# Annotating Diverse Scientific Data with HAScO

**Paulo Pinheiro**[1] **, Marcello P. Bax**[1,2] **, Henrique Santos**[1,3] **,**
**Sabbir M. Rashid**[1] **, Zhicheng Liang**[1] **, Yue Liu**[1] **,**
**Yarden Ne'eman**[1] **, James P. McCusker**[1] **, Deborah L. McGuinness**[1]

[1]Rensselaer Polytechnic Institute, Troy, NY – USA

[2]Universidade Federal de Minas Gerais, Belo Horizonte, MG – Brazil

[3]Universidade de Fortaleza, Fortaleza, CE – Brazil

{pinhep,oliveh,rashis2,liangz4,liuy30,neemay,mccusj2}@rpi.edu,

bax@ufmg.br, dlm@cs.rpi.edu

***Abstract.*** *Ontologies are being widely used across many scientific fields, most notably in roles related to acquiring, preparing, integrating and managing data resources. Data acquisition and preparation activities are often difficult to reuse since they tend to be domain dependent, as well as dependent on how data is acquired: through measurement, subject-elicitation, and/or model-generation activities. Therefore, tools developed for preparing data from one scientific activity often cannot be easily adapted to prepare data from other scientific activities. We introduce the Human-Aware Science Ontology (HAScO) that integrates a collection of well-established science-related ontologies, and aims to address issues related to data annotation for large data ecosystem, where data can come from diverse data sources including sensors, lab results, and questionnaires. The work reported in the paper is based on our experience developing HAScO, using it to annotate data collections to facilitate data exploration and analysis for numerous scientific projects, three of which will be described. Data files produced by scientific studies are processed to identify and annotate the objects (a gene, for instance) with the appropriate ontological terms. One benefit we realized (of preserving scientific data provenance) is that software platforms can support scientists in their exploration and preparation of data for analysis since the meaning of and interrelationships between the data is explicit.*



## 1. Introduction

Scientific studies are designed and executed with the goal of acquiring new knowledge about a given domain area. They are complex activities composed of more specialized activities (steps), like sampling, data acquisition and data analysis. Studies are repeatable and reusable if their data acquisition activities are described in a systematic and comprehensive way. The re-usability and repeatability of scientific studies is widely recognized as a requirement of validating and reusing previous work in data-intensive domains [Mayer et al. 2014]. Ontologies are being widely used to represent science terminology, often in settings related to integrating data resources and activities [Brodaric and Gahegan 2010]. Scientific communities have developed a significant number of ontologies to describe scientific data (e.g., OBO Foundry [Smith et al. 2007]), but far less effort has been done to describe scientific studies

themselves (SIO [Dumontier et al. 2014] is an example of such ontology), and even less to describe data acquisition activities (none to the best of our knowledge). Therefore, with existing ontologies, it can be challenging to write systematic and comprehensive descriptions of data acquisition activities. This challenge has been even more evident in large data ecosystems with a wide range of content and integration needs. In this paper, we introduce the Human-Aware Science Ontology (HAScO), designed for the specific use of encoding metadata of scientific studies. We define its scope from requirements gathered from use cases. HAScO leverages community-approved foundational ontologies as much as possible. HAScO is used today for modeling cross-domain experiments, for data annotation, semantically rich query support, and for producing data driven views for specific user groups. We claim and demonstrate that HAScO can be used to semantically annotate data from a wide variety of diverse scientific studies with the aim of supporting data integration for services such as data pooling and preparation for data analysis. One important goal of HAScO is to support the design and implementation of the Human-Aware Data Acquisition Framework (HADatAc[1]) [Pinheiro et al. 2018], an extensible platform aimed at supporting broad scientific data acquisition activities. Besides introducing HAScO, we further describe how the ontology was validated, being successfully used to represent data acquisition activities of numerous large-scale scientific projects, three of which we mention below. We use these experiences to discuss the ontology's strengths and weaknesses.

The rest of the paper is organized as follows. In Section 2, we present three data intensive projects as use cases for describing data acquisition activities. Section 3 includes identified requirements for describing scientific data acquisitions from the use cases. In Section 4, we discuss supporting science ontologies that have been integrated into the HAScO ontology. Core concepts of HAScO that fulfill the identified requirements are introduced in Section 5. In Section 6, we evaluate HAScO in the context of the use cases introduced in Section 2. Finally, in Section 7 we discuss the main properties and benefits of HAScO, including potential research impacts.

## 2. Use Cases

We describe projects that cover a variety of scientific studies and highlight a range of diverse requirements. Data is generated from a wide range of instruments, observations vary across many levels of abstraction and many types of subjects (human, environmental), and time and spatial granularity varies significantly as well. Representing this varied data in what appears to be an integrated data structure (a knowledge graph) has many challenges, a couple of the largest ones being semantic integration of the data and alignment to a number of vocabularies/ontologies.

**HAScO in Environmental Projects:** HAScO has been used in environmental projects that aim to create a deep understanding of physical and biological systems composing the overall ecological system of a lake [McGuinness et al. 2014]. The project includes content related to climate, run-off, lake circulation, lake water content, and food webs. These systems are composed of observational and experimental data acquired from sensor networks, and simulation data generated from the execution of computer models. HAScO's terminology is used to annotate project's data with descriptions on how the data were acquired from a broad range of sensing devices including sensor networks and

[1] `http://hadatac.org`

laboratory instruments.

**HAScO in Human Health Projects:** The Child Health Exposure Analysis Resource (CHEAR) aims to support investigations into exposure science and relationships to health outcomes through the collection and analysis of data from multiple studies. The ontology [McCusker et al. 2017] developed for the project thus contains terms from many domains including epidemiology, chemistry, metabolomics, toxicology, and health. HAScO is imported into the CHEAR ontology with the goal of providing uniform terminology to describe studies and their objects (e.g., subject, samples).

**HAScO in Building Sciences Projects:** HAScO has been used in an interdisciplinary project involving architects, environmental scientists, cognitive scientists, and health professionals investigating the impact of plants on humans indoors, with special interest in green walls. In this project, humans locked in an air tight room were exposed to higher concentrations of $CO_2$ with the goal of investigating a green wall's mitigations effects on the $CO_2$ concentration as well as on executive functions of the subjects. HAScO is used to support multi-criteria data alignment including time, subject, and sample.

## 3. Requirements For Describing Data Acquisition Activities

Data acquisition activities undertaken in the above projects usually involve the process of semantically correlating multiple domain variables, especially when a common element for those variables is identified. When data comes from different projects or datasets, variables may be labeled with common names, which may mean different things. For instance, $elevation$ in one dataset may mean $elevation$ "with respect to the terrain" while in another it may mean $elevation$ "with respect to the sea level." Therefore, annotating data (and data acquisition activities) can be time consuming even in simple cases when the data is from one variable in one study. However, data acquisition is often done across studies including multiple variables. Such diversity generates many requirements, for instance, representing and controlling the source or provenance of the data, appropriately identifying and encoding which instrument to use to obtain data, and appropriately representing the the quality of the data captured. These requirements should be considered when performing data acquisition activities.

**Requirements for describing sensing infrastructure metadata**: The effort of developing a vocabulary in support of environmental projects with particular attention to the distinction between measured data and model-generated data helped shape the requirements for HAScO [Pinheiro et al. 2015]. The environmental projects provided a motivating use case that generated requirements for a semantic web foundation that could support the representation and integration of observational, modeling, and simulation data. The studies had heavy measurement requirements, including being able to handle measurements of variables by more than one instrument that had different accuracies and resolutions. Additionally measurements from the same instrument would be expected to imply common accuracy and resolution, however instrument calibration and/or expertise of the operator may impact measurement results.

HAScO support for building science projects introduced other data acquisition challenges. For example, when samples (i.e.,saliva) are collected from all the subjects to measure their cortisone levels, some instruments are deployed to just some of the subjects, thus requiring the additional piece of information of the study subject of an instrument

(individual or group). Also, multiple physical properties of the room such as $CO_2$ and temperature are measured under differing conditions, e.g., air conditioner on or off, thus requiring additional context to be represented. Any data analysis of these experiments requires intensive data preparation of more than one million data points coming from over twenty instruments, operating under a range of conditions. Let's assume the measurement of the variable $room\_temperature$ indicates that, for a period of a week, two groups of human subjects will be exposed to different temperatures. In the context where room temperature is controlled by an air conditioner (AC), each event (e.g., turning on the AC, opening the door) impacts the data variable. Building Sciences projects placed additional granularity requirements on context representation.

**Requirements for describing scientific study metadata**: Processes of acquiring and organizing data are central for scientific advancement, and they may range from the short-term performance of a single scientific activity to the long-term performance of many complex scientific activities. A scientific study needs to have some specific goals, a well-defined plan, and many other components, like a leader and a funding source among others things. The description of objects related to studies is very important, like the description of subjects, samples, sampling locations, periods of sampling and measurements. In some projects, such as observing weather conditions, there is no need for air locations to be related to other air locations as long as the weather is in the right geospatial location. In some projects, the level of interrelationships between study objects is so complex that important relations need to be explicitly described. For example, one study may involve biometrics and blood samples extracted from both the mother and her child. In this case, it is essential for the data acquisition activity to describe how mothers, children, samples from mothers, and samples from children are related. The same kind of variation may be related to the importance of temporal and spatial dependencies between these objects. Prior to HAScO's support for human health projects, HAScO focused more on supporting sensor network and observational data that was not directly related to human samples. These human health projects introduced the need to record data from human (subjects) rather than focusing primarily on environmental observational and sensor data. HAScO then began to support epidemiological data from questionnaires concerning humans. Questionnaires were also considered data capture instruments. We view subject data as the collection of all data acquired from a subject. This may involve the measurement of some physiological indicators such as blood pressure or heart rate as well as the elicitation of some qualitative attributes such as smoking habits. As the samples and habits are from human subjects, subject data needs to be connected to the samples and the related questionnaire content.

Summarizing, the HAScO ontology should meet the following three basic requirements: (1) Be able to represent metadata from data acquisition activities, supporting both Study Metadata and Sensing Infrastructure Metadata. (2) Support organizing and integrating heterogeneous and diversified scientific data and allow for fusing them with domain metadata so that knowledge delivery from data to knowledge can be facilitated. (3) Enable extensibility: allow ontologists to extend it by freely grouping the concepts to support faceted visualizations that do not need to be based on the logical classifications existing within it or any other ontology imported or referenced by it.

## 4. Supporting Science Ontologies

Supporting reuse and repeatability of scientific studies is key to science in general, but even more important in data-intensive domains. As described above, many studies contain complex chains of activities, involving various data sources, computing infrastructure, software tools, or external and third-party services, making repeatability a challenging task. Another important aspect of many experiments is the social and organizational dimension - often the knowledge of how the experiments are performed is tacit and remains with the researcher, and the collaborative and distributed aspects, especially of larger experiments, contribute to this challenge [Mayer et al. 2014]. In order to fulfill the above requirements, we have carefully selected a set of foundational ontologies appropriate for use in modeling scientific data. We have aligned those ontologies and HAScO leverages them to provide a high-level common vocabulary for use across multiple studies. Using HAScO terminology, we are then able to describe values annotating them with (common) entities, attributes and units, which is key to enabling data integration between studies.

### 4.1. Foundational ontologies for data acquisition instruments and units

The "Virtual Solar-Terrestrial Ontology - Instrument model" (VSTO-I) [Fox et al. 2009] is an ontology that contains concepts that describe entities capable of collecting data (e.g., instruments, detectors and platforms) and activities related to those entities, such as deployment of an instrument on a platform. The VSTO-I ontology's development was led by the National Center for Atmospheric Research's High Altitude Observatory[2], in collaboration with McGuinness Associates, and has been used and refined by a number of organizations, including collaboration with Woods Hole Oceanographic Institute's BCO-DMO effort[3]. To represent units of measure, HAScO uses the Units Ontology (UO) [Gkoutos et al. 2012]. UO is an ontology from the OBO Foundry that provides URIs, labels, definitions, and a hierarchy for all of the International Systems of Units (SI). UO units are commonly used with data aligned with either SIO or the OBO-Foundry ontologies, making it one of the more widely-adopted unit measurement ontologies available.

### 4.2. Foundational ontologies for provenance and sensing networks

Provenance knowledge is crucial for scientific data, enabling one to understand and thus answer many important questions about the data: what is the data about? was the data measured, elicited or computed? which instrument was used to acquired the data? what was the main reason for the data to be acquired? The relationships and structure of the ontology make these question easy to answer, because a computer can deduce that, due to the relationships in the ontology and their transitivity rules, any annotations made to the parts of the cell cycle are also annotations to the cell cycle itself. HAScO's support for provenance is also tailored for science and for data quality. The use of VSTO-I, UO and PROV is strategic for HAScO since these ontologies are well-established and used by a large community of scientists to characterize the context in which data acquisition activities took place.

### 4.3. Foundational ontologies for entities and their attributes

The semantic annotation of scientific concepts in HAScO is based on the Semanticscience Integrated Ontology (SIO), which defines the types and relations currently used in HAScO

[2] https://www2.hao.ucar.edu/

[3] http://www.bco-dmo.org/

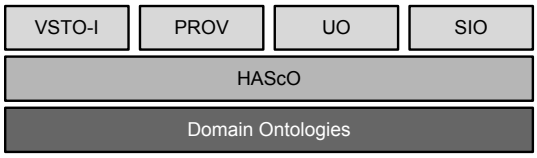


**Figure 1. HAScO and its Supporting Ontologies**

for objects, attributes and processes, and therefore provides the integrated framework from which the ontology is rooted. SIO is centered around descriptions of objects, processes, their attributes, and time. The use of SIO together with domain ontologies allows scientists to characterize the set of entities and attributes that are the objects of study in more specific scientific domains. HAScO has been designed to be extended and specialized to form domain-specific ontologies, exemplified by the use of HAScO in building the CHEAR Ontology [McCusker et al. 2017] for exposure science and health.

## 5. HAScO's Core Concepts Implementation

Figure1 depicts HAScO's comprehensive alignment of the instrument module of the VSTO-I and PROV-O, while making use of UO and SIO ontologies to further characterize domain-agnostic scientific data and related activities. We organize HAScO's Core concepts in three categories: Scientific Activities, Instruments for data acquisition and Data Organization, which is subdivided into Study Objects and Data Schema.

### 5.1. Scientific Activities

HAScO uses the view that "science is organized knowledge" and recognizes that many events may be required to acquire and organize knowledge. One of HAScO's goals is to support the identification and categorization of these events, viewed as scientific activities, along with supporting deeper representation of the events and their interdependencies in order to enable queries and integration across inter-related events. Figure 2 shows the three essential scientific activities defined in HAScO: `Study`, `DataAcquisition`, and `Deployment`. HAScO scientific activities are defined as subclasses of W3C PROV's `Activity`. That means that they are "something that occurs over a period of time and acts upon or with entities; that may include consuming, processing, transforming, modifying, relocating, using, or generating entities." The exact nature of the entities depends on the kind of scientific activity, as described below.

**Studies**: In HAScO a study can be specialized into five categories: ExperimentalStudy, FieldStudy, LaboratoryStudy, ObservationalStudy and SubjectStudy, as shown on the right side of Figure 2. Each study may be composed of several steps (StudyStep), which can be DataAcquisition, DataAnalysis and Sampling. HAScO provides a high-level classification of studies to be expanded as needed. On the top of this hierarchy of studies, studies are classified as observational when no variable in the study is controlled, or experimental when at least one variable is controlled. In observations, HAScO-annotated datasets (the entire collection of data from a study) can be represented as a single data acquisition if no control, such as instrument calibration, is taken into consideration. In experiments, having the dataset broken down into data acquisitions is an effective way of describing variable control. For example, if an experiment is measuring the effects of light on human subjects, each data acquisition may be characterized by events like turning the

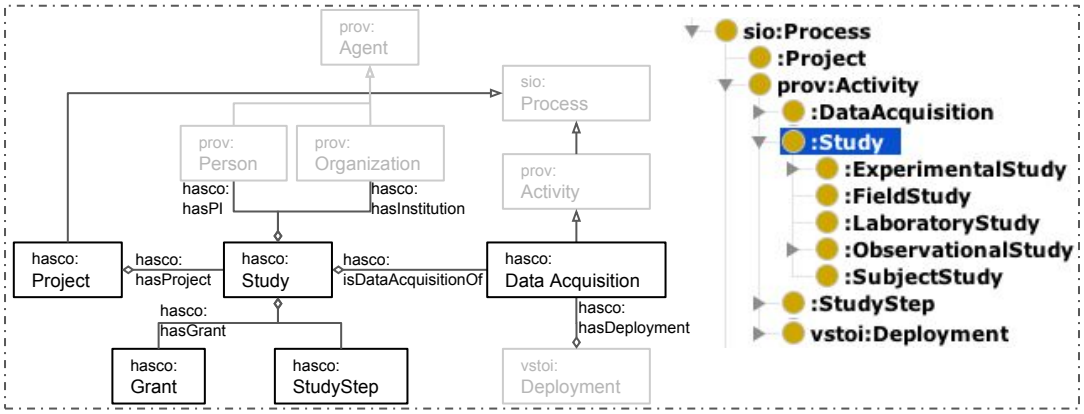


**Figure 2. HAScO's Study Representation**

lights on from off, or off from on. The subclasses of a study are not disjunctive. A study that is a FieldStudy requires instrument management data like conditions of deployment and configuration set. A study that is not a laboratory study may not have uncertainty management, i.e., computing accuracy (limit of detection), resolution, or reliability of detectors. A study involving humans is subject to IRB[4] regulations that may be manifested in terms of PROV's `Agent` involvement in the project and the questions in a questionnaire instrument. HAScO does not aim to fully define a scientific study. Instead, it aims to describe what is required for the representation, integration and analysis of study data; how such data relates to supporting the achievement of study goals, and how data that was originally acquired in support of one study may be reused in other studies.

**Data Acquisitions**: In HAScO `DataAcquisition` is both an event using an instrument to acquire data, as well as the overall collection of data values acquired by the instrument during its deployment, where all the points belonging to the collection have exactly the same quality. HAScO defines data quality as the entire configuration set and property set of the instrument and corresponding deployment that were used to acquire the data. For instance, instrument properties are key to defining data accuracy and resolution, and instrument/deployment configuration parameters are critical to data precision. In Figure 3(A), we see that a study, in terms of data, is characterized by its associated collection of data acquisitions.

As shown in Figure 3(A), each data acquisition is associated with a single deployment. Data acquisitions only exist in the context of deployments. This means that the start date/time of a data acquisition cannot occur before the start date/time of its associated deployment, and the ending date/time of a deployment is also the ending date/time of any open data acquisition associated with the deployment.

**Deployments**: A `Deployment` is the placement of an instrument in a platform so that it can be ready for acquiring data. At any given time, more than one instrument can be placed at a single platform, meaning that many deployments may occur at a single platform at any given time. A deployment must have a start time and may have a stop time. If a deployment has no stop time, it is assumed to be ongoing. A triggering event

[4]Under FDA regulations, an IRB is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects.

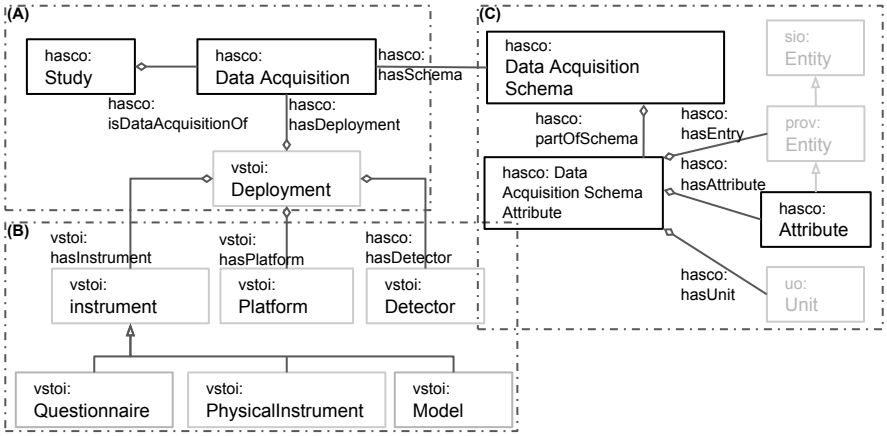


**Figure 3. (A) Scientific Activities; (B) Instruments; (C) Data Schema.**

indicates a change in the deployment configuration, which may be a change in the instrument itself. Any configuration change during an ongoing deployment means that, within the deployment, data acquired before the change should only be compared or analyzed against data acquired after the event, if there is a clear understanding and consideration of the change event in the quality of acquired data.

**Instruments (of data acquisition)**: HAScO's `Instrument` is a concept imported from the VSTO-I ontology, and is a key building block of sensor networks. Figure 3 shows a specialization of `Instrument` into `Questionnaire`, `PhysicalInstrument` and `Model`. The `PhysicalInstrument` is the concept that is currently used as sensor network's building block. The use of other non-physical instruments along with deployments and data acquisitions have shown that the HAScO generalization of questionnaires and models as instruments enable uniform characterization of data in the sense that each data point, regardless of its provenance, was acquired by an instrument deployed to a platform, and the quality of the data is defined by instrument/deployment configurations and settings. Questionnaires may be viewed as instruments for eliciting human knowledge as shown in Figure 3(B). For HAScO, simulation `Models` that are capable of generating data semantically equivalent to physical instruments are considered subclasses of `vstoi:Instrument`.

## 5.2. Scientific Data Organization in a Study

Over the course of a scientific activity (either a single data acquisition activity or multiple data activities of a study, or even of multiple studies), data is constantly acquired from attributes of study objects of interest. A significant description of the design and structure of a study is done through the modeling of objects related to the study. An internal identifier, an optional investigator-managed identifier, and relations to other study objects minimally compose the set of objects of a study. Study objects can be subjects, samples from subjects, and samples from the environment. Time events are examples of more abstract study objects.

**Study Objects and Semantic Object Collections**: In order to describe and manage study objects, they are grouped into semantic object collections (SOCs) that provide a convenient way of referring to all objects in a study that play the role of, for example, being subjects. SOCs can be used to describe potentially complicated inter-

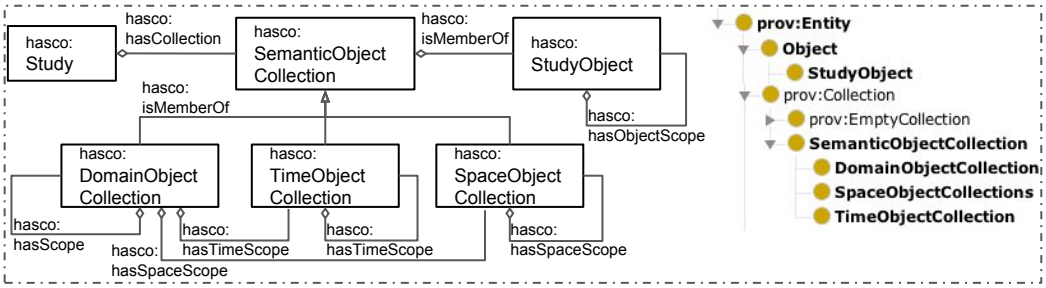


**Figure 4. Study Objects and Object Collections (Semantic Study Design)**

relationships between samples and/or subjects as well as requirements for collections. For instance, in a given study, it may be specified that two blood samples are collected from each subject. In HAScO, a `prov:Collection` is a `prov:Entity` and provides a structure (e.g. set, list, etc.) to some constituents (which are themselves Entities). The `prov:Collection` class can be used to express the provenance of the collection itself: e.g. who maintained the collection, which members it contained as it evolved, and how it was assembled. SOCs can also be used to capture spatial and temporal relations between study objects. For instance, for the blood samples collected from each subject, one may be collected during the subject's first interview, while the second sample may be collected at the subject's third interview. In this case, a Time Object Collection, with two abstract study objects ("first interview" and "third interview"), can be created and associated with the SOC for samples. As shown in Figure 4, Study objects, instances of `StudyObject`, are organized in collections (`SemanticObjectCollection`) that constitute a `Study`. HaScO specifies subclasses of `SemanticObjectCollection`: `DomainObjectCollection`, `TimeObjectCollection`, `SpaceObjectCollections`.

**Data Acquisition Schema**: Datasets conveying scientific data are frequently shared together with human-readable descriptions of their format, as a way of enabling data understanding by new users. HAScO understands that reusing a data schema for multiple datasets is a common practice and it provides a representation for schemas that can then be reused for multiple datasets, multiple data acquisitions or even multiple studies. HAScO calls this a "Data Acquisition Schema" since it is used to identify the portions of a dataset that are relevant to a study, and to specify how these portions of the dataset content should be semantically represented. Since in HAScO every scientific data point is always part of a data acquisition activity, it is assumed that a study is comprised of at least one `DataAcquisition` and that each data acquisition is described by one `DataAcquistionSchema`. Figure 3(C) shows that a data acquisition schema is comprised of a collection of `DataAcquisitionSchemaAttributes`, and that each schema attribute is associated with three classes: a subclass of `hasco:Entity`, a subclass of `sio:Attribute`, and a subclass of `uo:Unit`. With the characterization of these three classes, one can verify if any two data points are semantically related or equivalent, e.g., if they measure the same attribute of the same entity using the same unit.

## 6. Discussion & Evaluation

HAScO is an upper-level ontology that has been developed to provide a comprehensive description of data acquisition activities performed within the context of scientific stud-

| Data Acquisition Activity Metadata | | | | HAScO's Use Cases (Research Projects) | | |
|---|---|---|---|---|---|---|
| | | | | Environmental Studies | Human Health Studies | Building Sciences Studies |
| Study Metadata | Study Type | Observations | | X | | X |
| | | Experiments | | X | X | X |
| | Study Description | Identification of objects and their inter-relations | | | X | X |
| | | Data quality management at study level | | X | | X |
| | | Temporal support for study description | | | X | X |
| | | Spatial support for study description | | X | | X |
| Sensing Infrastructure Metadata | Instruments (Data Acquisition Methods) | Measurement Data | sensor networks | X | | X |
| | | | lab controlled | | X | |
| | | Elicited Data | questionnaires | | X | X |
| | | | documents as source | | | X |
| | | Model Generated Data | simulation | X | | |
| | Activities (Quality Control) | Uncertainty Provenance | lab managed | | X | |
| | | | deployment managed | X | | X |

**Table 1. Requirements for Scientific Data Representation**

ies. It is designed to leverage both ontologies describing scientific studies and ontologies describing scientific data. It has the goal of aligning the terms from leveraged ontologies, e.g., HAScO aligns PROV's Activity with SIO's Process, and aligns VSTO-I's Instrument with SIO's Device. HAScO has been under development for more than four years and a comprehensive infrastructure based on it supports the entire management of data in a number of major scientific projects, each one of them composed of tens of studies from multiple principal investigators. HAScO plays a number of roles in the process of describing scientific data acquisition activities: (1) It is designed to be extended by domain ontologies in multiple application areas - for example, we have described some of our work using it in areas including lake science, exposure science, and built environment science in addition to health science; (2) It integrates terms from high level ontologies required to describe data acquisitions in the context of scientific studies - for example, HAScO uses provenance terms from W3C PROV, the instrument and deployment concepts from VSTO-I, the hierarchy of units from OBO Foundry's Unit Ontology, and the hierarchies of entities and attributes from SIO; (3) It works as a framework to integrate ontologies describing scientific data – for example, HAScO classes have been designed to be extended with terms coming from ontologies like ChEBI [Degtyarenko et al. 2008], HP [Robinson et al. 2008], and COGAT [Poldrack et al. 2011] that are used to describe scientific data in the area of biochemistry, human phenotype, and cognitive measurements; and (4) It introduces a number of concepts not found in other ontologies - for example, it introduces the "data acquisition" term for handling data quality, and the "semantic object collection" and "study object" terms to describe study design. Currently the HAScO ontology is being used to represent metadata from various research projects that are using the data acquisition platform resulting from the HaDatAc project [Pinheiro et al. 2018]. Hadatac is a tool that merges data from many different studies, such as those that take place in the projects mentioned in Section 2. With this tool, a user can browse and compare the annotations of objects from different studies because the annotations are made with a common ontology. To demonstrate the applicability of the HAScO ontology, and as a way to validate the ontology, Table 1 illustrates the coverage of the semantic data annotation requirements of data acquisition activities from each of the major scientific projects that we introduced in this paper and that are served by HAScO, which were briefly described in Section 4.

**Support for Representing Study Metadata**: In addition to allowing the annota-

tion of different types and descriptions of scientific study, HAScO supports the identification and encoding of relationships between domain data involved in studies and represented as objects in a RDF graph. For example, HAScO is capable of encoding complex relationships among and between samples. In an environmental study, scientists need to decide where measurements and simulations are made to be able to understand and thus predict environment behavior. In a clinical study, epidemiologists and health professionals need to select cohorts of subjects using the subjects' property values. Moreover, when actual material samplings occur, if they occur directly from an environment or if they are sampled from other samples, for example, when a saliva sample or a blood sample is collected from a human subject, it is essential to understand how these objects (i.e., the samples) are related to understand the relationship between data from sample's properties.

**Support for representing Sensing Infrastructure Metadata**: In HAScO every data point is defined in the context of a data acquisition and every data acquisition with the same context has the same data quality, i.e., the same combination of contextual values defined in Figure 3(A) and (B). In this way, for instance, the quality of any two data points in a large data collection of values for a single variable can be compared through the inspection of their corresponding HAScO-annotated provenance graph. HAScO addresses the shortcomings of other science driven ontologies that are not capable of simultaneously supporting preparation of data that was acquired through measurements, data elicitation from human subjects, and data simulation with the use of a computer model.

## 7. Conclusion

The complex task of extracting knowledge from data involves the now popular use of data analysis and the often-ignored (or at least underestimated) laborious task of preparing data. In this paper, we introduced the Human-Aware Science Ontology (HAScO) that was developed and applied to major scientific projects with the immediate goal of helping teams of scientists with data preparation in support of data analysis. HAScO is domain-agnostic, and leverages a combination of well-established foundational ontologies including SIO, OBO Foundry's UO, W3C's PROV, and VSTO-I. In addition to supporting some high-level scientific concepts such as `Studies` (including subclasses' `Observations`, and `Experiments`), `Subjects`, `Samples` and others, HAScO provides a quality dimension of data based on a new generalizable concept called `Data-Acquisition`. As pointed out in [Brodaric and Gahegan 2010], an effort to encode scientific findings in a structured, knowledge-enhanced way using ontologies, can support research exploration and potentially identify novel connections, thereby increasing the overall research impact. The ontology is available[5] under MIT license. We are maintaining and evolving the ontology through its use in the HaDatAC infrastructure in support of numerous sponsored research projects. To the extent that the HaDatAc framework is used as a basis for the implementation of new research projects, HAScO will evolve accordingly, guaranteeing the necessary support for the progress of the ontology.

## Acknowledgements

This work was partially supported by the National Institute of Environmental Health Sciences (NIEHS) Award 0255-0236-4609 / 1U2CES026555-01, National Science Foundations Award DBI 1625044, the Gates Foundation through the Healthy Birth, Growth, and

[5] `http://hadatac.org/ont/hasco/`

Development knowledge integration (HBGDki), the RPI Tetherless World Constellation, and CAPES Award 88881.120772 / 2016-01.

## References

Brodaric, B. and Gahegan, M. (2010). Ontology use for semantic e-science. *Semantic Web*, 1(1, 2):149–153.

Degtyarenko, K. et al. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350.

Dumontier, M. et al. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5:14.

Fox, P., McGuinness, D. L., Cinquini, L., West, P., Garcia, J., Benedict, J. L., and Middleton, D. (2009). Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. *Computers & Geosciences*, 35(4):724–738.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The Units Ontology: a tool for integrating units of measurement in science. *Database*, 2012.

Mayer, R., Miksa, T., and Rauber, A. (2014). Ontologies for Describing the Context of Scientific Experiment Processes. In *2014 IEEE 10th International Conference on e-Science*, volume 1, pages 153–160.

McCusker, J. P., Rashid, S. M., Liang, Z., Liu, Y., Chastain, K., Pinheiro, P., Stingone, J. A., and McGuinness, D. L. (2017). Broad, Interdisciplinary Science In Tela: An Exposure and Child Health Ontology. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 349–357, New York, NY, USA. ACM.

McGuinness, D., Pinheiro, P., Patton, E., and Chastain, K. (2014). Semantic escience for ecosystem understanding and monitoring: The jefferson project case study. In *AGU Fall Meeting Abstracts*, volume 1, page 3712.

Pinheiro, P., McGuinness, D. L., and Santos, H. (2015). Human-Aware Sensor Network Ontology: Semantic Support for Empirical Data Collection. In *Proceedings of the 5th Workshop on Linked Science. Bethlehem, PA, USA*.

Pinheiro, P., Santos, H., Liang, Z., Liu, Y., Rashid, S., McGuinness, D., and Bax, M. (2018). HADatAc: A Framework for Scientific Data Integration using Ontologies. In *Proceedings of the ISWC 2018 Posters & Demonstrations Track*.

Poldrack, R. A. et al. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in neuroinformatics*, 5:17.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Smith, B. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.