2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing

# A Linked-data Model for Semantic Sensor Streams

Payam Barnaghi, Wei Wang
Centre for Communication Systems Research
The University of Surrey
Guildford, GU2 7XH, United Kingdom
Email: {p.barnaghi, wei.wang}@surrey.ac.uk

Lijun Dong, Chonggang Wang
InterDigital Inc.
781 Third Avenue
King of Prussia, PA 19406, USA
Email: {lijun.dong, chonggang.wang}@interdigital.com

***Abstract*—This paper describes a semantic modelling scheme, a naming convention and a data distribution mechanism for sensor streams. The proposed solutions address important challenges to deal with large-scale sensor data emerging from the Internet of Things resources. While there are significant numbers of recent work on semantic sensor networks, semantic annotation and representation frameworks, there has been less focus on creating efficient and flexible schemes to describe the sensor streams and the observation and measurement data provided via these streams and to name and resolve the requests to these data. We present our semantic model to describe the sensor streams, demonstrate an annotation and data distribution framework and evaluate our solutions with a set of sample datasets. The results show that our proposed solutions can scale for large number of sensor streams with different types of data and various attributes.**

## I. Introduction

The rapid increase in number of network-enabled devices and sensors deployed in the physical environments is changing the information communication networks. It is predicted that within the next decade billions of devices (Cisco predicts that the number of the Internet connected devices will be around 50 Billion by 2020 [1]) will generate myriad of real world data for many applications and services in a variety of areas such as smart grids, smart homes, e-health, automotive, transport, logistics and environmental monitoring [2]. The related technologies and solutions that enable integration of real world data and services into the current information networking technologies are often described under the umbrella term of the Internet of Things (IoT) [3].

Network-enabled sensor devices (and the wireless sensor networks) are the key technologies that enable capturing and communicating the observation and measurement data collected from the physical environments. The heterogeneity and complexity of the sensor devices and their underlying networks make seamless integration of their data and services into the existing higher-level applications and services a challenging task. A potential solution to address this heterogeneity issue is using service-oriented mechanisms to provide common interfaces and to develop scalable and loosely coupled applications on that represent the IoT devices, networking resources, and the IoT data [4], [5] [6]. This represents sensors or other devices as services in the cyber-world and enables efficient provisioning and management of these devices and their data [7]. However, as most of the IoT devices operate in the real world environments, the exposed services are not as reliable and stable as those well-engineered and maintained business services and the quality of information and the services in IoT domain can vary over the time. The heterogeneity of underlying devices and networks also makes it difficult to provide one-fit-all solutions to represent data and services that emerge from the IoT networks. This brings significant challenges to data integration, data fusion and discovery mechanisms that require interoperable and machine-interpretable data and quality descriptions. In recent years a number of efforts have been made to model sensor networks and their data using machine-interpretable and interoperable formats. The existing work often use solutions that are adapted from the semantic Web and semantic data modelling to overcome the interoperability issues and to provide semantically rich descriptions for the IoT data. The recent advancements in this area are discussed in several existing works including the W3C's Semantic Sensor Network Incubator Group's ontology (SSN Ontology) [8], and other research reports such as [9], [10], [11]. The research on the IoT data so far has largely focused on knowledge representation, i.e. how to semantically describe capabilities of IoT devices and services [12], [7], data annotation and publications, i.e. how to create and publish semantically annotated IoT data and linked data models [13]. However, modelling and integrating the observation and measurement data, streaming sensor data and providing discovery mechanisms to enable distributed query mechanisms are other key issues to enable end-to-end solutions for publications and consumption of the sensory data emerging from IoT resources.

In this paper we describe a semantic modelling framework to annotate streaming sensor data. Sensor streams are data sources that represent observations and measurements collected by sensory devices. The data is collected and provided through an observation period and in a sense the sensor stream data can be seen as time series data. We discuss the modelling scheme and the linked-data approach to create lightweight and expressive descriptions for the semantic sensor streams. An annotation tool is provided to describe the spatial, temporal and thematic attributes of the data. The proposed model uses a *geohashing* mechanism to describe the spatial attributes of the data and we discuss a clustering mechanism that uses *"geohash"* specifications to distribute the stream data among different repositories. The efficiency of the representation by considering the size and flexibility of the data descriptions, and the data distribution and clustering methods are also described and evaluated. The remainder of the paper is organised as follows. Section 2 describes the related work. Section 3 discusses semantic annotation of streams and demonstrates the proposed semantic model. In Section 4 naming, distribution and resolution are described. Section 5 provides an evaluation of the proposed framework and discusses the open issues. Section 6 concludes the paper and describes the future work.

978-0-7695-5046-6/13 $26.00 © 2013 IEEE
DOI 10.1109/GreenCom-iThings-CPSCom.2013.95

## II. Related work

The Sensor Web Enablement working group[1] at the Open Geographical Consortium[2] has created an XML scheme and a standard model, called SensorML, to describe sensors systems and processes related to sensor observations [14]. The XML descriptions in SensorML, however, provide limited means to describe and link the domain knowledge and external annotation concepts to describe spatial, temporal and thematic attributes of the observations and measurements. In [10] a data description model that is adapted from SensorML observation and measurement scheme is discussed. The model is represented in the RDF form. While the RDF representation of the SensorML model enabled linking and annotating the data using external domain knowledge, the data description model was complex and unsuitable for large-scale data annotation and processing (specially in the constrained environments).

The SSN ontology [8] defines a higher-level scheme to link the observation and measurement data to sensor systems and device related attributes. However, the SSN ontology is developed for semantic sensor network descriptions and does not provide detailed descriptions for the observation and measurement data. In [15] an observation model and around 20,000 annotated data from the weather stations in the United States are described. The observation model in [15]captures the time, location and type attributes of the observation data and also provides links to locations in GeoNames that are near the weather stations. However, the model does not describe any mechanism to query the data based on the location proximity and it also does not provide an annotation framework to describe the detailed attributes of the data (e.g. quality and sensing device related attributes). In our previous work described in [13] and in a similar work called the Linked Sensor Middleware (LSM) [16] that is developed at DERI, the sensor descriptions and their data are annotated using relevant links from DBPedia and GeoNames concepts. These two platforms mainly focus on annotating the data and sensing resources and provision the data via common interfaces.

The work described in the current paper proposes an optimised observation and measurement data that uses a linked-data approach to annotate the streams and the observation and measurement data in them. The model which is described in the following section provides a flexible annotation scheme and the distribution and resolution of data can be also provided by processing the attributes described in the core model. In the next section, we describe the naming and distribution mechanisms to enable efficient query and resolution of the data.

## III. Data modelling and annotation

The sensory data represents physical world observation and measurement and requires time and location and other descriptive attributes to make the data more meaningful. For example, a temperature value of *15 degree* will be more meaningful when it is described with spatial (e.g. Guildford city centre) and temporal (e.g. *8:15AM GMT, 21-03-2013*), and unit (e.g. Celsius) attributes. The sensory data can also include other detailed meta-data that describe quality or device related attributes (e.g. Precision, Accuracy). Using semantic descriptions can provide machine-interpretable and interoperable data descriptions for sensor streams. The semantic sensor streams will include raw sensory data annotated with semantic descriptions that specify spatial, temporal and thematic and other attributes of the data.

[1]http://www.opengeospatial.org/projects/groups/sensorwebdwg

[2]http://www.opengeospatial.org

TABLE I: Comparing IoT data streams with conventional multimedia streams

| **Attributes** | **IoT data** | **Conventional data streams** |
|---|---|---|
| *Size* | often very small; some IoT data can be a real number and unit of measurement; the meta-data is usually significantly larger than the data itself | usually much larger than IoT data (video data) |
| *Location dependency* | most of the time location dependent | normally not location dependent |
| *Time dependency* | time dependent; need to support various queries related to temporal attributes | normally not time dependent |
| *Life span* | usually short lived or transient | long lived |
| *Number* | often very large | usually smaller than IoT data items |
| *Persistency* | some of the data needs to be archived | usually persistent |
| *Resolution* | names created from meta-data for resolution could be longer than conventional data (taking into account temporal and spatial dimensions) | resolution is usually based on names |

As most of the network-enabled sensor devices and sensor networks are resource constrained (i.e. often have limited power, bandwidth, memory, and processing resources), the semantic sensor streams should also support in-network data processing to aggregate or summarise the data to reduce the communication overload. To reduce the amount of information that needs to be transmitted across networks, the data model should be lightweight while accurately and sufficiently capturing the key attributes of the data. In the cases that the semantic annotation is considered to be performed on a more powerful intermediary node (e.g. a gateway node) there still will be vast amount of streaming data where the size of meta-data is significantly larger than the original data. In such cases, there must be a balance between expressiveness, level of details and complexity and size of the meta-data descriptions. In this paper, we define a core model for semantic sensor streams and propose a linked-data approach to provide an optimized semantic description model while keeping it expressive and flexible for various types of data with different quality requirements and specifications.

### A. Semantic sensor streams

To design our model we review some of the distinctive features of the IoT data that can be identified by comparing them to the conventional multimedia data streams (shown in Table 1).

As shown in the table, the streaming model for sensory data (or in general IoT data) should consider the volume, variety, velocity of change and time and location dependencies while describing observation and measurement values. Another aspect that should be taken into consideration is how the data will be used and queried. The sensor data queries can consist of information with one or several attributes such as location

(e.g. location tag, latitude and longitude values), type (e.g. temperature or light), time (e.g. timestamps, freshness of data), value (including observation and measurement value, value data-type and unit of measurement) and other meta-data (e.g. quality of information, what is described by the observation and measurement data).

It should be also noted that the queries may combine information from several attributes and also from several sources. The possible types of queries from an individual stream can be identified as:

- *Exact queries* - where the key data attributes are known. For example, Type, Location, or Time attributes are defined for a requested data. Other meta-data attributes such as quality of information (QoI) or Unit of measurement can be also provided.
- *Proximate queries* - where data from an approximate location or with a threshold of quality of information is queried.
- *Range queries* - where a time range or location range is used to query the data.
- *Composite queries*- where the result of the query should be provided by integration (and processing) of data from different sources and sometimes with different types.

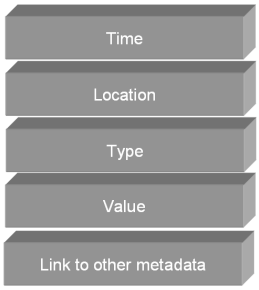


Fig. 1: The observation and measurement data attributes

The primary step to query, discovery and/or integration of data streams is being able to access and interpret the observation and measurement data and their attributes. Given the features of the sensory data and possible query types, we propose a semantic model for describing the sensor streams. The model describes the streaming data with the following main attributes (as shown in Figure 1):

- Location (e.g., location tag, location area)
- Time (e.g., timestamps)
- Type (e.g. temperature, humidity)
- Value (i.e, actual observation and measurement value and data-type)
- Links to other meta-data (i.e. linked to descriptions that provide source or quality of information related attributes)

A graph demonstration of the proposed model is also shown in Figure 2. Some other attributes such as ID, unit and data-type shown in Figure 2 are described in the following section.

### B. Semantic annotation

Semantic descriptions of the sensor streams in the proposed model follow a linked-data approach: the data items can be linked to existing concepts that are defined in commonly used ontologies or vocabularies; and the detailed meta-data and source related attributes can be provided as links to other sources. However, it is important to note that designing a model without providing efficient tools and mechanisms to annotate the data will not solve the interoperability and data description issues. In other words, the model provides a schema for describing the data and sensor streams but without having an effective solution to provide the detailed attributes that are designated in the semantic model, the annotations will still vary from one source to another. This will make the interpretation of the meta-data still a challenging and error prone task [7]. In this section we describe our annotation framework for the proposed model and demonstrate a tool that is designed to create semantic descriptions and templates for data annotations according to the semantic model.

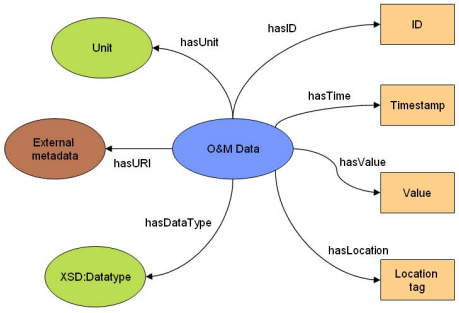


Fig. 2: The observation and measurement data attributes

To describe the location attribute, we use *geohash* tagging. Geohash is a mechanism that uses Base-N encoding and bit interleaving to create a string hash of the decimal latitude and longitude value of a geographical location [17]. It uses a hierarchical structure and divides the physical space to grids. Geohashing is a symmetric technique that can be used for geo-tagging. An interesting feature of the Geohash is that the nearby places will have similar prefixes in their string representation (with some exceptions[3]). In our annotation framework for semantic sensor streams, we use a Geohashing algorithm that employs Base32 encoding and bit interleaving to create 12bytes hash string representation of latitude and longitude geo-coordinates. For example the location of Guildford that has latitude value of "51.235401" and longitude value of "-0.574600" is represented as *"gcpe6zjeffgp"*.

Figure 3 shows 4 sample locations at the University of Surrey campus marked on a Google Map. On the left side of Figure 3 the *geohash* string tags for the shown locations are presented. As highlighted in the figure, the locations

[3]In Geohashing algorithm, the locations that are close to each other but on opposite sides of the Equator and the nodes that are a meridian can result in Geohash codes that do not have a common prefix.

with close proximity have similar prefixes. As the distance becomes closer the length of the prefix similarity increases (e.g. locations 2, 3 shown in the figure). This concept and using a simple string similarity method can provide location based search in querying and discovering the data. We also use the location prefixes to create an aggregated prefix when several data are integrated or accumulated from different locations with close proximity. In this case, the longest prefix string that is shared between all the data items is used to represent an aggregated location prefix tag for the data.


Fig. 3: Sample locations on a Google Map and their equivalent geohash strings

However, in some cases the stream providers or data consumers may know the higher-level location name (i.e. concept) but do not have the geo-coordinate values. We use the location concepts that are available via DBPepdia[4] and GeoNames[5] resources that are publically available as a part of the Linked Open Data cloud[6]. GeoNames contains over 10 million geographical names and 5.5 million alternate names of popular places. For example the latitude and longitude values of the location "Guildford" can be obtained by querying DBPedia and GeoName repositories (see Figure 4). For the query and inference of the semantic data that is available on DBPedia and GeoNames, we use public SPARQL-end points provided for DBPedia[7]. We have developed an online client and inference mechanisms that use AJAX technology [18] to query DBPedia and process the semantic descriptions to obtain the longitude and latitude values of the location concepts. This method is discussed in detail in our previous work in [13].

Fig. 4: The available information for a sample location (i.e. "Guildford") on DBPedia

For the unit of measurement, we use concept from the NASA's Sweet ontology[8]. The type of measurement attribute can be linked to the existing concepts on a common vocabulary. In the case of our prototype development, we have used concepts from DBPedia to describe type of sensors that enable the streams (e.g. temperature sensor, smoke sensor) but in a real world application a more specific ontology or for describing the type of sensory data can be employed.

The above attributes and the observation and measurement value, data type and timestamps create the core description model. Additional features such as the source related data (i.e. how the data is measured, using what device or quality of information) can be added in a modular form (e.g. adding a new semantic description module to describe the quality of information attributes or measurement range properties, etc. and linking them to the core descriptions) as they can be linked to information available on other sources such as the provider device itself, gateway, etc. This approach provides a flexible solution to describe the streaming sensor data where the model captures the core attributes of the data and the additional information can be provided as linked-data. Figure 5 shows a sample measurement that is annotated according to the proposed model.


Fig. 5: A sample measurement data represented in RDF

Figure 6 shows a screen capture of the annotation tool that creates the semantic representations according to our proposed model. The annotation tool uses direct SPARQL queries to allow selection of concepts from external sources such as DBPedia and Sweet Ontology and provides basic inference functions to extract geo-coordinate values from the available location concepts. The use of DBPedia concepts for annotating the data, however, is only an example to shown how our proposed framework can be utilised for describing the data using existing knowledge and external resources. In practice, multiple public or proprietary (depending on the application and requirements) sources can be used to annotate the data. The main advantage of this method is that the more users and stakeholders in an application use the same existing concepts and knowledge to describe their data according to a common model, the better understanding of the properties of data and their relation to domain knowledge will be provided. This will enable developing more enhanced inference and processing mechanisms to interpret the data and to integrate them into or utilise them in different applications.

The annotation tool enables to create sample data sets or design templates for the semantic streams; in the real world applications, however, it is not feasible to expect that each individual item will be annotated manually or using an online tool. To annotated streaming data, the providers that publish/submit the data can use a template that is created semi-automatically or manually by submitting the spatial and thematic attributes of the stream. The template can then be altered when any of the location, type or attributes of the sensor stream change.

[4]http://dbpedia.org/
[5]http://www.geonames.org/
[6]http://linkeddata.org/
[7]http://dbpedia.org/sparql
[8]http://sweet.jpl.nasa.gov/ontology/

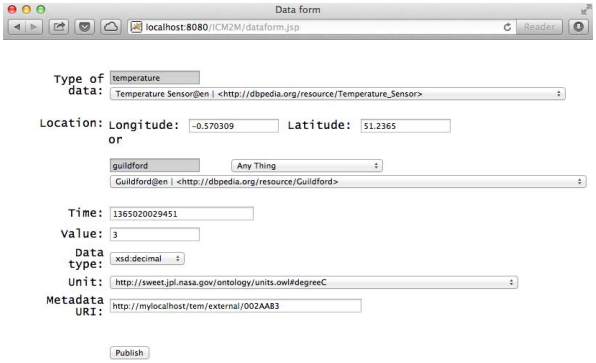


Fig. 6: The semantic stream annotation tool

Another important issues is that the observation and measurement data originated from the semantic streams are usually large number of individual values and annotating each individual item will create large volumes of meta-data that are not suitable for constrained environment and will also make processing and archiving these data less efficient. In this paper we have proposed two approaches to address this issue. The first approach is to annotate an individual item and then link the subsequent items in the stream (or a time window of a stream) to that item. A sample description of a measurement value that is linked to the item shown in Figure 5 is demonstrated in Figure 7.

```
- <rdf:RDF>
  - <rdf:Description rdf:about="kppvc5dv6ds98e2bf9e032aea0b0532dca2af1ef3e9cfa1b7b5aad99d8aa827922dd3f0c086b">
      <wot:time>1361891922203</wot:time>
      <wot:value>22</wot:value>
    - <wot:stream>
        kppvc5dv6ds9b7bf0323221e4579dd00d5f9bc73eb11a8bce20df29546bf3846efda80b8fce7
      </wot:stream>
    </rdf:Description>
  </rdf:RDF>
```

Fig. 7: A sample measurement data that is linked to the example shown in Figure 5

The above representations for streaming data reduce the size of descriptions significantly by using the core attributes and providing the details as linked-data (which can be retrieved upon request) but there are still repetitive data such as stream ID, RDF headers that repeat over each annotated item. In the second approach, we create one representation for a series of data (within a time frame or for the whole stream) that emerges from a semantic stream. A sample representation for the integrated semantic stream is shown in Figure 8.

The example shown in Figure 8 demonstrates annotation of staionary data. However in some applications the stream provider can be mobile. The proposed model can be also adapted for mobile streams. Figure 9 illustrates the data annotation for a mobile streams where the location of data is changed through the time.

## IV. Naming, Distribution and Resolution

The data collected from sensor streams can be stored on the sensing device (i.e. network-enabled sensor nodes), intermediary nodes (i.e. gateways) for short-term access or the data

```
- <rdf:RDF>
  - <rdf:Description rdf:about="ug8gvkvmg0g1c28540c96568e287b3371abb91119acba0af384a00ef3cb7901178a987ab1c32">
      <wot:type>http://dbpedia.org/resource/Pressure_sensor</wot:type>
      <wot:location>ug8gvkvmg0g1</wot:location>
    - <wot:series>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>11</wot:value>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>26</wot:value>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>8</wot:value>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>9</wot:value>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>21</wot:value>
        </wot:item>
      - <wot:item>
```

Fig. 8: A set of measurements in a semantic stream

```
- <rdf:RDF>
  - <rdf:Description rdf:about="ug8gvkvmg0g1c28540c96568e287b3371abb91119acba0af384a00ef3cb7901178a987ab1c32">
      <wot:type>http://dbpedia.org/resource/Pressure_sensor</wot:type>
    - <wot:series>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>11</wot:value>
          <wot:location>ug8gvkvmg0g1</wot:location>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>26</wot:value>
          <wot:location>ug8gvkvmg0g1</wot:location>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>8</wot:value>
          <wot:location>ug8gvkvmg0b1</wot:location>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>9</wot:value>
          <wot:location>ug8gvkvmg0g1</wot:location>
        </wot:item>
      - <wot:item>
          <wot:time>1361891926765</wot:time>
          <wot:value>21</wot:value>
          <wot:location>ug8gvkvmgfg1</wot:location>
        </wot:item>
```

Fig. 9: A set of measurements in a mobile semantic stream

can be archived on repositories and stored on directory servers. Access to the data can be also provided using common interfaces and (web) services. The data access methods and service interfaces are discussed in several existing work including [6], [16], [19], [20]. While the existing solutions provide efficient solutions to represent the services and facilitate accessing the sensory data, there is a lack of naming conventions for the sensor streams and providing solutions for distribution and resolution of the stream data when there are a number of stream sources in a domain. Figure 10 demonstrates a generic architecture for publishing and accessing the streaming sensor data.

In the proposed architecture the data can be accessed by referring to stream or individual data ID attribute (if the source of data is known) or the data can be queried based on attributes such as spatial, temporal and type. The ”#” sign in Figure 10 shows where the indexes of data can be stored. Before discussing the resolution and distribution mechanisms, we describe how the streams and data items are provided with unique names (i.e. IDs). Providing a unique naming scheme will make the stream data publication and integration of the streaming sensor data into the current Internet networks more data-centric as the data can be accessed by referring to their

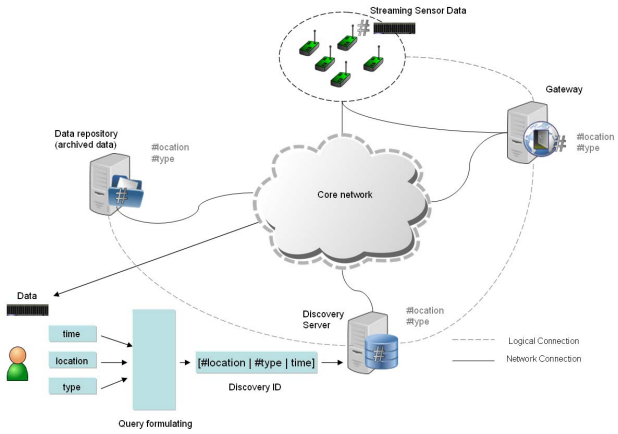


Fig. 10: An architecture for sensor stream data publication, storage and resolution

attributes or to their IDs regardless of the source that provides the data. However, if the source related information were important for an applications (and if the relevant security and privacy procedures were in place), this information can be accessed using the external meta-data links that are included in the semantic description model described in Section III-B.

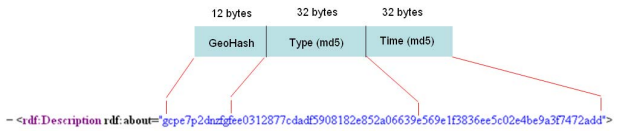


Fig. 11: Naming and ID construction

To name a stream or a data item, we use location, time (for a stream this will be starting time of the measurements in the current window of the stream) and the type information. We create a long string to represent ID of a stream or a particular data item. The ID is generated using the geohash tag of the location information, and the MD5 digest of the type and time values. A sample ID construction is shown in Figure 11.

To distribute the data among different repositories (or short-term cache on the gateways) we use a clustering method that distributes the data into different clusters. Each data cluster can be assigned to (and stored in) a repository or the clusters can be used to divide the data within a repository to provide fast query and resolution mechanisms to access the data. To query the data, we use SPARQL queries and semantic processing of the annotations. These methods are adapted from our previous work reported in [13], [21]. Using the SPARQL queries for semantic data is not novel on its own; the challenge that is addressed in the paper is how to make semantic queries more efficient while dealing with large-scale annotated data. One approach that is mainly taken by the researchers in information retrieval and the semantic web community is to enhance the query processing, storage functions, and more efficient query processing techniques [22].

In this work, we use a K-means clustering mechanism [23] to distribute the data among different repositories and then use a prediction method based on the clustering model to identify the repositories that maintain each part of the data. This enables dividing large volumes of semantically annotated data to smaller clusters and then running standard SPARQL queries within each cluster to find the relevant data according to user queries. Section V discusses the evaluation work and demonstrates the results of our preliminary experiments using the proposed mechanisms.

## V. Evaluation

To evaluate the stream modelling scheme, the naming and data distribution mechanisms, we have implemented a dataset generator that creates sample annotated data according to the proposed semantic model. The evaluation data includes three sample datasets that each include 500 samples of the semantically described sensor streams. The annotated data were created according to the normal data descriptions, linked streams (where one item is annotated with all the core attributes and other the subsequent data items in the stream are linked to this item), and stream annotation where all the data items for a stream (or a window of a stream) are included in one annotated construct. The annotations for these sample data sets are provided using 30 different location tags for the streams. The dataset are represented in RDF form and are available at: http://tinyurl.com/ckv7fdl.

A comparison of the size of descriptions that are required to represent each dataset is shown in Figure 12. As can be seen from the Figure 12 including the series of observation and measurement data from a stream into one construct significantly reduces the size of descriptions (by 84% compared to the noraml form where all the items are annotated individually).

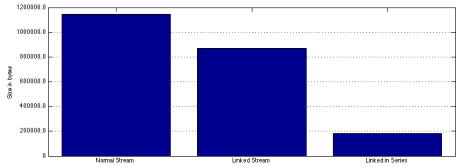


Fig. 12: The size of stream data using different representations

Annotation of all data items in a stream in one construct also enables addressing and processing the associated and linked attributes of the data using one value or concept for each attribute for all the items. In this form of representation, the size of stream series can be defined according to window size (that is used to divide the streaming data into frames). The size of the stream series can depend on the application requirements and the caching/buffering, freshness, bandwidth and communication resources and requirements. The stream processing and window based division of the streaming sensor data is not in the scope of this paper. In a previous work we have provided some discussions and demonstrations regarding the window based processing of the streaming sensor data. More information can be found in [24].

The name and ID descriptions in all the dataset are created according to the naming convention described in Section IV. To distribute the data, we use the *geohash* location tags and a

K-means clustering algorithm to divide the data into different clusters (shown in Figure 13 based on using 3 clusters). In our experiment we have assigned three directory servers to index and store the data. Each directory server maintains the data for one cluster. Within the directory servers, the location prefixes and "*type*" attributes of data are used to index and resolve the data requests. Once a query is narrowed down to a location and type, a standard SPARQL query is run on the directory server to retrieve the data. The *geohash* tags are represented as 12byte strings in our dataset. To cluster the *geohash* tags, the string representations are converted to equivalent ASCII code (which create a 12 columns vector for each *geohash* tag). The vectors are then fed to a K-means clustering algorithm. The clustering is preformed by using a Singular Value Decomposition (SVD) [25] to reduce the dimensionality of data before the clustering and the results are compared to using the normal 12 dimensions *geohash* vectors.

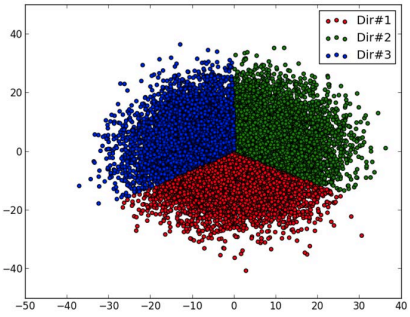


Fig. 13: Dividing the data into 3 clusters where each cluster can be stored on a directory server (shown as Dir#1,#2,#3)

The experiments are performed using 20,000 sample annotated stream data (the dataset is available at: http://tinyurl.com/boshz7e). The experiment is run on a Desktop Computer with a Pentium 4 CPU (2.4GHz) and 2GB RAM. Figure 14 shows the processing time for the training and predictions. As can be seen from Figure 14 the processing time in the learning phase and creating the clusters is significantly higher than the prediction time.The results also show that using SVD before the training reduces the time of clustering process.

Figure 15 shows the V-measure values when the number of clusters in the model are increased. The V-measure describes the mutual information (NMI) normalised by the sum of the label entropies. The V-measure values are shown for both using the original *geohash* vectors and also by using the SVD vectors for training the model. As can be seen from the figure, as the number of clusters increase the entropy level decreases and the homogeneity and completeness increase (i.e. higher V-measure).

### A. Discussion

As shown in the experiments, the training phase of the clustering process takes a significantly higher time than the prediction. However, the clustering process will only happen once to train the model and the subsequent data publication or query requests to identify the cluster labels can be decided

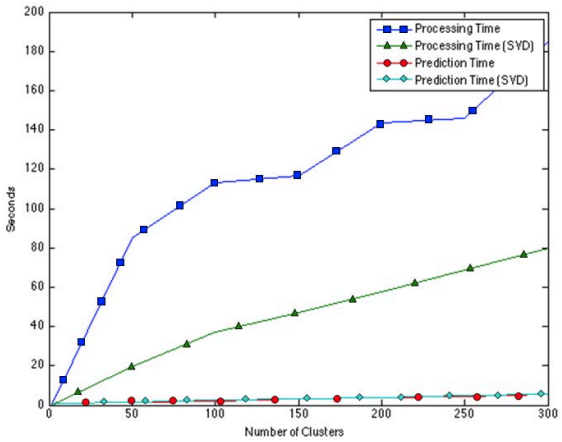


Fig. 14: Processing time for the clustering process

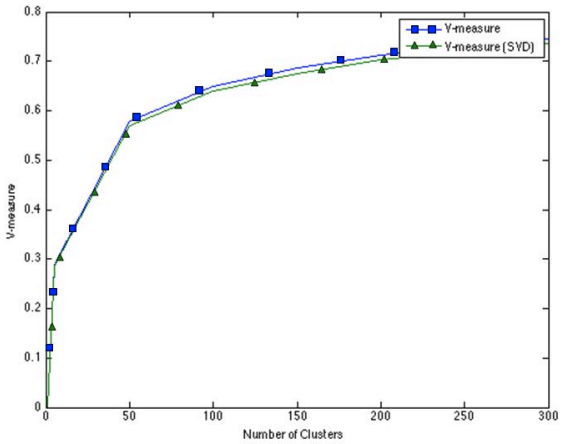


Fig. 15: V-measure evaluation when the number of clusters increases

by using the prediction function. In a practical setup, the training process for the model to perform the clustering can be performed off-line and as long as the number of location tags that are used in the training phase are not significantly changed, the model can be used for performing the predictions. If a significant number of new location tags are introduced (i.e. new resources are included from different locations or the existing resources move to new locations) then a new model can be trained to replace the existing model. It is also important to note that all the gateways or data publication and query provider nodes need to share the same clustering model in a domain to be able to distribute the data publication and query requests according to the cluster labels.

Increasing the number of clusters enhances the entropy and completeness of the data distribution in the cluster. However in real world applications, there will be a trade-off between the number of clusters and the size and variation of data in each cluster. The more clusters are used the more homogeneous distribution data will provided in each cluster; however, the latter will increase the load of repository allocation and query

handling. A solution to address this issue is distributing the data in a domain into clusters according to the number of gateways or directory servers/repositories that store the data. The clusters can be also used to create logical division between data stored in a gateway/directory server to distribute the query processing and data publication requests to different clusters.

In describing the naming scheme we stated that the generated IDs are unique. However in cases that two or more different sensor devices from the same location produce data, the naming mechanism will generate similar IDs for these data. To avoid this issue, a device ID can be also included in the ID descriptions; in the current work as we mainly focus on the data, the source and quality related attributes can be regarded as additional metadata to help differentiating data that are generated from different sources.

## VI. Conclusions

The paper describes a linked-data approach to annotate the sensor streams and employs a geohashing mechanism and type and time hash digest to name the data items and streams. A clustering mechanism is provided to distribute the data into different (logical or physical) directories. The linked-data model for annotating the streaming sensor data is demonstrated using a tool for creating the annotation templates. The proposed modelling scheme uses a set of core attributes to describe the stream data. The detailed attributes such as quality and source (device) related information are provided as linked-data. We have shown that the proposed model can efficiently reduces the size of the representations and can describe different attributes of the stream data while being optimised for storage and query processing purposes.

An architecture is proposed to demonstrate how data from sensor streams can be published, indexed, queried and discovered in a distributed network. A novel naming scheme is also introduced that uses a combination of *geohash* location tags, type and time digests and constructs a unique ID for the streams or the individual observation and measurement data items. The location tags are then used in a clustering mechanism to distribute the data among different directory servers and to predict the query or data publication destination based on the trained clustering model. We have evaluated our proposed methods with a dataset which is also made available online.

The future work will focus on hierarchical clustering for creating multi-domain solutions. We will also investigate creating interfaces to perform automated annotation and template alteration when the stream providers move from one source or one location to another.

## References


[1] D. Evans, "The internet of things: How the next evolution of the internet is changing everything," 2011.

[2] R. v. Kranenburg, E. Anzelmo, A. Bassi, D. Caprio, S. Dodson, and M. Ratto, "The internet of things," *Draft paper Prepared for the 1st Berlin Symposium on Internet and Society, Berlin, Germany (October 2011)*, 2008.

[3] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Comput. Netw.*, vol. 54, pp. 2787–2805, Oct. 2010.

[4] D. Guinard, V. Trifa, S. Karnouskos, P. Spiess, and D. Savio, "Interacting with the soa-based internet of things: Discovery, query, selection, and on-demand provisioning of web services," *IEEE Trans. Serv. Comput.*, vol. 3, pp. 223–235, July 2010.

[5] P. Spiess, S. Karnouskos, D. Guinard, D. Savio, O. Baecker, L. M. S. d. Souza, and V. Trifa, "Soa-based integration of the internet of things in enterprise services," in *Proceedings of the 2009 IEEE International Conference on Web Services*, ICWS '09, (Washington, DC, USA), pp. 968–975, IEEE Computer Society, 2009.

[6] W. Wang, P. Barnaghi, G. Cassar, F. Ganz, and P. Navaratnam, "Semantic sensor service networks," in *Sensors, 2012 IEEE*, pp. 1–4, 2012.

[7] P. Barnaghi, W. Wang, C. Henson, and K. Taylor, "Semantics for the internet of things: Early progress and back to the future," *Int. J. Semant. Web Inf. Syst.*, vol. 8, pp. 1–21, Jan. 2012.

[8] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor, "The SSN ontology of the W3C semantic sensor network incubator group," *Journal of Web Semantics*, vol. 17, pp. 25–32, 2012.

[9] A. Sheth, C. Henson, and S. Sahoo, "Semantic sensor web," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 78–83, 2008.

[10] P. Barnaghi, S. Meissner, M. Presser, and K. Moessner, "Sense and sens ability: Semantic data modelling for sensor networks," *Conference Proceedings of ICT Mobile Summit 2009*, 2009.

[11] M. Compton, C. Henson, L. Lefort, H. Neuhaus, and A. Sheth, "A survey of the semantic specification of sensors," *Proc. Semantic Sensor Networks*, vol. 17, 2009.

[12] S. De, P. Barnaghi, M. Bauer, and S. Meissner, "Service modelling for the internet of things," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pp. 949–955, IEEE, 2011.

[13] P. Barnaghi, M. Presser, and K. Moessner, "Publishing linked sensor data," *ISWC 2010*, 2010.

[14] M. Botts and A. Robin, "Opengis sensor model language (sensorml) implementation specification," 2007.

[15] H. Patni, C. Henson, and A. Sheth, "Linked sensor data," in *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*, pp. 362–370, 2010.

[16] D. Le-Phuoc, H. Q. Nguyen-Mau, J. X. Parreira, and M. Hauswirth, "A middleware framework for scalable management of linked streams," *Web Semant.*, vol. 16, pp. 42–51, Nov. 2012.

[17] B. Beatty, "Compact text encoding of latitude/longitude coordinates," Patent number: 7302343, November 2007.

[18] S. M. Lauriat, *Advanced Ajax: Architecture and Best Practices*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2007.

[19] L. Gurgen, C. Roncancio, C. Labbé, A. Bottaro, and V. Olive, "Sstreamware: a service oriented middleware for heterogeneous sensor data management," in *Proceedings of the 5th international conference on Pervasive services*, ICPS '08, (New York, NY, USA), pp. 121–130, ACM, 2008.

[20] W. Yu, T. N. Le, J. Lee, and D. Xuan, "Effective query aggregation for data services in sensor networks," *Comput. Commun.*, vol. 29, pp. 3733–3744, Nov. 2006.

[21] W. Wei and P. Barnaghi, "Semantic annotation and reasoning for sensor data," in *Smart Sensing and Context*, pp. 66–76, Springer Berlin Heidelberg, 2009.

[22] I. Filali, F. Bongiovanni, F. Huet, and F. Baude, "Transactions on large-scale data- and knowledge-centered systems iii," ch. A survey of structured P2P systems for RDF data storage and retrieval, pp. 20–55, Berlin, Heidelberg: Springer-Verlag, 2011.

[23] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. M. L. Cam and J. Neyman, eds.), vol. 1, pp. 281–297, University of California Press, 1967.

[24] P. Barnaghi, F. Ganz, C. Henson, and A. Sheth, "Computing perception from sensor data," in *Sensors, 2012 IEEE*, pp. 1–4, 2012.

[25] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Indust. Appl. Math.: Ser. B, Numer. Anal.*, vol. 2, pp. 205–224, 1965.