles. The specification was modified further based on the air quality expert's inputs.

3.4 Community review of version 1.0:

The SCMS is available at https://github.com/uofu-ccts/prisms-sensor-model for public review and comments.

4. Results

The SCMS provides a metadata representation for harmonizing and storing vast amounts of air quality data from different sensors. This model could be used in data integration platforms such as OpenFurther to support the study of effects of the environment (exposome) on health and well-being. The SCMS consists of three domains (Figure *4*).

4.1 Instrument:

The instrument domain contains data elements that describe a physical inventory of models, along with its documentation, data transport, validation tests, measurement features, and owning and manufacturing organizations. It can be used to maintain a library of instruments using which researchers can make informed selections of instruments for different research purposes.

4.2 Deployment:

The deployment domain contains data elements that describe how a physical instrument is deployed in real world and includes characteristics such as the instruments deployment environment, setting, data transport, and calibration.

4.3 Output:

The output domain contains data elements that describe the measurement of the sensors or the physically collected samples of different species.

5. Discussion

The three domains of the SCMS may be implemented with various database technologies (e.g. relational, graph, or document databases). Here are some examples of how you may implement the three domains.

5.1 Instrument:

This is a functional description of each sensor including its ownership and manufactory specifications. This metadata can be stored as a library for supporting investigator selection of appropriate sensors and deploying for different studies.

This metadata may be best as a relational or a graph metadata repository[12,13]. Document databases are not recommended here since this library portion of the model is highly relational. For direct storage of large amounts of binary objects such as PDF documents as a part of the metadata, a relational database would be better suited since graph databases have very limited support for large binary objects. On the other hand, storing file paths to external files then a graph database may be more suitable as graphs have better support for hierarchical structures such as instruments containing a hierarchy of sensors. Graph databases also have better support for web-linked data. In other words, graphs provide better support for ternary or more degrees of relationship types, along with many-to-many cardinality as in the case of the Instrument's self-referencing relationships and the ternary degree relation between the Instrument, Organization, and Inventory entities. The Utah implementation of this uses a graph.

5.2 Deployment:

This is the metadata regarding how each physical instrument or sensor is deployed. Deployment provides information on how the sensor was configured and setup, and informs the investigator on using the Output for specific study analysis.

The deployment model is simple with one-to-many relationships surrounding the Deployment entity type. Therefore, a relational or graph database may be equally well suited.

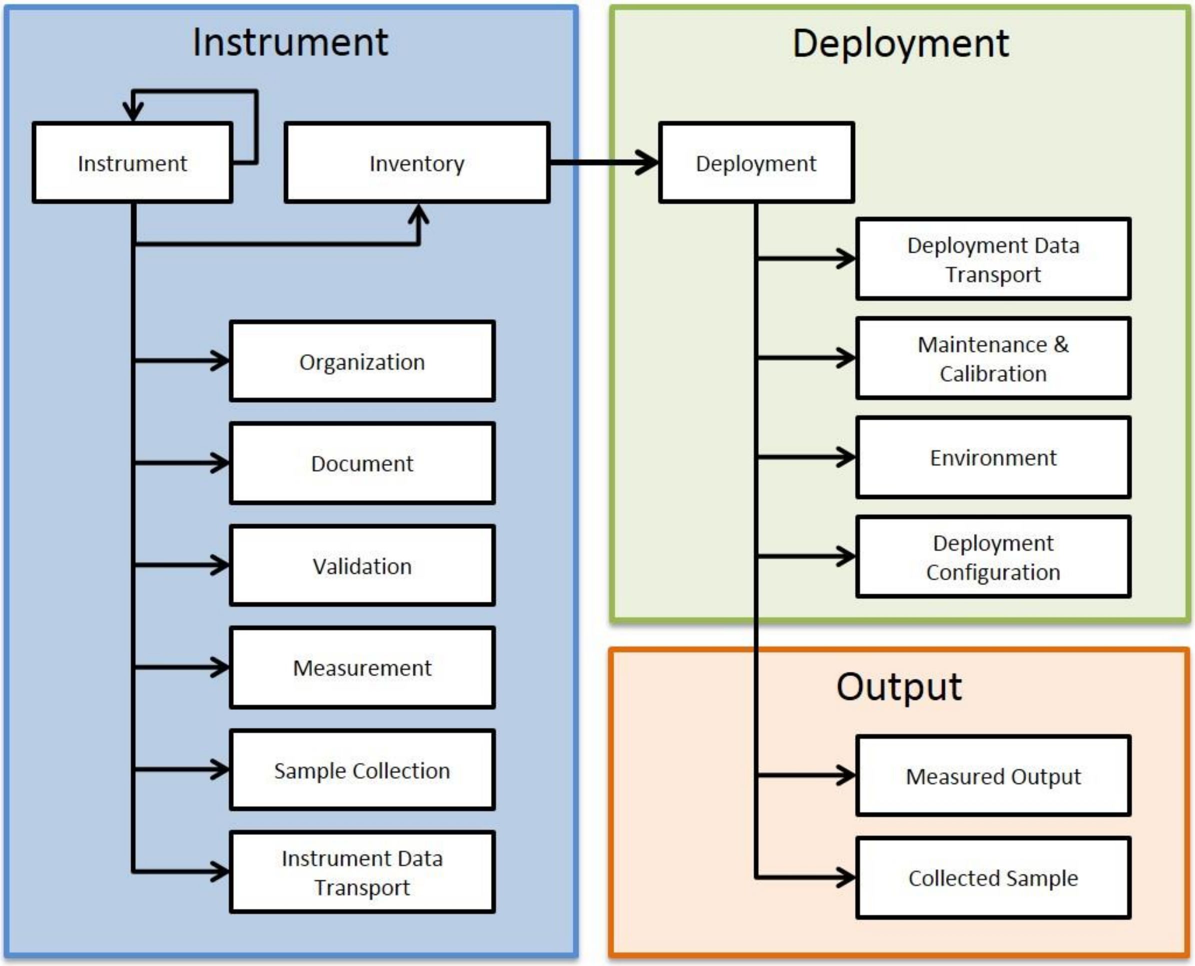

Figure 4: Domains and entities of the SCMS.

5.3 Output:

The actual output received from each deployed instrument. A document store database may be better suited for this purpose. Sensor output may be generated with a high frequency and are usually in JSON, XML, or text file formats. These features fit well with the nature of document data storage. The high throughput of data may also require the use of Big Data technologies such as a distributed file systems and a framework for parallel data processing. Document databases are generally well suited and designed for Big Data Technologies.

For example, 100 sensors collecting data every minute, over a period of one year would generate 100 x 1440 x 365 = 52,560,000 records. Although this may sound reasonable for a relational database, scaling this up to 1000 sensors collecting data every 10 seconds, for ten years would generate 1000 x (1440 x 6) x 3650 = 31,536,000,000 records. This would become much more difficult to transform, process and store in a single large relational table.

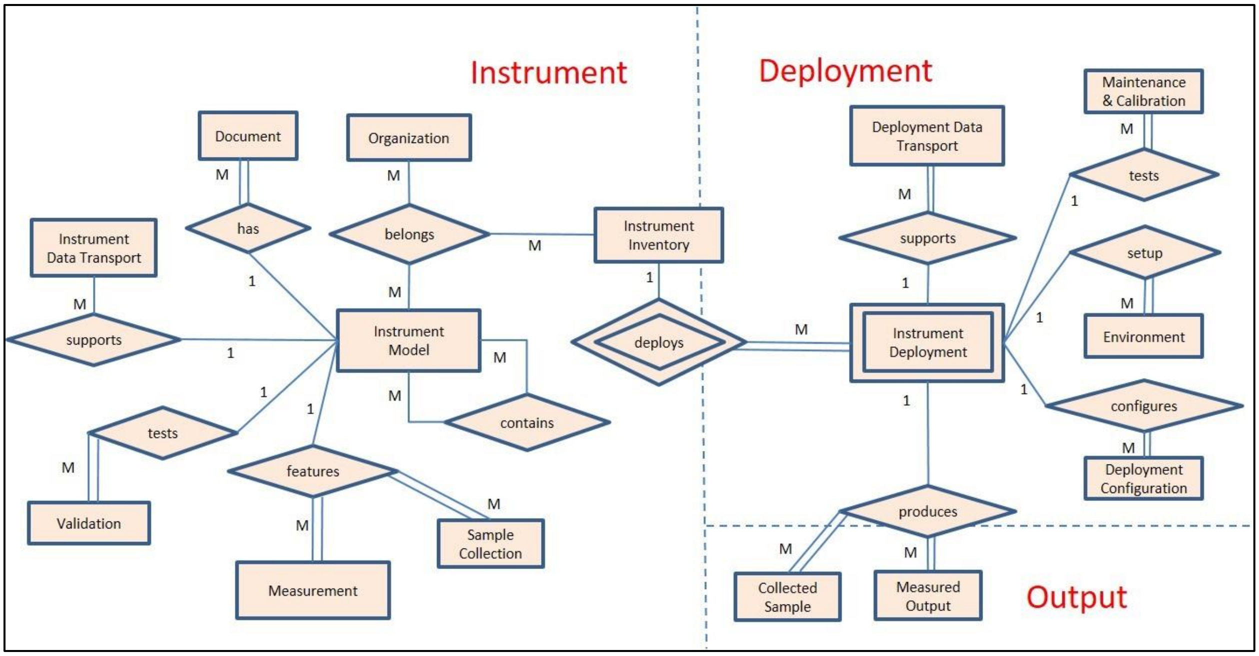

Figure 5: Entity relationship diagram of SCMS.

Figure 5 illustrates the relationships between the entities to one another within their domains and when implemented in data store. Details of SCMS including attributes within each entity is illustrated in Figure 6. The developed SCMS has been found to be generalizable to other sensors (e.g. physiological sensors) beyond air quality sensors. The SCMS is currently being implemented in the Utah Informatics platform[14] to develop a logical data model for storage and harmonization of metadata from heterogeneous sensors. The Utah PRISMS Informatics platform stores the SCMS in OpenFurther's metadata repository[12,13] to support metadata driven semantically consistent integration of all data. Using these SCMS the Utah PRISMS Informatics Platform support data from many different sources and types, and for different research use-cases (Figure *1*).

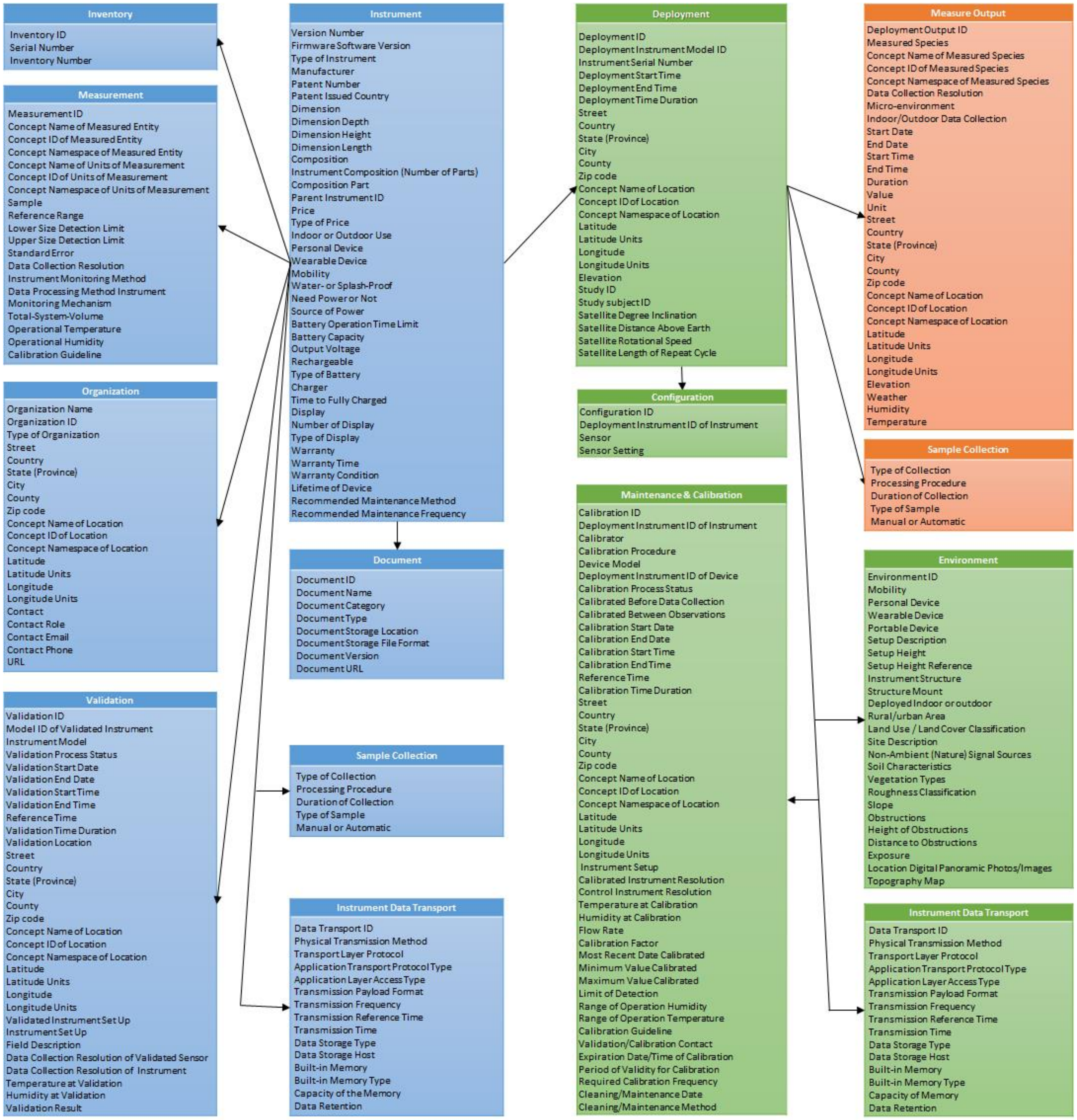

Figure 6: Entities and attributes of SCMS separated into their domains by color.

6. Acknowledgements

The author wishes to express their appreciation to Jingran Wen, PhD, Peter Mo, MS, Randy Madsen, BS, Katherine Sward, RN, PhD, FAAN, and Julio Facelli, PhD, FACMI. This study was performed with funding from a University of Utah Undergraduate Research grant. This effort is also supported by U54EB021973, National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health.

7. Reference

1. Fleisch, A. F. et al. Air Pollution Exposure and Abnormal Glucose Tolerance during Pregnancy: The Project Viva Cohort. Environ. Health Perspect. (2014). doi:10.1289/ehp.1307065
2. Weisel, C. P. Assessing exposure to air toxics relative to asthma. Environ. Health Perspect. 110, 527–537 (2002).
3. Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J. & Schwartz, J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmos. Environ. 45, 6267–6275 (2011).
4. Lioy, P. & Weisel, C. Exposure Science: Basic Principles and Applications. (Academic Press, 2014).
5. Weisel, C. P. et al. Relationships of Indoor, Outdoor, and Personal Air (RIOPA). Part I. Collection methods and descriptive analyses. Res. Rep. Health Eff. Inst. 1-107-127 (2005).
6. Pediatric Research Using Integrated Sensor Monitoring Systems | National Institute of Biomedical Imaging and Bioengineering. Available at: http://www.nibib.nih.gov/research-funding/prisms.
7. Welcome! | SynopticLabs and MesoWest. Available at: https://synopticlabs.org/. (Accessed: 14th June 2017)
8. PurpleAir.org. Available at: http://www.purpleair.org/. (Accessed: 14th June 2017)
9. Yanosky, J. D. et al. Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. Environ. Health 13, 63 (2014).
10. Silcox, G. D., Kelly, K. E., Crosman, E. T., Whiteman, C. D. & Allen, B. L. Wintertime PM2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley. Atmos. Environ. 46, 17–24 (2012).
11. McMillan, N. J., Holland, D. M., Morara, M. & Feng, J. Combining numerical model output and particulate data using Bayesian space-time modeling. Environmetrics n/a-n/a (2009). doi:10.1002/env.984
12. Gouripeddi, R. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. In American Medical Informatics Association 2013 Annual Symposium, Washington, D.C.
13. Mo, P. et al. Real-time Federated Data Translations using Metadata-driven XQuery. in (2014).
14. Gouripeddi, R. An Informatics Architecture for an Exposome. in American Medical Informatics Association Spring 2016 (2016).