Towards an extensible ontology for streaming sensor data for clinical trials

Robert Lyons
Medidata Solutions
New York, NY
rlyons@mdsol.com

Geoffrey Ross Low
Medidata Solutions
London, UK
glow@mdsol.com

Clare Bates Congdon
Medidata Solutions
New York, NY
ccongdon@mdsol.com

Melissa Ceruolo
Medidata Solutions
New York, NY
mceruolo@mdsol.com

Marissa Ballesteros
Medidata Solutions
New York, NY
mballesteros@mdsol.com

Steven Cambria
Medidata Solutions
New York, NY
scambria@mdsol.com

Paolo DePetrillo
Medidata Solutions
New York, NY
pdepetrillo@mdsol.com

ABSTRACT

The use of wearable sensors for clinical trials can lead to better data collection and a better patient experience during trials, and can further allow more patients to participate in trials by allowing more remote monitoring and fewer site visits. However, extracting maximum value from the data collected via streaming sensors presents some specific technical challenges, including processing the data in real time, and storing the sensor data in a representation that facilitates the use of biomarker algorithms that can be used and reused with different similar sensors, at different scales, and across different clinical trials. Here we present our initial work on SORBET, a Sensor Ontology for Reusable Biometric Expressions and Transformations. Our design strategy is presented, along with the initial design and examples. While this ontology has been created for the Medidata Sensor Cloud product, it is our hope that others working in this space will join us in extending and hardening this ontology, as we expand it to incorporate more sensors and more needs for clinical trials research.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; *Health care information systems*; • **Information systems** → **Resource Description Framework (RDF)**; **Web Ontology Language (OWL)**; • **Computing methodologies** → **Ontology engineering**; **Semantic networks**.

KEYWORDS

Health informatics, Ontology engineering, Semantic networks

BCB '21, August 1–4, 2021, Gainesville, FL, USA

ACM ISBN 978-1-4503-8450-6/21/08.
https://doi.org/10.1145/3459930.3469562

ACM Reference Format:
Robert Lyons, Geoffrey Ross Low, Clare Bates Congdon, Melissa Ceruolo, Marissa Ballesteros, Steven Cambria, and Paolo DePetrillo. 2021. Towards an extensible ontology for streaming sensor data for clinical trials. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3459930.3469562

1 INTRODUCTION

Wearables and other high-quality remote sensors provide detailed objective biometric data, collected directly from people in the context of their daily lives, and enable a better patient experience and greatly improved data collection in clinical trials. As sensor technology expands and more clinical trials utilize remote sensors, the scientific and commercial landscape for the wearable sensors that supply the data underlying this digital health transformation is rapidly evolving and fragmented; device providers come and go, each with their own APIs, data formats, and algorithms. This situation leads to a semantic divergence of data that confounds our ability to extract scientific results from data collected in clinical trials, and slows the progress of novel biomarker discovery. Analyses of similar biometrics done with data streams from different devices, at different times, or by different organizations are challenging to compare. With the increasing use of wearable sensors, the process of conducting clinical trials has become a big-data endeavor, and we need to develop data representations to keep pace with this growth and enable us to achieve data harmonization and to maximize the knowledge that can be extracted from this valuable data.

This paper proposes SORBET, Sensor Ontology for Reusable Biometric Expressions and Transformations, a semantic model for wearables and other sensors used in clinical trials, designed as part of the Medidata Sensor Cloud. SORBET builds on established biomedical ontologies to provide an extensible representation to facilitate the design and application of reusable and scalable biomarker algorithms for clinical trials. As we extend this ontology to cover more sensor types and more clinical areas, we seek to

share this ontology with others working in this space, and to build a community around its continued development and use.

In the following, we will present related work in Section 2, and the needs for our ontology in Section 3. Section 4 will describe the ontology that we have developed, and Section 5 will provide examples of use. Section 6 will outline plans for future developments.

2 BACKGROUND AND RELATED WORK

As argued by Lehne et al., [11], the notion of interoperable data and systems is essential for modern digital medicine, based on needs of big-data analytics, as well as other concerns. In order to pursue digital-medicine analytics to its full potential, there must be clear data structures and semantics. Miron et al. [13] look at clinical trials data in particular, and note that the use of structured data and ontologies is necessary in order to make data reusable for downstream analyses. In this paper, our motivation is to be able to conduct meaningful analyses of biometric data collected from wearable sensors, where possibly different sensors are used for the same clinical trials. Additionally, the goal is to be able to pool data from different clinical trials for downstream analyses. The SORBET ontology is designed to enable these sorts of analyses, while also being an extensible framework that can be extended to new biometric data in the future. Our efforts exist in the context of other ontologies for streaming sensor data, of course. This section will describe some of this related work. First, seemingly similar ontologies in the sensor space, and second, the established ontologies that have been incorporated into or have inspired our work.

2.1 Other ontologies for medical sensor data

Internet of Things (IoT) technologies are of course relevant to streaming sensors for clinical trials, and substantial W3C Semantic Web projects described in the next section are relevant to and incorporated into our work. However, many IoT semantic projects are not specific to clinical trials, and are not discussed here. We are aware of little work on ontologies designed for streaming sensors used in clinical trials, but the following two publications are of note.

Hennessy et al. [9] includes a lightweight health sensor ontology as part of a larger system, focused on providing the minimal amount of data needed from the sensors. This is an interesting early project in this space, but does not address our somewhat opposite needs of wanting to maximize the potential for reuse of all data captured.

El-Sappagh et al. [3] talks about mobile health technologies and semantics of sensor data, and is concerned with efficient patient monitoring via sensors, focused on integration with electronic health records to create a remote integrated monitoring and treatment system. This project bears some similarities to our goals and implementation, but it focused in particular on diabetes; thus, the resulting ontology is narrower in scope than is needed for our work.

2.2 Established frameworks and ontologies

In creating SORBET, we are building on ideas from multiple established frameworks for clinical trials and for remote sensors. We have designed SORBET to incorporate relevant ontologies and to allow incorporation of others as the ontology grows. This section describes some of the established standards that are related to or directly incorporated into our work.

The World Wide Web Consortium has multiple groups that relate to the semantic web. The Simple Knowledge Organization System (SKOS) [12] establishes a data model for sharing and linking knowledge organization systems via the web. This sets a standard for data harmonization efforts. An ontology for sensors, observations from sensors, and related procedures is the Semantic Sensor Network Ontology (SSNO) [7]. While SSNO is quite broad, within SSNO, there is the Sensor, Observation, Sample, and Actuator (SOSA) ontology [4, 10], which we have incorporated into SORBET.

The QUDT ontology [14], is used to represent quantity and unit standards, designed specifically for scientific and engineering work. QUDT implements international standards and therefore facilitates interoperability. Relatedly, the Unified Code for Units of Measure (UCUM) [15], is intended to include all units of measures being contemporarily used in international science, engineering, and business. These systems have also been incorporated into SORBET.

The Logical Observation Identifiers Names and Codes (LOINC) [1] terminology is the international standard for the exchange of clinical health information, including measurements, observations, and documents. This is a hierarchical set of codes for clinical assays, and does not directly integrate with the semantic web, although its structure informs our own.

There are additional established ontologies and standards that could be brought into the SORBET model; a few of these are mentioned here. The Observational Medical Outcomes Partnership (OMOP) Common Data Model [16], designed to help unify disparate medical observational databases. The Clinical Data Interchange Standards Consortium (CDISC) [2] includes models, domains, and specifications for data representation in medical research and related areas of healthcare. The Biomedical Research Integrated Domain Group (BRIDG) model [5] has a goal to represent "basic, pre-clinical, clinical, and translational research and its associated regulatory artifacts."

2.3 Representational framework used

The Resource Description Framework (RDF) is described as "a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed." [6]. This provides exactly the representational power that we need for our model.

2.4 A note about sensors and their data

In thinking about sensor data, it is important to think about and model in the ontology the different forms of data possible. A sensor device may be constructed with multiple hardware-level sensors. The data from the hardware level may then be processed to produce the biomarkers (observations) reported for that sensor. This processing might happen physically on the sensor device itself, or might happen in the sensor vendor's cloud, and the biomarkers reported might or might not include the hardware-level data. Some reported biomarkers may have been calculated over a span of time, for example, an average for a 24-hour day. For the purposes of downstream analyses (future studies that might not have been anticipated) in practice it can be helpful to capture and store both raw and processed data when this is an option.

3 GOALS FOR THE SORBET ONTOLOGY

The primary purpose of SORBET is to facilitate the design and application of reusable discovery algorithms for clinical trials conducted with wearable sensors. While we have started the design of this ontology as described in this paper, it is our hope that others in academia and industry who are working in this space will join us in the development of this ontology. In this section, we describe some of our guiding principles.

3.1 Distinguish the observable from the observations

Sensor companies create the representation for the data coming from their sensor; these are the observations. A clinical biomarker is an observable, potentially captured by many different devices. We need to be able to identify common observables apart from the observations, so that we can identify comparable observables in the data. As a trivial example, the observable "heart rate" can be recorded as the observation "HR", "heart_rate", or "HeartRate".

3.2 Maintain flexibility in the representations of observable properties

Some devices record observations with qualifiers, for example, "step count" vs. "step count while moderately active", or "heart rate" vs. "heart rate while asleep". The representation must allow for annotations that record the qualifications on the sensor reading, as well as the cutpoints for examples such as these, and additional annotations on the semantics of those cutpoints. As another example, heart rate can be captured using electrical signals (electrocardiograph), or using light (photoplethysmography), and to maximize the potential for reuse of the data, we must represent the technology used for the device. It may be further important to represent the algorithm used to calculate a biometric from raw sensor data when there are multiple possible algorithms.

3.3 Create a semantic model that is flexible enough to evolve

As the spaces of medical sensors and of biomarkers for clinical trials continually evolve, our representation must evolve as well. We need to be able to capture new sensor data, new biomarkers, and new semantic structures, while still being able to make full use of data captured in the past. For Sensor Cloud, we have created a sensor ingestion pipeline for wearables and other sensors used in clinical trials. We have incorporated sensors from MC10, Actigraph, BioIntellisense, BioBeat, Indie Health, and Oxitone, and are continually onboarding additional sensors. (Ingestion involves capturing the data sent from the sensor company cloud or from their devices, including validating the correct receipt of the data and normalizing the stream into the SORBET representation. The original data streams are also saved, so that they can be reprocessed if necessary as the ontology changes.)

3.4 Our design and implementation path

To achieve our three goals, we crafted a custom ontology broadly representative of the LOINC model. Using RDF, we are able to make use of established relevant ontologies such as SKOS, SOSA, and QUDT to capture details of sensors, observations, and observables, and to allow an evolvable representational structure.

4 SORBET DESCRIPTION

As mentioned previously, the design of SORBET incorporates established ontologies to develop one specifically for use with wearable sensors used for clinical trials. Here is a discussion of the thinking behind some of the structural components.

- A biomedical sensor device may contain one or more sensors, so we must distinguish the device from the separate sensors that it contains. For example, a device may contain both an accelerometer and a gyroscopic sensor. Using the SOSA structure, the device is considered the "Platform" (sosa:Platform) and each sensor is a "Sensor" (sosa:Sensor).
- It is further important to note that there are biometrics that may be derived from sensor data, and may involve more than one sensor. This connects with the distinction between Observables (the biometrics) and the Observations (the data as reported from the sensor). The capabilities of the sensors are modelled based on the sosa:ObservableProperty, but in recognition that there are commonalities within and across platforms we abstract using an Observable class.
- The Observable class performs as a common concept for something to be recorded against the subject of the investigation (the patient); examples include heart rate, steps, and minutes of activity.
- The ObservableProperties apply a layering of context qualifiers on top of the Observables to account for a few uses. Aggregation properties for sets of observations over an epoch; examples include the mean, maximum, minimum, first, and last values Disposition properties for the scenarios in which the measures were taken; these are primarily driven by other measurements or calculations; examples are number of seconds of activity at a particular level or cutpoint, heart rate while sleeping, etc. Collection properties for the devices themselves, such as the 'Device Wear State', which might be any of 'worn', 'unworn', 'not specified', or 'unknown'.
- For a given sensor observation, the actual values may be reported in different contexts. For example, 'average heart rate' may be contextualized as 'average heart rate while sleeping', 'average heart rate while resting', and 'average heart rate while moving'.
- For each quantifiable value, the units need to be supplied. We chose to use a system where conversions are a first-order component (as the devices are generally not intended to support ranges of units). For this reason we used the UCUM (by way of QUDT).
- We saw parallels between the LOINC model and the sensor observations. For example, test codes and panels in LOINC are contextualised as a pre-coordinated vocabulary; a LOINC concept can be decomposed into parts such as 'Component/Analyte'and 'Kind'. The 'Components/Analyte' can be further decomposed into Analyte, Challenge, and Subcomponent; we modelled the Observable on the Analyte (i.e., it is a core term that gets reused) and have gone some

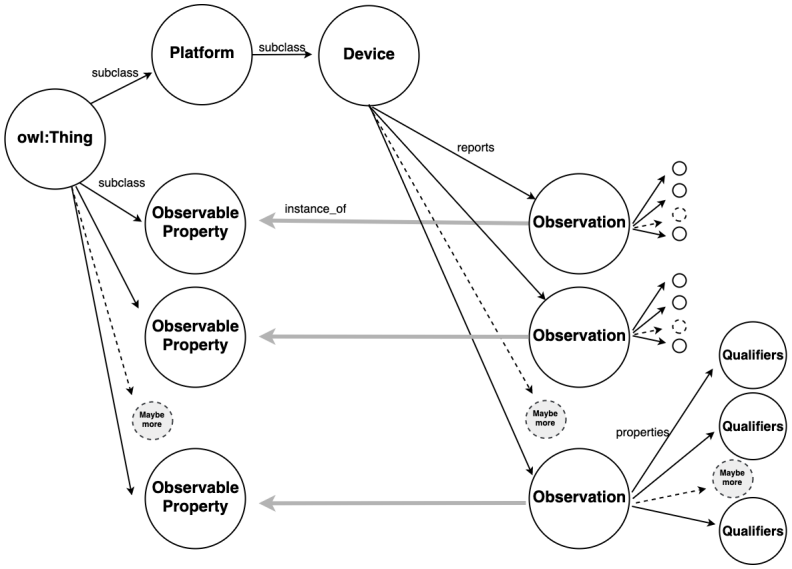

Figure 1: A generalized illustration of the SORBET ontology structure.

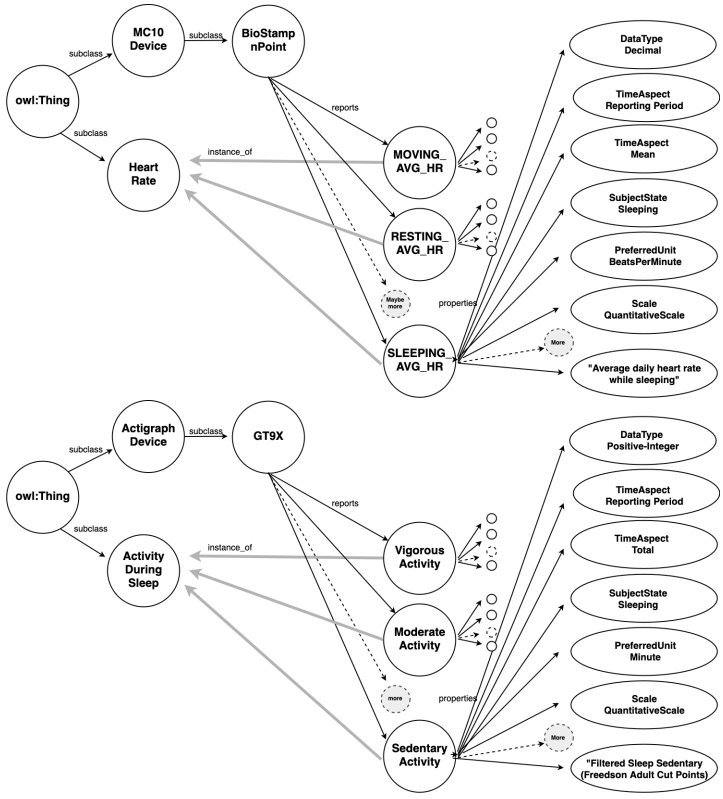

Figure 2: Illustrations of part of the of the SORBET ontology instantiated for the MC10 BioStamp nPoint sensor and the Actigraph GT9X sensor.

way towards providing a taxonomic structure (ie Subject-State.Sleeping is a narrower form of the Subject Physiological State) for the Observables; where we can unambiguously assign it, we've linked Observable to a discrete concept in the UMLS vocabulary through the Controlled Unclassified Information (CUI) term.

Figure 1 is an abstract illustration of the general structure the SORBET ontology. In terms of class structure, SORBET is a shallow ontology. The extensive additional annotations we seek to capture for sensor data is represented as properties. While the set of properties used is too large to list in this paper, here are some example properties:

Datatype Using the QUDT datatypes, we record the type of this property.

Time aspect One can either measure a Property at a moment (point) in time or measure it over a time interval and integrate, in the mathematical sense, over time. In the latter case, we combine a series of physiologic states into a single scalar value that reflects some aggregate property, such as "mean" or "sum".

Patient state For sensors that calculate the activity level of the patient at the time of a reading, the state they are categorized to be. For example, MC10 categorizes a patient as one of "moving", "resting", or "sleeping". These labels for activity level are determined using one of many cutpoint systems, such as Crowther or Staudenmayer, and the cutpoint system in use is also part of the annotations for the observation.

Cutpoint A specified value used to sort continuous variables into discrete categories. It may be set according to its demonstrated usefulness in predicting abnormal clinical events or arbitrarily.

Method The method of the measurement (e.g., radioimmunoassay, immune blot, etc.).

Unit A reference to the unit of measure of a quantity (variable or constant) of interest.

Adjustment The data element contains calculations that adjust or correct some measured value.

Again, there is a large number of available properties currently in SORBET and these are just representative examples. Additional properties are easily added as the ontology expands. We seek to record these details with the data to the extent possible to facilitate later biometric analysis. For example, if there is a wish to pool data from different previous clinical trials, it is essential to convert common units – e.g., mmol/L and mg/dL for blood glucose are easily resolved – and is just as important to know when the measurements are not directly comparable for an intended future use.

5 EXAMPLES OF USE

Section 4 provided a a description of the SORBET ontology; in this section, we would like to provide some examples of the ontology in use. First, we will describe and illustrate the "instanced" ontology for example sensors, and secondly, we will describe and illustrate how SORBET is designed to facilitate biomarker analysis.

5.1 Example SORBET representations

Figure 2 illustrates some of the specific structure for the MC10 BioStamp nPoint sensor and the Actigraph GT9X sensor. The MC10 illustration shows the observable property Heart Rate and the BioStamp nPoint observations for heart rate in each of three different subject states. Each of these observations is annotated with additional properties, such as the type of data, the units used, and a definition for the observation. The Actigraph illustration shows the observable property Activity During Sleep and three of the GT9X observations for this. Again, each of these observations is annotated with more details about the observation.

5.2 Example of comparing sensor streams

SORBET is designed to facilitate biomarker analysis, especially in the case of a single clinical trial that uses different sensors to collect the same biomarker data and in the case of downstream analysis, where data collected from previous clinical trials is repurposed to search for potential additional clinical outcomes. As we have only recently completed our ingestion pipeline to bring sensors online for use in clinical trials, we are still developing complex examples of using SORBET for combining multiple wearable devices into a single study. Here, we present a simple example using two different sensors to collect data on the same observable.

The Actigraph GT9X device reports a sleep duration observable property as the observation "TimeAsleepInMinutes"; as can be deduced from the name, the unit of measurement for this is minutes. The MC10 BioStamp nPoint device reports the same observable as the observation "SLEEPING_DURATION", using seconds as the unit of measurement. Both of these observations are labeled as the SORBET observable: "Epoch_Sleeping_Duration", with associated annotations for the units of measurement. Additional annotation properties clarify that for the Actigraph device, the observation is defined as "A daily aggregate of the non-partial epochs for subject that represent when the subject was asleep", while for the MC10 device, the observation is defined as "Total time in a 24-hour period that the patient is classified as sleeping." Thus, in this example, one can see that a simple conversion of units is likely all that is needed to combine data sets using the two different sensors when the biomarker of interest is the recorded sleeping time of the patient.

Figure 3 is an illustration of the MC10 BioStamp nPoint SORBET structure for sleep as compared to the Actigraph GT9X structure.

5.3 Assessing Autonomic Function by Heart Rate Variability

A recent review of the literature on the use of heart rate variability in assessing autonomic function [8] casts some doubt on the prevailing framework linking low-frequency (LF) and high-frequency (HF) components of heart rate variability (HRV) with sympathetic and parasympathetic autonomic divisions. The authors cite a study of their own [17], where they combine LF/HF HRV data with accelerometer data to show that long-term HRV are affected by contemporaneous physical activity and posture and that the autonomic functions linked to short term HRV measurements can not be generalized to the interpretation of long-term HRV. Having both HRV and accelerometer measurements from multiple studies in a single database with a common semantic model aids greatly in being able to identify such correlations.

6 FUTURE WORK

We are still early in the journey of building SORBET. With each new device integrated, we learn how to usefully extend it to better cover the domain. There are many stakeholders in the clinical trial process that have an interest in the development of a data model that facilitates faster, more ubiquitous biomarker algorithms, as well as the

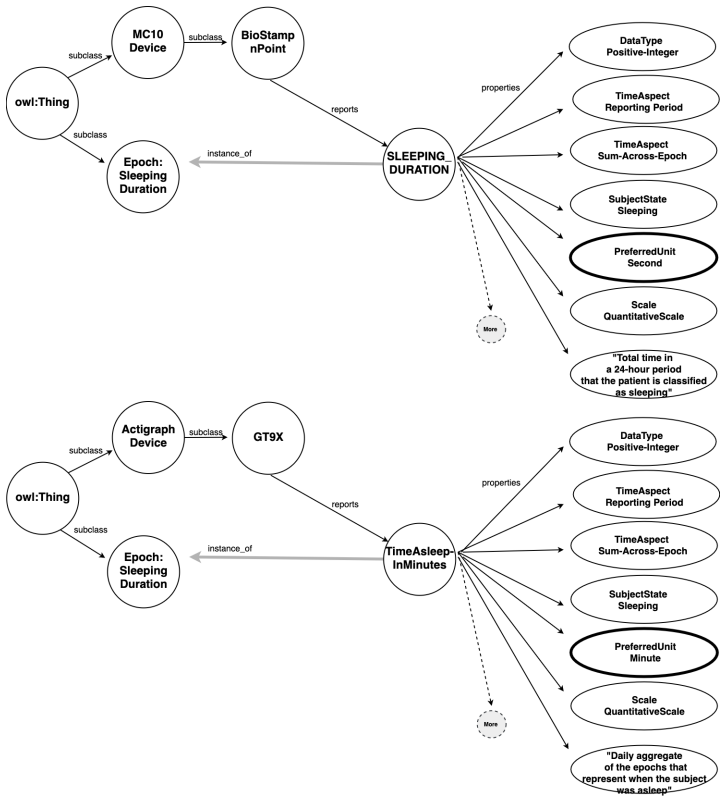

Figure 3: Illustrations of the MC10 BioStamp nPoint sleep graph and Actigraph GT9X sleep graph, showing some of the structure relative to reporting sleeping duration. The observation for MC10 is SLEEPING_DURATION, captured in seconds, while the observation for Actigraph is TimeAsleepInMinutes, captured in minutes.

discovery of new biomarkers. These stakeholders include patients, investigators, sponsors, CROs, data scientists, academic institutions, and others, all of whom stand to gain from this standardization effort. We are actively developing partnerships with many of these parties now in hopes of collaboratively and openly advancing the agenda of patient-centric research through scalable, semantically sound, digital biomarker analysis and discovery. We welcome any and all partners in this effort and hope when it matures sufficiently, to develop SORBET as an open industry standard.

REFERENCES

[1] LOINC Committee. 2021. LOINC homepage. https://loinc.org/ "The international standard for identifying health measurements, observations, and documents.".

[2] Clinical Data Interchange Standards Consortium. 2021. CDISC Standards. https://www.cdisc.org/standards

[3] Shaker El-Sappagh, Farman Ali, AAbdeltawab Hendawi, Jun-Hyeog Jang, and Kwak Kyung-Sup. 2019. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. *BMC Med Inform Decis Mak* 19, 97 (2019), 36. https://doi.org/10.1186/s12911-019-0806-z

[4] Raúl García-Castro, Armin Haller, and Nandana Mihindukulasooriya. 2021. *On the usage of the SSN ontology.* W3C internal document, 18 March 2021. W3C. https://w3c.github.io/sdw/ssn-usage/#analysis-ontologies

[5] Biomedical Research Integrated Domain Group. 2021. BRIDG: An international standard for Biomedical research concepts designed to support computable semantic interoperability. https://bridgmodel.nci.nih.gov/

[6] W3C RDF Core Working Group. 2021. Resource Description Framework (RDF). https://www.w3.org/RDF/

[7] Armin Haller, Krzysztof Janowicz, Simon Cox, Danh Le Phuoc, Kerry Taylor, and Maxime Lefrançois. 2017. *Semantic Sensor Network Ontology.* W3C Recommendation, 19 October 2017. W3C. https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/

[8] Junichiro Hayano and Emi Yuda. 2019. Pitfalls of assessment of autonomic function by heart rate variability. *J Physiol Anthropol.* 38, 3 (2019), 8.

[9] Mark Hennessy, Chris Oentojo, and Steven Ray. 2013. A framework and ontology for mobile sensor platforms in home health management. In *2013 1st International Workshop on the Engineering of Mobile-Enabled Systems (MOBS).* IEEE, San Francisco, CA, 31–35. https://doi.org/10.1109/MOBS.2013.6614220

[10] Krzysztof Janowicz, Armin Haller, Simon J. D. Cox, Danh Le Phuoc, and Maxime Lefrançois. 2019. SOSA: A Lightweight Ontology for Sensors, Observations, Samples, and Actuators. *Journal of Web Semantics* 56 (2019), 1–10. https://doi.org/10.1016/j.websem.2018.06.003

[11] Moritz Lehne, Julian Sass, Andrea Essenwanger, Josef Schepers, and Sylvia Thun. 2019. Why digital medicine depends on interoperability. *NPJ Digital Medicine* 2, 19 (August 2019), 5. https://doi.org/10.1038/s41746-019-0158-1

[12] Alistair Miles and Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference.* W3C Recommendation, 18 August 2009. W3C. https://www.w3.org/TR/2009/REC-skos-reference-20090818/

[13] Laura Miron, Rafael S. Gonçalves, and Mark A. Musen. 2020. Obstacles to the reuse of study metadata in ClinicalTrials.gov. *Scientific Data* 7, 443 (12 2020), 14. https://doi.org/10.1038/s41597-020-00780-z

[14] QUDT.org. 2021. QUDT Home Page. http://www.qudt.org/pages/HomePage.html

[15] Gunther Schadow and Clement J. McDonald. 1999-2014. The Unified Code for Units of Measure. https://ucum.org/trac

[16] Observational Health Data Sciences and Informatics. 2021. OMOP Common Data Model. https://www.ohdsi.org/data-standardization/the-common-data-model/

[17] Yoshida Y, Ogasawara H, Yuda E, and Hayano J. 2016. What does LF/HF of heart rate variability in ambulatory ECG mean? Effect of time in lying position during monitoring. *Eur Heart J.* 37(suppl) (2016), 2.