

Views from the coalface: chemo-sensors, sensor networks and the semantic sensor web

Jer Hayes*, Edel O'Connor¹, John Cleary¹, Harry Kolar², Robert McCarthy*, Richard Tynan³, Gregory M.P. O'Hare³, Alan Smeaton¹, Noel E. O'Connor¹, Dermot Diamond¹

¹ CLARITY: Centre for Sensor Web Technologies, Dublin City University, Glasnevin, Dublin 9, Ireland

²IBM Systems & Technology Group, East Fishkill, New York

^{*}IBM Innovative Environmental Solutions, Mulhuddart, Ireland

³CLARITY: Centre for Sensor Web Technologies, University College Dublin, Belfield, Dublin 4, Ireland

{hayesjer@ie.ibm.com; edel.oconnor@computing.dcu.ie; john.cleary@dcu.ie; kolar@us.ibm.com; rjmccarthy@ie.ibm.com; richard.tynan@ucd.ie; gregory.ohare@ucd.ie; alan.smeaton@dcu.ie; oconnorn@eeng.dcu.ie; dermot.diamond@dcu.ie}

Abstract. Currently millions of sensors are being deployed in sensor networks across the world. These networks generate vast quantities of heterogeneous data across various levels of spatial and temporal granularity. Sensors range from single-point in situ sensors to remote satellite sensors which can cover the globe. The semantic sensor web in principle should allow for the unification of the web with the real-world. In this position paper, we discuss the major challenges to this unification from the perspective of sensor developers (especially chemo-sensors) and integrating sensors data in real-world deployments. These challenges include: (1) identifying the quality of the data; (2) heterogeneity of data sources and data transport methods; (3) integrating data streams from different sources and modalities (esp. contextual information), and (4) pushing intelligence to the sensor level.

Keywords: Environmental sensor networks, chemo-sensors, metadata standards, sensor intelligence

1 Introduction

The semantic sensor web offers the unique opportunity to unify the real and virtual world [1]. The notion of unifying the real world with the virtual world has been described before as *internet-scale control*, a concept that which originated with IBM researchers Ron Ambrosio and Alex Morrow [2]. The notion of *internet-scale sensing* and how it relates to chemo-sensors has also been examined [3], and this suggestion made that the realization of large-scale sensor networks should be based on internet enabled sensors that allow for external browsing of the sensor's status, provide command and control, and facilitate feedback of information to individuals and other

devices (see Fig. 1). Such a view suggests pushing analytics down to the sensor level. The *internet-scale sensing* concept is very similar to current proposals for the Semantic Sensor Web (SSW) where such a system will need to be automatically deployed, automatically configured and have tailored delivery of information for a variety of users [1].

At its simplest a SSW is one where sensor data is annotated with semantic metadata to increase interoperability as well as to provide contextual information essential for situational knowledge [4]. According to the ‘Sensor Model Language (SensorML) for In-situ and Remote Sensors’ discussion paper a sensor is "an entity capable of observing a phenomenon and returning an observed value. A sensor can be an instrument or a living organism (e.g. a person)...", p12, [5]. Clearly, manual sampling still takes place in many parts of the world (where it may be a legal requirement) and this results in under-sampling. Under-sampling can be overcome either by using remote sample collection devices (auto-samplers) or by using sensor networks to continuously measure over long time periods. However, we should note that the SSW will have to deal with data from manual sampling (perhaps via a lab-based data management system) as well as near real-time data streams. This is important as “events” may be detected via SSW but confirmation may require autosamplers to grab samples from the same source at the same time, and flag that they are available; an operator has to go to the sample, and collect for a more sophisticated lab-based analysis that generate data acceptable in a court.

The variety of existing sensor networks is extensive but the core consistent with that of SSW in that it envisages a world in which the status of the real world is monitored by large numbers of distributed sensors, forming a sensor ‘mesh’ that continuously feeds data into integration hubs, where it is aggregated, correlations identified, information extracted, and feedback loops used to take appropriate action [6]. Sensor networks provide a web of interconnectivity that provides the multiple sources of information that will underpin more accurate decision making. Decision-making in the SSW will be absolutely essential. Decision making is a complex and demanding process which is often constrained in a number of possibly conflicting dimensions including quality, responsiveness and cost. We suggest that analytics in the SSW will happen across a continuum from one edge of the SSW to the other, from the sensors right through to the end user. By this we mean data stream analytics can occur on the sensor itself (see Section 4) and from any point in SSW to the end user who may apply his or her own analytics to a variety of data streams from the SSW.

Sensor networks are composed of sensor nodes which are the smallest component of a sensor network that has integrated sensing and communication capabilities and these can be wired (as in some coastal observatories) or wireless (as in wireless sensor networks). The sensor node has basic networking capabilities through communications with a base-station and sometimes other nodes. The simplest sensor nodes will have a microcontroller to perform basic processing operations but sensor nodes can also be highly complex, e.g. remote sensing instruments on satellite platforms, and which perform complex processing operations locally.

At present the culture in SSW/wireless sensor network (WSN) research is very

heavily biased towards transducers like thermistors which is understandable as they exhibit almost ideal behaviour - low cost, long-life, very low-power, small form factor, high accuracy and precision, rugged, reliable, etc. This bias colours the expectations of SSW/WSN researchers in that they expect all sensors to conform to this ideal. The physical sensor bias, in short, is the notion that *all* sensors act like thermistors and thus are held to be reliable or at least reliable enough that questions of data quality are relatively straightforward, and the sensors are simple to use and require little maintenance. In reality this is invariably not the case for a variety of reasons including, leaching of active components from sensing membranes (see [7]), physical damage, lack of selectivity, non-linear performance, baseline drift and biofouling (particularly in the marine environment). And as such the data stream generated by sensors (especially environmental sensors) is prone to data quality (trust) issues [8]. Physical sensors which are encapsulated can also be affected by extreme changes in temperature. The SSW system itself must be capable of examining the streams of data being imported, and both observing environmental events as they take place (e.g. pollution event) or observing erratic behaviour from a particular sensor and flagging it as unreliable and requiring attention. An end user who accesses data from the SSW should be given information on the quality of the data from a sensor or set of sensors and ultimately be given enough information to ascertain whether they wish to trust the data or not. But how is the analytics to take place? Should there be a standard? The solution to this may lie in current research on the interaction between the rule and the ontology layers of the Semantic Web.

In dealing with raw data streams we can ask – what does this data stream mean? Generally speaking data streams are not self identifying and we require outside information, metadata, to understand the stream. The main driver for the use of metadata has been data sharing. Scientists generate large amounts of data and often we wish to share this data with other researchers. This "data sharing" is made easier when the data formats are the same or at least interoperable. However, it is often the case that "data sharing" is difficult due to competing standards and a general lack of metadata. In the marine area there has been much work on establishing metadata standards. This work has been driven by practical needs as often because researchers are interested in phenomena that cover large areas where several groups or institutes are gathering data. To get a full picture of a particular phenomenon, e.g. an algal bloom along a coast line, a researcher may need to augment the in-situ sensor data streams with information from a variety of other sources (e.g. satellite information). Therefore in considering SSW applications it is worth examining current standards to see if they can fit into a SSW.

Another goal of *internet-scale sensing* and the SSW is to allow for plug'n'play sensors (or alternatively deploy-and-forget sensors). It is worth noting how difficult this is in the real world. Consider the example of an ocean observatory such as the one being developed by the Monterey Bay Aquatic Research Institute (MBARI). MBARI highlight that the network for the ocean observatory will use a wide variety of communication links: optical fibre, microwave, packet radio, satellite and acoustic [9]. This results in a diversity of throughput, latency and intermittence throughout the network. These are factors which will also affect the SSW in the real-world, namely –

can the SSW handle a diversity of throughput, latency and intermittence ? There has been work done on data transport protocols that guarantee delivery (e.g. [IBM's MQTT]) but where a sensor is continuously sending data it may be the case that dropping several packets is not considered a large problem. This is major challenge (and perhaps the most obvious) for the SSW – given the heterogeneity of data sources and data transport methods how can they all fit neatly into the SSW ?

Therefore, from the perspective of sensor developers (especially chemo-sensors) and with respect to integrating sensors data in real-world deployments there are four main challenges to the unification of the real and virtual world:

1. The heterogeneity of data sources and data transport methods that all must neatly fit into the SSW.
2. The quality of the data must be described and understood.
3. Data streams from different sources and modalities (esp. contextual information) which vary in across many dimensions, including spatial, temporal, granularity of data, must be integrated.
4. The SSW must be capable of supporting analytics (e.g. decision making) across the SSW nodes.

Some of these challenges are inter-related e.g. a solution to integrating data streams from different sources and modalities must respond to challenge (1).

2 Integrating data from heterogeneous sensors and modalities: a marine example

The question of what constitutes a sensor must not be constrained when considering the SSW. For example in the context of environmental monitoring, in-situ wireless sensor networks (WSN's) substantially reduce the need for costly labour-intensive on-site sampling and data collection. However WSN's pose some distinct disadvantages and we are often required to consider alternative or complementary sensing modalities. In order to identify and highlight some of the issues for the SSW in relation to integrating data from heterogeneous sensors, we will examine the scenario of marine monitoring. However many of the issues outlined here can equally be applied to other application scenarios.

Marine monitoring includes the observation of various aspects of the marine environment. It ranges from the detection of pollution and the development of harmful algal blooms to the monitoring of coastal features and coastal erosion. A recent analysis of developments in mote-based wireless sensor networks with respect to environmental monitoring [6] suggests that there are still many limitations with the current capability of these platforms for sustainable environmental sensing. As previously mentioned, in-situ sensors which are in direct contact with the environment are subject to the problem of bio-fouling and require regular maintenance. This can result in unreliable and noisy data or gaps in the sensing data. Furthermore, the more advanced chemo-sensors are still quite expensive to produce,

and at present require regular maintenance (from days to weeks). Therefore only a limited number of these may be deployed in the environment and are subject to regular maintenance. Due to the expense and logistical difficulties associated with in-situ networks and some of the problems outlined above, it is currently not possible to monitor a wide area over long periods of time with current mote-based wireless sensor networks. Also, in-situ sensor networks may not be suited to certain types of applications. For example Alexander and Holman [10] used an alternative sensing mechanism (video cameras) to quantify near-shore morphology of a coastal location since the turbulent nature of the surf zone often makes it difficult to maintain in-situ instrumentation.

2.1 Satellite-based remote sensing

Due to the limitations outlined above, alternative sensing modalities are often considered for the purposes of marine monitoring. Sophisticated satellite sensors are very effective for monitoring many parameters such as sea surface temperature, sea surface height, ocean currents, turbidity, and chlorophyll pigment concentration (which subsequently can be used to determine the amount of algal growth in the water). A number of these sensors are orbiting the earth on various satellite platforms. These sensors have differing spatial resolutions and operate on satellites with varying orbits and orbit cycle times (which subsequently affect geo-spatial comparison and temporal resolution).

Some of these sensors only operate in the solar reflective spectral range; hence they only gather useful data on cloud-free days during periods of daylight (i.e. when illumination conditions are suitable). For example, MERIS (Medium Resolution Imaging Spectrometer) onboard ESA's Envisat platform is a programmable, medium-spectral resolution, imaging spectrometer, which operates in this range. Its primary purpose is to measure the colour of the ocean and subsequently derive estimates of the concentration of parameters such as chlorophyll and suspended sediments. It has a spatial resolution of 1200m over the ocean and 300m over land and coastal zones and it completes coverage of the Earth in 3 days [11].

Other sensors are not subject to these restrictions and can provide data during periods of darkness and cloud-cover. An example of such a sensor is a radar altimeter which transmits microwave frequency pulses to the sea surface and receives the reflected echoes [12]. This type of remote sensor is often used for measuring sea surface height. The Ocean Surface Topography Mission (OSTM)/Jason-2 is an international satellite mission that was launched in June 2008 to extend the continuous climate record of sea surface height measurements. The main instrument onboard Jason-2 is an altimeter that measures the distance from the satellite to the ocean surface. It repeats its ground track every ten days, covering 95 percent of the world's ice-free oceans [13].

Many of these satellite sensor streams also arrive in various formats. Furthermore data is not overlaid on one common grid which affects geospatial comparison. There exists a number of European and international projects aimed at improving the interoperability of satellite sensor data; an example of such a project is the Global High resolution Sea Surface Temperature pilot project (GHRSSST-PP) which was

initiated by GODAE – Global Ocean Data Assimilation Experiment. GODAE identified that numerical ocean forecasting models require a near real-time supply of SST data, sampled often enough to resolve the diurnal cycle, along with an accuracy better than 0.2K and a spatial resolution better than 10 km which is only possible by combining the best capabilities of different types of sensors.

In 2002 GODAE initiated GHRSSST-PP [14] and the data products from this programme satellite provide SST observations from various satellite sensors in a common format (netCDF) together with a measure of uncertainty for each observation. This means that all satellite SST data are presented in a common format and the user doesn't have to re-code for the ingestion of different satellite data. The ancillary data provided allows the user to filter data based on the criteria outlined to their specific application. A combined analysis of all available SST data is carried out enabling the benefits of using in situ, microwave satellite SST and infra-red satellite SST in synergy. Diagnostic datasets are also produced for a number of sites around the globe. This is where all available data for a number of areas are gathered and subsequently resampled onto a common grid to assist intercomparison and characterisation of the various input data streams [15].

2.2 Alternatives to Satellite-based remote sensing

Another alternative sensing modality is that of optical airborne remote sensing. Its major benefit as opposed to satellite remote sensing is that the user can define its operational and deployment characteristics. It generally can provide much higher spatial resolution data and be deployed when atmospheric (i.e. cloud free), environmental, and solar conditions are acceptable to study specific phenomenon [16]. This can also be coordinated with in-situ sampling for algorithm validation or development. In coastal aquatic environments, many processes occur over space and time scales that cannot be adequately monitored using satellite remote sensing systems. The use of airborne remote sensing offers unique capabilities that enable specific coastal events to be studied. Coastal video systems have also been identified as effective tools for coastal monitoring and can be used to monitor coastal erosion, sea conditions, etc. A prime example of this is a major European research project entitled CoastView [17]. This is an alternative to the more expensive satellite and airborne remote sensing data which can also provide data over long periods of time at high spatial and temporal scales which is suitable for monitoring inland and coastal marine locations. Web cams and CCTV cameras are cheap and easily deployed. In fact, there are an abundance of web-cams available on the World Wide Web that can be used for monitoring purposes.

The singular use of video and images from cameras can act as a powerful sensing tool but even more so when used in combination with other sensing modalities. Davidson et al. [17] point out that despite the potential to improve monitoring of coastal zones with coastal video systems, that there are many coastal management issues that may only be addressed adequately through the integration of additional data sources and expert knowledge alongside the image data. For example, O'Connor et al have investigated the use of multiple sensing modalities in a river location [18] using video feeds with data streams on pH, temperature, turbidity, conductivity and

depth. It can provide some context around what is being sensed by monitoring various parameters. The conditions surrounding certain events may subsequently be deduced e.g. what are the environmental conditions surrounding an algal bloom event, what were the prior conditions to increased phosphate detected in the water. This can subsequently be used to automatically control the sampling frequency of in-situ sensors. Multiple sensing signals can also be used to deduce the quality of data and provide this information to the user.

2.3 Problems facing SSW from the coalface of the marine environment

The SSW should be aiming at providing similar functionality to GHRSSST-PP on a much larger scale. In effect the SSW must be able to provide an awareness of the capabilities, limitations and differences of the sensors and associated data streams. This is necessary in order to select appropriate data streams, from the diverse array currently available, to meet the needs of specific applications. Other problems outlined in this section were that: (a) the phenomena sensed is broad and this requires a broad suite of sensors / instruments which all have to be described / classified within the SSW; (b) In the real world we will have unreliable and noisy data or gaps in the sensing data and the SSW must account for this; (c) data is often in different formats (which currently results in users having to recode) and these must neatly fit into the SSW; (d) SSW should allow reasoning over heterogeneous multimodal sensor data and push intelligence to the sensor level i.e. if condition (x), condition(y) and condition(z) are met, start sampling more frequently or alert the responsible authorities that samples need to be taken immediately, (e) the SSW should accommodate interoperability of data streams and be able to deduce that the sensing signal in question is faulty or offline and subsequently provide the alternative.

All of these problems relate to the four challenges outlined in section 1. Challenge (1) which relates to the heterogeneity of data sources covers problems (a), (c), and (e) and so appears to be the largest challenge to SSW. While challenge (2) which relates to data quality is highlighted by problem (b), and (d) relates to challenge (4) where analytics occur across the SSW.

3 Precursors & building blocks of the Semantic Sensor Web

Considering the maxim that it is best not to reinvent the wheel we will outline a number of attempts at describing data and sensors. These attempts can be viewed as the precursors to the SSW or as possible components in future architectures. One movement in sensor research is towards making sensors web-resident thus making it possible to remotely discover, access, and use real-time data taken directly from the sensors. These three activities, discovery, access and use are fundamental to working with sensors. We will briefly describe three standards of data description and sensor description (1) the CDI XML schema [19], (2) MarineXML [20] / CSML and (3) SensorML [21]. SensorML falls into the category of sensor description rather than just data description. We should stress that these three standards are just three among

a larger community of standards, e.g. KeelyBricks [22], MBARI [23] , MIML [24], ESML [25] , and OBIS [26]. The common data index (CDI) is designed to be used as an index to the individual datasets held by sea-search partners and as such could be described as a description of data sets [19]. This metadata about data sets has been implemented in an XML format known as the CDI XML schema. . The CDI XML format has adopted the ISO19115 metadata standard which is an ISO standard that defines the schemas for describing geographic information and services [19]. The CDI is supposed to provide enough information to answer the following questions: (1) Where? - What is the geographical location of the captured data? (2) When? - When the observation began and when did it end? What was the sampling interval? (3) What? - What was measured? (4) How? - What instruments were used? What platforms were involved? (5) Who? - Who is the originator of the data? (6) Where to find data? - Which partner holds the data? Is there web access? Are there restrictions?

MarineXML gives a common framework for the data and its structure in terms of a catalogue of feature types but does so by largely using the Climate Science Markup Language (CSML) to tag data. In fact, it is probably more correct to say that MarineXML is a framework for allow interoperability of marine data. The framework requires the implementation of a common vocabulary for measurement systems by use of parameter dictionaries for storing agreed definitions of phenomena and the units used to measure them. This is the same solution that the CDI uses for instruments and the respective measurements. However, MarineXML adapts the unit and phenomena dictionary definitions inherent in the Climate Science Markup Language (CSML) in applying a GML encoding of CFStandardNames for referencing phenomena dictionaries and UDUnits for unit definitions. Essentially, phenomena (things that can be sensed/measured) are measured in terms of units. GML has a dictionary of phenomena and associated units (of measurement). Whereas CDI schema XML is based on ISO19115 and MarineXML is based on ISO 19136 both these ISO standards fall into the general ISO 19100 category of geographic standards. Thus we use the term "MarineXML/CSML" to refer to the data standard rather than the whole data interoperability framework of MarineXML.

As with the CDI XML schema there is the conception that a common framework and grammar for expressing the data and its structure is needed and that this also necessitates a common vocabulary of measurement systems and feature types. But abstractly all marine data in general should have the following attributes: (1) Position: all data will have associated positional information; (2) Time; (3) Units; (4) Tolerances: accuracy, precision, resolution; (5) Source; (6) Agent: what person / organisation carried out the data recording? (7) Method: method by which the data was obtained, (8) Promoter: entity that initiates data collection, e.g. a government agency; (9) Original purpose: what was the original purpose of the data collection? (10) Restrictions: are there copyright restrictions etc? (11) Errors; (12) Quality control; (13) Form; (14) Format; (15) Metadata.

SensorML is an XML based description of the process or processes of measurement that a sensor or sensor systems performs. Processes are entities that take one or more inputs and through the application of well-defined methods using specific parameters, results in one or more outputs. In addition there is a large amount of metadata related to the sensor with respect to system location, capabilities,

characteristics, contacts, time constraints, legal constraints, security constraints amongst others.

3.1 The problem of granularity and metadata standards

One problem which reoccurs in the creation of SSW systems is - what metadata is to be used and how specific should it be? However, a second problem arises when organizations use different types of metadata and this problem is a question of semantics in the broadest sense – what do the metadata terms actually mean? It may be the case that different groups may represent the same data in different ways or use codes that have different levels of granularity. The SSW system has to be able to deal with all these problems. In Table 1 are listed two parameters which list the GF3 codes. In this example the code “DRYT” refers to “Dry bulb temperature”. In Table 2 we list a number of BODC codes related to “air temperature” using a “dry bulb thermometer”. As can be seen the BODC codes list the instrument used, in this case a “dry bulb thermometer”, as well as what is being measured. The BODC also appears to have more entries for “air temperature” using a “dry bulb thermometer” than GF3. The BODC in this case is more detailed and thus has a higher level of granularity and different levels pose a problem which data with different tags have to be integrated. Does “DRYT” map onto all three BODC codes or is it just equivalent to one of the BODC codes? This type of question must be answered by developers of the SSW system but also more importantly is must be answered by the community of users. The problem of granularity is a general problem for ontology-builders and those who wish to map ontologies. This problem also falls under challenge (2), the heterogeneity of data sources, when the SSW uses descriptions to aid in classifying data sources the choice of appropriate metadata standards is fundamental.

4.1 Chemo-sensors & the semantic sensor web

Chemo/bio-sensor Networks employ emerging molecular sensing technologies in order to monitor specific targets in the environment, and in some cases develop linked proxies for predictive use. The Adaptive Sensors Group (based in Dublin City University) have developed a ground-based sensing device, in this case, an autonomous phosphate analyzer [27]. This is a field-deployable system for long-term monitoring of phosphate levels in natural waters was developed incorporating sampling, pumping, reagent and waste storage, optical detection, and wireless communication in a robust and portable device. The analyzer is more complex in design than common sensors such as thermistors and passive infrared sensors as it uses “wet chemistry” to analyse phosphate which involves pumps, valves and fluid handling, the use of reagents and storage of waste.

The phosphate monitoring analyser is designed to operate autonomously in long-term field deployments. Figure 2 depicts trial results from a waste water treatment plant obtained during a 30+ day trial. A trial of this length is a major achievement as many chemo-sensors operate over days rather than weeks to months. Comparable data

were obtained by the plant's monitoring system suggesting that the wireless phosphate analyser can produce reliable data and is sufficiently robust to be operated in a completely autonomous manner for at least seven weeks.

4.1 Analytics: pushing intelligence to the edge of the semantic sensor web

SmartBay is a program of national infrastructure investment with the aim of enabling the development of next generation advanced coastal and marine monitoring and management technologies [28]. The Marine Institute (Ireland) and IBM are engaged in a multiyear collaboration to develop and provide advanced capabilities for global water management solutions. This collaboration is multifaceted but two major outcomes have been the development of an advanced embedded sensor platforms and the development SmartBay information portal. The advanced embedded sensor platforms are based on hardware that is ultra low power and embedded software builds IBM technology (e.g. J9 JVM, Lotus Expeditor components, MQ Microbroker). This platform has been developed to push intelligence down to the sensor level where real-time decision making can take place.

The autonomous phosphate analyzer [27] was used as a testbed for the advanced embedded sensor platforms. The core idea here is to push intelligence onto the sensor. The new system has the following capabilities: (1) Self monitoring - the system is able to monitor its consumables and change sample rates in response to outside events; (2) Scalability - the command and control can be updated remotely and multiple units can be updated simultaneously; (3) Verifiable data transmission - MQTT is used to ensure data is delivered.

The concentration of phosphate in the treatment plant is affected by the local ambient weather conditions and so it is desirable to vary the sampling rate as local weather conditions change. Rainfall can result in increases in phosphate levels in water bodies due to increased run off from agricultural/forestry land where manure/fertilizers are used. Heavy rainfall or storm events can also lead to large increases in the flow in a river which can also increase phosphate levels. The inclusion of satellite meteorological data, which is usually available directly on the web or can be acquired from the local meteorological institute, can give a more complete picture of the reasons behind the changes of the pollutants measured. This is exemplified by the causal link between deterioration in water treatment effectiveness and the waste volume throughput. A major increase in water volume input for example due to heavy rainfall in the local catchment, may overwhelm a plant's capacity and lead to a deterioration in the treated water quality. For a chemo-sensor such as the phosphate system on board analytics may identify events based on local changes in phosphate level but data from contextual sources are required to provide the full picture. The seamlessly joining of plug-n'-play sensors into the SSW requires the system to be able to handle (1) the sensor and (2) contextual information. Can the SSW provide a sensor with contextual information that can allow for predictive modeling?

Recent developments in wireless sensor node technologies have resulted in devices with increased CPU, memory and transmission capabilities. Such developments have lead to the possibility of deploying goal based reasoners onto the leaf nodes of the

network to engage in real time, in-situ and intelligent decision making. Given the remote operation, potential latency in message transmission and data volume, such abilities may be crucial to the successful operation of the sensing system. Each entity resident on a node is termed an agent and there may be multiple agents on a single node. An example of one such system is AgentFactory Micro Edition (AFME) [29]. AFME has been successfully deployed on a wide range of devices with varying capabilities. For example, it has been used on SunSPOTs to provide adaptive sensing capabilities. AFME has also been deployed to the SmartBay Phosphate Monitoring system. Using AFME provides a common programming model for the wide range of sensor devices that may possibly be deployed to compliment the core system. As mentioned previously, it also provides in network decision making so for example, decisions based on trade-offs between system accuracy and power consumption can be taken without human intervention in the field. The degree of cleaning of the device will impact power consumption, as well as sample quality. In some cases it may be vital to have a very precise reading when, for example, no other sensors are within the locality. However, when numerous other sensors are also participating, then minor inaccuracies may be tolerated as they can be averaged out by using a combination of all sensor readings. Further standard energy saving decisions such as adaptive transmission and sampling frequencies can also be taken by the agents.

In addition to network based decisions, the agents can also provide some analytic mechanism to signify important trends in the data. For instance, if a phosphate level is breached as in Figure 2, the agent might decide to notify a local or government authority. Such thresholding and event detection can be disseminated to the agents in a similar way to the policy level considerations such as prioritizing power consumption discussed previously. The thresholds may be automatically adjusted on a daily or even hourly basis depending on the cumulative levels detected over a given period of time. A code snippet from AFME which would classify three high phosphate events in the Figure 2 but more importantly a series of actions can occur from this ongoing event detection is given below:

```
newThreshold(?t) > setThreshold(?t)
threshold(?x), reading(?y) > checkReading(?x, ?y);
thresholdBreached(?amount) > informUserAgent(?amount);
severeThresholdBreached(?amount), strictPolicy() > informPlantAgent(?amount);
```

Further code from AFME could also detect sensor drift, diagnose operational issues and identify further user-defined events and integrate outside data sources.

However, Figure 2 also highlights the challenges to the SSW. The reference sensor (in red) does not identify the first event that is flagged by the prototype phosphate sensor. Which raises the question - is this event real or is it a false positive? Many sensors will need access to contextual information and have sophisticated on-board intelligence to assist in the process of deciding whether detected events are true or false. The SSW should allow sensors to discover, access, and process relevant contextual information - even sophisticated instruments such as the prototype phosphate sensor can benefit from contextual information that improves the quality of event detection. The contextual information should also be quality tagged (e.g. via metadata) to identify whether it should be used or not by other nodes in the SSW and this quality checking may further require access to other sets of related contextual

information; which highlights the challenge of quantifying data quality. It will also be the case that false negatives can occur (what events have been missed?) and so as the numbers of devices scale up, the complexity of decision-making also scales up. However, in both cases (false positives and false negatives), the quality of event detection, and dependent decision making, can be improved. In the case of false positives, the confidence in a positive decision is enhanced through, for example for the phosphate sensor, correlation of sampling rate with rainfall level; i.e. water quality decreases when there is a heavy rainfall event in the local catchment; therefore increase sampling rate to get more independent measurements for cross-validation. On the other hand, if an event is predicted from contextual information but not detected (possible false negative), the instrument could be instructed to check the data using more sophisticated algorithms to see if there is any evidence of an event.

4.2 Problems facing SSW from the coalface of analytics and chemo-sensors

To summarise our discussion of analytics and chemo-sensors, it is clear that more sophisticated decision making tools are needed to ensure that the incidence of false positives and false negatives is minimized. If this is not done then the usefulness of the aggregated information will be unacceptably compromised, and WSN effectively useless. In short, decision-making tools are required to if we are to achieve workable, functioning internet-scale sensing. This problem falls under challenge (4) where analytics may occur across the whole SSW and challenge (3) where contextual information will have to be accessed from different data streams (and different data sources and modalities).

5 Conclusions

Currently millions of sensors are being deployed in sensor networks across the world. These networks generate vast quantities of heterogeneous data across various levels of spatial and temporal granularity. The semantic sensor web will handle sensor data ranging from networks to single-point in-situ sensing to remote sensing which can cover the globe. This will result in the unification of the web with the real-world. In this position paper, we discussed the major challenges to this unification from the perspective of sensor developers (especially chemo-sensors) and integrating sensors data in real-world deployments. These challenges are:

1. The heterogeneity of data sources and data transport methods that all must neatly fit into the SSW.
2. Identifying the quality of the data.
3. Integrating data streams from different sources and modalities (esp. contextual information).
4. Analytics (e.g. decision making) may occur across the SSW.

These challenges were discussed in relation to current metadata standards, integrating data sources in the marine environment and in relation to a chemical analyzer. These challenges cannot be dealt with separately as we have seen in the marine environment that the heterogeneity of data sources makes integrating data streams from different sources and modalities extremely difficult, and makes analytics based on contextual information problematic. The identification data quality will also rely on contextual information that is difficult to automatically process given the heterogeneity of data sources. Thus heterogeneity of data sources (and data transport methods) is the core challenge but the other challenges must be dealt with for the SSW to offer a fully scaleable, integrated solution to environmental monitoring.

Acknowledgements

We gratefully acknowledge the financial support from Science Foundation Ireland under the CLARITY CSET award (07/CE/I1147), the Marine Institute for the Beaufort Marine Research Award in Environmental Sensing (BEAU-SENS-10) and Enterprise Ireland Technology Development Award (No: TD/08/111).

References

1. Manfred Hauswirth and Stefan Decker, "Semantic Reality - Connecting the Real and the Virtual World," Microsoft SemGrail Workshop, Redmond, Washington, June 21-22, 2007.
2. R. Ambrosio, "Internet-Scale Data Acquisition and Control Systems — Programming Paradigm Challenges", Paper presented at the conference, Creating An Expanded DER Industry, November 28–30, Loews L'Enfant Plaza Hotel, Washington, DC. (2001).
3. D. Diamond, "Internet-scale sensing", *Anal Chem.*, 15, 278A-286A (2004)
4. Amit Sheth, Cory Henson, and Satya Sahoo, "Semantic Sensor Web," *IEEE Internet Computing*, July/August 2008, p.78-83.
5. Sensor Model Language (SensorML) for In-situ and Remote Sensors portal.openeospatial.org/files/?artifact_id=11516
6. D. Diamond, S. Coyle, S. Scarmagnani, and J. Hayes, "Wireless Sensor Networks and Chemo-/Biosensing", *Chem. Rev.*, 108, 2, 2008, pp. 652-679
7. Sonia Ramirez-Garcia and Dermot Diamond. Internet-scale Sensing: Are Biomimetic Approaches the Answer?, *Journal of Intelligent Material Systems and Structures*, 18 (2) (2007) 159-164.
8. G. M. P. O'Hare, D. Diamond, K. T. Lau, J. Hayes, C. Muldoon, M. J. O'Grady, R. Tynan, G. Rancourt, H. R. Kolar and R. J. McCarthy, *IBM Journal of Research and Development* (2009), submitted for publication.
9. O'Reilly, T.C., et al., 2001: "Smart Network" infrastructure for the MBARI Ocean Observing System, *Proceedings of the Oceans 2001 MTS/IEEE Conf.*, Honolulu, Hawaii, November 5-8, 2001.
10. P. Alexander and R. Holman. Quantitative analysis of nearshore morphological variability based on video imaging. *Marine Geology*, 208(1):101-111, 2004.
11. Christopher W. Brown, Laurence N. Connor, John L. Lillibridge, Nicholas R. Nalli and Richard V. Legeckis. *Remote Sensing of Coastal Aquatic Environments*, Chapter 2, An introduction to satellite sensors, observations and techniques, 21-49. Springer, 2007.
12. Ocean Surface Topography Mission/Jason-2, http://www.nasa.gov/mission_pages/ostm/overview/index.html

13. MERIS Product Handbook, <http://envisat.esa.int/handbooks/meris/>
14. GODAE – Global Ocean Data Assimilation Experiment, <http://www.godae.org/>
15. GHRSSST – Group for High Resolution Sea Surface Temperature, <http://www.ghrsst-pp.org/index.htm>
16. J. S. Myers and R. L. Miller. Remote Sensing of Coastal Aquatic Environments, Chapter 3, Optical Airborne Remote Sensing, 51-66. Springer, 2007.
17. M. Davidson, M. V. Koningsveld, A. de Kruijff, J. Rawson, R. Holman, A. Lamberti, R. Medina, A. Kroon, and S. Aarninkhof. The coastview project: Developing video-derived coastal state indicators in support of coastal zone management. Coastal Engineering, 54(6-7):463-475, 2007.
18. E. O'Connor, A. F. Smeaton, N. E. O'Connor, and D. Diamond. Integrating multiple sensor modalities for environmental monitoring of marine locations. In SenSys '08: Proceedings of the 6th ACM conference on Embedded network sensor systems, pages 405-406, New York, NY, USA, 2008. ACM.
19. Common Data Index (CDI) - Metadata Format and full description of XML schema -version 2.04. http://www.sea-search.net/cdi_documentation/
20. MarineXML, <http://www.iode.org/marinexml/>
21. Tutorial 1: Using SensorML to describe a Complete Weather Station (2006). <http://vast.uah.edu/SensorML/tutorial/SensorML%20Tutorial%201%20-%20Weather%20Station%20System.pdf>
22. Keeley, R., Isenor, A., Linguanti, J. (2003), XML Bricks., <http://ioc.unesco.org/marinexml/contents.php?id=19>
23. Monterey Bay Aquarium Research Institute, <http://www.mbari.org/ssds/ReferenceDocuments/MOOSMetadataSchema.xsd>
24. Marine Information Mark-up Language, <http://www.rdc.uscg.gov/iws/pubs/miml.pdf>
25. Earth Science Mark-Up Language, <http://esml.itsc.uah.edu/index.jsp>
26. Ocean Biogeographic Information System, <http://iobis.org/obis/obis.xsd>
27. C.M. McGraw, S.E. Stitzel, J. Cleary, C. Slater and D. Diamond. Autonomous microfluidic system for phosphate detection, Talanta 71 (2007) 1180–1185.
28. SmartBay, <http://www.marine.ie/home/services/operational/SmartBay/>
29. Muldoon, C., O'Hare, G.M.P., O'Grady, M., Tynan, R., Agent Migration and Communication in WSNs, 1st International Workshop on Sensor Networks and Ambient Intelligence, December 1-4, 2008, Dunedin, New Zealand.

Table 1. GF3 codes for a number of parameters captured by M3A buoys.

Parameter	Unit	GF3 codes
Air Temperature	Celsius	DRYT
Wind Speed	m/sec	WSPD
...

Table 2. BODC codes for 'Air temperature' using a dry bulb thermometer.

Parameter	Description	BODC codes
Air Temperature	"AirTemp":Temperature of the atmosphere by dry bulb thermometer	CDTBSS01
Air Temperature	"AirTemp":Temperature of the atmosphere by dry bulb thermometer	CDTASS02
Air Temperature	"AirTemp":Temperature of the atmosphere by dry bulb thermometer	CDTASS03

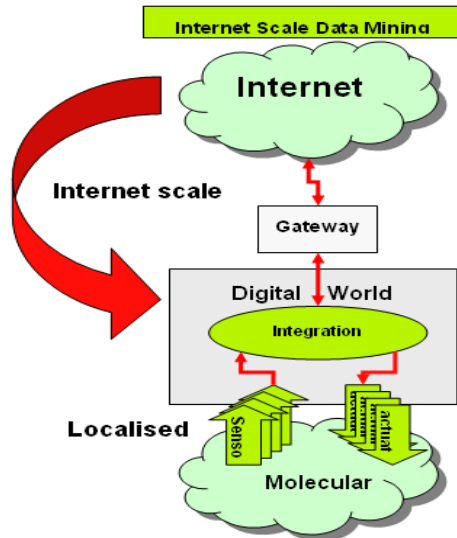


Fig. 1. Establish the chain. All analytical measurements must be linked to realize the concept of Internet-scale sensing. Localized control of important parameters is maintained, but the information is shared with external users via the Internet.

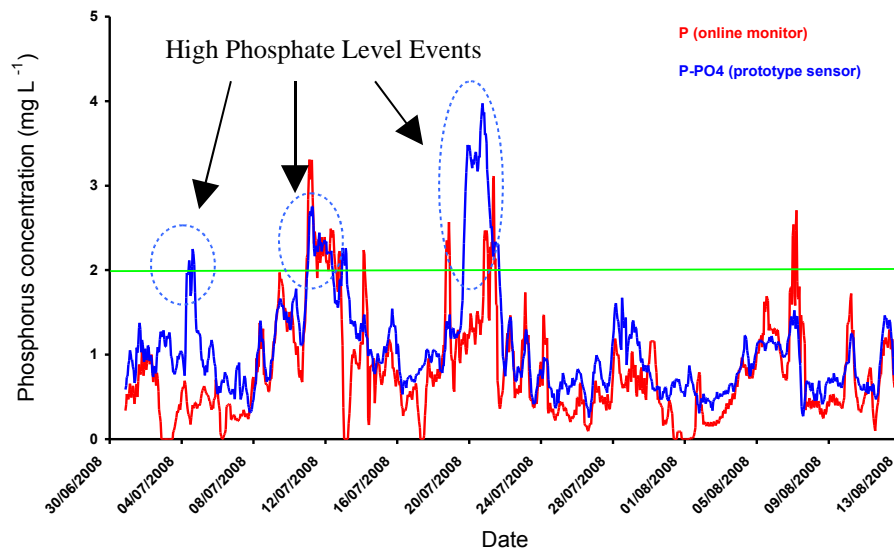


Fig. 2. Chemo-sensor in action: trial results. The phosphate levels of a waste water treatment plant obtained with a prototype analyzer during a 30+ day trial are shown. Comparable data were obtained by the plant's monitoring system for reference values (on-line monitor). The AgentFactory Micro Edition operating on the phosphate analyzer detects three high phosphate events based on the prototype analyzer data using the indicated threshold, which can set off a chain of remedial action. However, the first event is possibly a false positive as the reference system does not indicate high levels. Furthermore, high levels of phosphate are indicated towards the end of the trial by the reference monitor, but not by the prototype system, which is potentially a false negative.