Big Data Metadata Management in Smart Grids

Trinh Hoang Nguyen, Vimala Nunavath, and Andreas Prinz

Abstract Smart home, smart grids, smart museum, smart cities, etc. are making the vision for living in smart environments come true. These smart environments are built based upon the Internet of Things paradigm where many devices and applications are involved. In these environments, data are collected from various sources in diverse formats. The data are then processed by different intelligent systems with the purpose of providing efficient system planning, power delivery, and customer operations. Even though there are known technologies for most of these smart environments, putting them together to make intelligent and context-aware systems is not an easy task. The reason is that there are semantic inconsistencies between applications and systems. These inconsistencies can be solved by using metadata. This chapter presents management of big data metadata in smart grids. Three important issues in managing and solutions to overcome them are discussed. As a part of future grids, some concrete examples from the offshore wind energy are used to demonstrate the solutions.

1 Introduction

Advanced technologies are making the vision for living in smart environments become realistic. Recently, several concepts within the smart environments have been introduced, such as smart home, smart transport, smart grids, smart museum, and

T.H. Nguyen (✉)· V. Nunavath · A. Prinz
Department of Information and Communication Technology, University of Agder, Norway
e-mail: trinh.h.nguyen@uia.no

V. Nunavath
e-mail: vimala.nunavath@uia.no

A. Prinz
e-mail: andreas.prinz@uia.no

smart cities. These smart environments are built based upon the Internet of Things (IoT) paradigm where lots of devices, sensors, appliances are connected through the Internet. These devices produce vast amounts of data, thus making the management of data a highly challenging task. Another common feature and an important problem of these smart environments is that each of them involves data modeling, information analysis, integration and optimization of large amounts of data coming from various smart appliances in diverse formats. The data are then processed by different intelligent systems with the purpose of providing efficient system planning, power delivery, and customer operations. Even though there are known technologies for developing most of these smart environments, putting them together to make intelligent and context-aware systems is not an easy task. The reason is that there are semantic inconsistencies between applications and systems. These inconsistencies can be solved by using metadata.

Typically, data are a collection of raw and unorganized symbols that represent real-world states. The information is the processed, organized, and structured data according to a given context [2, 60]. The context of related data and processes will decide the role as information of the captured data. Principally, information is the structured data with semantics. For example, if data are used for documentation or analysis, the data become information. Without metadata, the data cannot easily become information and incomplete or inaccurate metadata or too much metadata can cause misinterpretation of data [55]. Metadata should be therefore managed in a way that data can be easily interpreted and transformed to information.

Metadata management is a key to make data integration successful [25]. It has to be taken into consideration in the development of systems since it helps in making the systems scalable. For formal metadata management, semantic technologies have been developed. Ontology, which is a part of semantic technologies, plays a significant role in managing metadata of a domain. Ontologies can be used to support data integration in terms of facilitating knowledge sharing and data exchange between participants in a domain. In ontologies, concepts, properties, relations, functions, constraints, and axioms of a particular domain are explicitly defined [19]. We use semantic technologies to exploit the semantics of data, and hence ease metadata handling in smart environments.

In this chapter, we discuss how to manage big data metadata in smart grids with a particular focus on (1) knowledge sharing and data exchange, (2) derived data from relations between concepts, and (3) data quality as metadata. We will present a developed ontology model for offshore wind energy metadata management as an example of domain concept descriptions. IEEE P2030 points out that ontology might be a good option to create formal representation of real-world systems or objects composing these systems within smart grids [1]. As the number of devices is increasing tremendously, and many of them will be used in smart environments, it is important to make sure that any future system is scalable enough to keep pace with the technologies. Metadata models, as a backbone of any system, also need to be considered thoughtfully. The models need to be developed so that the following requirements are fulfilled.

- The models need to be compatible with existing data resources and future applications.
- Minimum effort is used to modify the models when integrating new devices.
- New devices' metadata are described in a way that discovery and access to them are easy.
- It must provide a guide to structuring, sharing, storing, and representing the big data in smart grids.
- The semantics of data needs to be exploited and clearly defined.
- Since it is not feasible to attach metadata with individual data, the metadata models must be related to data sources.

The rest of the chapter is organized as follows. Section 2 gives some background information about the areas that we discuss in this work. Section 3 presents some challenges of big data metadata management that we attempt to tackle. Section 4 describes solutions and approaches to overcoming the challenges. Section 5 discusses our solutions and gives some remarks on future work. Finally, section 6 concludes the chapter.

2 Background

This section describes the background of metadata, semantic technologies, IoT and smart grids. The relations between these areas are also highlighted.

2.1 Metadata

The term "metadata" was first introduced in 1968 by Philip R. Bagley to refer to descriptive data that provided information about other data in a database environment [51]. In different contexts, the term metadata is interpreted in different ways, for example, metadata are data about data; or metadata are machine-readable information about electronic resources or other things; or metadata are structured information that describes, explains, locates an information resource [54]. Basically, metadata are descriptors that describe a way of identifying information. Data without metadata result in blind decision making [55]. In other words, without metadata, data have no identifiable meaning. For instance, when a user searches for information, he will receive a list of search results from a search engine. The search engine looks up for requested information from huge amounts of data based on search terms, tagging content, and other metadata associated with data. Metadata provide the necessary documentation for users by answering who, what, when, where, why, and how questions upon the users' requests.

Metadata put data into a context so that the data can be understood by users and become information. Besides the general role as descriptors, metadata can be used for:

- information classification - information is classified into different categories based on content, purpose, location, area, etc;
- information discovery - a large amount of time is used to look for things, and many of them cannot be found due to the lack of descriptions. Metadata therefore enhance information discovery and knowledge sharing;
- information interpretation - a poor description of data may lead to wrong decision making or business loss due to wrong interpretation of the data;
- data integration - when we integrate data from various sources in different formats and platforms, metadata are the only option that can make a foundation for data integration [55];
- device discovery - based on metadata of devices such as location, type, and other features devices can be discovered either automatically or semi-automatically by a system.

2.2 Big Data Metadata Management

Big data is characterized with volume, variety and velocity [61]. Volume is considered as a huge amount of data which can hold terabytes to petabytes of data which come from different devices, applications, and systems. Velocity is the speed at which the data comes in, and variety means many data types and data formats. Structured, semi-structured and unstructured data are involved in big data [15]. Data often come from machines, sensors, social networks such as Facebook, Tweets, smart phones and other cell phones, GPS devices and other sources making it complex to manage [45]. According to a report from McKinsey Global Institute, every year, over 30 billion original documents with data are created. 85% of the data will never be retrieved, 50% of the data is duplicates, and 60% of stored documents are obsolete. $1 and $10 are the costs to create a document and to manage it, respectively [31]. As the amount of data increases, the cost of management also increases. It is important to describe and manage metadata so that only important and necessary data are stored and provided to users when requested. Since data are used for making decisions by different applications and systems, the quality of data is one of concerns.

Not all of the data captured from sensors or devices are useful, only a part of the data is. Data are transformed to information only if the data are used for particular purposes, e.g., modeling, documentation. Part of the information will become knowledge in terms of abstraction and perception. Users are not interested in information (numbers), they are interested in knowledge, i.e., what can be derived from the information. For example, if a user wants to know about the temperature in a wind turbine hub, he will probably not expect to get a number or set of numbers as a response, but he will probably want to get either “Normal”, “Cold”, or “Hot”. Eventually, only part of the knowledge will be transformed to wisdom if the knowledge is used to serve some actionable intelligence [46]. Every step of the transformation involves management of data, information, and knowledge. Management of big data

metadata concerns a way to manage big data metadata such that metadata are good enough to enable knowledge extraction from big data.

2.3 Smart Grids and Internet of Things

Smart grids are the future generation of power grids where the energy is managed in a way that both consumers and energy producers will get more benefits from the grid in terms of reduction of expenditure on energy and reduction of carbon emissions. Indeed, it enables consumers to utilize lower tariff charges during off-peak periods and energy producers to react efficiently during peak periods. Smart grids are also used to effectively response to the fluctuations of renewable sources such as wind and solar when they are integrated in a power grid.

A smart grid is an electricity network that efficiently delivers sustainable, economic, and secure electricity supplies by intelligently integrating the actions of all users connected to it, including generators, consumers and those that do both [16]. On the consumer side, smart grids involve many smart meters and smart appliances, for example, smart washing machines, and dishwashers. The number of smart appliances is increasing dramatically. These devices are connected directly to the Internet. A large amount of sensors are used in these devices to make sure that every single change can be detected, managed and controlled. On the energy provider side, intelligent applications are used to maintain balance between demand and supply. Smart grids will bring the decision making gradually from a centralized level to local and finally to automatic.

In order to make a grid become smart, different technologies and applications are involved, e.g., advanced metering infrastructure (AMI), distribution management system (DMS), geographic information system (GIS), outage management systems (OMSs), intelligent electronics devices (IEDs), wide-area measurement systems (WAMS), and energy management systems (EMSs) [12]. These systems are driven effectively by IoT [56].

In IoT, things are connected in such a way that machines and applications can understand our surrounding environments better and therefore make intelligent decisions and respond to the dynamics of the environments effectively [6]. These things communicate to each other over the Internet. Advantages of IoT will contribute a lot to the effort of making smart grids in terms of real-time monitoring and control. Smart grid applications require quick response time no matter how big the data are. One example of such a system is an energy trading system which allows energy consumers or third parties to bid for energy prices in advance [13].

Due to characteristics of smart grids, a number of challenges are encompassed with the development of smart grids such as support heterogeneous participants, flexible data schema (e.g., add new or remove old appliances), complex event processing, privacy and security [57]. Thus, data from IoT alone are not enough. The data must be used together with the domain knowledge, machine interpretable metadata, services, etc. to become useful.

Figure 1 illustrates a conceptual model for smart grid communication with a focus on offshore wind as an energy generator. The model is based on the *Smart Grid Interoperability Panel* promoted by the National Institute of Standards and Technology (NIST) [38].

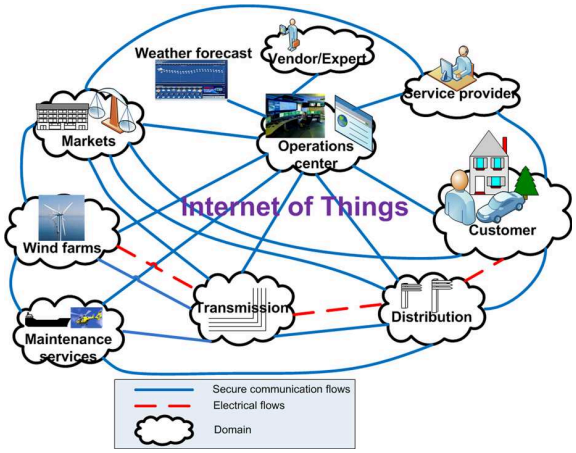

Fig. 1 An example of conceptual model for smart grid communication

Each domain is a high-level grouping of organizations, individuals, and systems of the offshore wind industry. Communication between stakeholders in the same domain may have similar characteristics and requirements. The communication flows are bidirectional. In this model, smart meters, smart appliances are installed at households, sensors are embedded on wind turbines, and intelligent programs are used at operations center.

Metadata are significant in the smart grid context. It is needed for organizing and interpreting data coming from energy market, service providers, customers, power grid, and power generators. Managing metadata in such a varied environment is a challenging task.

2.4 Semantic Technologies

Semantic technologies have been developed to make metadata understandable by a machine. Ontology is a part of semantic technologies that plays a significant role in managing metadata of a domain. There are several ontology languages such as SHOE, OIL, DAML-ONT, DAML+OIL, and OWL [30, 21]. Web Ontology Language (OWL), a language proposed by World Wide Web Consortium (W3C) Web Ontology Working Group, is being used intensively by research communities as well as industries. Ontologies can be represented by using Resource Description Framework (RDF)/RDFS (RDF Schema). However, a number of other features are missing in RDFS such as cardinality restrictions, logical combinations (intersec-

tions, unions or complements), and disjointness of classes. Let us examine some concrete cases within the offshore wind energy. The first case is that in RDF, we cannot state that *HydraulicSystem* and *HeatingSystem* are disjoint classes. The second case concerns the lack of cardinality restrictions, e.g., the fact that a wind power plant (WPP) can have more than one wind turbine converter component (WCNV) cannot be expressed in RDF, but it can be done in OWL using the following axiom $WPP \sqsubseteq (\geq 1\ hasWPPComponent.WCNV)$. OWL is an extension of RDFS, in the sense that OWL uses the RDF meaning of classes and properties [21, 8, 3]. The design of OWL was influenced by its predecessors DAML+OIL, the frames paradigm and RDF [21].

In OWL, Owl:Thing is a built-in most general class and is the class of all individuals. It is a superclass of all OWL classes. Classes are defined using owl:Class. A class defines a group of individuals that belong together. Individuals are also known as instances. Individuals can be referred to as being instances of classes. Note that the word concept is sometimes used in place of class. Classes are a concrete representation of concepts. Owl:Nothing is a built-in most specific class and is the class that has no instances. It is a subclass of all OWL classes. There are two types of properties in OWL ontology, they are object property and data type property. Properties in OWL are also known as roles in description logics and relations in Unified Modeling Language (UML). An object property relates individuals to other individuals (e.g., *hasWPPComponent* relates *WPP* to *WPP components*). An object property is defined as an instance of the built-in OWL class owl:ObjectProperty. A data type property relates individuals to data type values (e.g., *hasOilPressure*, *hasWindSpeed*). A datatype property is defined as an instance of the built-in OWL class owl:DatatypeProperty. A property in OWL can be transitive, functional, symmetric, or inverse.

OWL DL (DL stands for "Description Logic") is a variant of OWL. It was developed to support existing DL and to provide a possibility of working with reasoning systems. In this work, OWL DL is used to develop ontologies. The OWL DL semantics is very similar to the $\mathcal{SHOIN}^{(\mathcal{D})}$ Description Logic. It provides maximum expressiveness and it is decidable [21]. OWL DL abstract syntax and semantics can be found in [41].

2.5 Ontology Reasoning and Querying

A reasoner is a piece of software that is able to infer logical consequences from a set of asserted facts or axioms. It is used to ensure the quality of ontologies. It can be used to test whether concepts are non-contradictory and to derive implied relations. Reasoning with inconsistent ontologies may lead to erroneous conclusions [4]. There are some existing DL reasoners such as FaCT, FaCT++, RACER, DLP and Pellet. A reasoner has the following features: satisfiability, consistency, classification, and realization checking [49]. Given an assertional box $\mathcal{A}$ (ABox contains

assertions about individuals), we can reason w.r.t a terminological box $\mathcal{T}$ (TBox contains axioms about classes) about the following:

- Consistency checking: ensures that an ontology does not contain any contradictory facts. An ABox $\mathcal{A}$ is consistent with respect to $\mathcal{T}$ if there is an interpretation I which is a model of both $\mathcal{A}$ and $\mathcal{T}$.
- Concept satisfiability: checks if it is possible for a class to have any instances. Given a concept C and an instance a, check whether a belongs to C. $\mathcal{A} \models C(a)$ if every interpretation that satisfies $\mathcal{A}$ also satisfies $C(a)$.
- Classification: computes the subclass relations between all named classes to create the complete class hierarchy. Given a concept C, retrieve all the instances a which satisfy C.
- Realization: computes the direct types for each of the individuals. Given a set of concepts and an individual a, find the most specific concept(s) C (w.r.t. subsumption ordering) such that $\mathcal{A} \models C(a)$.

For relational database (RDB), Structured Query Language (SQL) is the query language of choice. But for ontologies, SPARQL and SQWRL (Semantic Query-Enhanced Web Rule Language) [39] are used to build queries. SPARQL is an RDF query language and SQWRL is a SWRL-based language for querying OWL ontologies. SPARQL extensions such as SPARQL-DL [48] and SPARQL-OWL [27] can be used as OWL query languages in many applications. But SPARQL cannot directly query entailments made using OWL constructs since it has no native understanding of OWL [39].

3 Challenges in Managing Big Data Metadata in Smart Grids

There are a number of challenges associated with management of big data metadata such as metadata quality, metadata provenance, semantics, and metadata alignment. In this section, we attempt to tackle three challenges in managing smart grids' big data metadata.

3.1 Knowledge Sharing and Information Exchange

In a diverse environment such as smart grids, meters, appliances, and applications are developed by different companies and vendors. Many of them use their own proprietary data formats, protocols, and platforms, thus data exchange is impeded. Using approved standards would contribute to solving such problems since they can make the data exchange unambiguous. The standards can be seen as a means of interoperability, a dictionary of data that can be used to manage, simplify, and optimize data models [10]. However, there are some problematic issues related to existing international standards for data exchange. For instance, it takes some years

to approve a standard internationally, but it seems that new technologies are proposed every year. As a result, novel concepts and terms are introduced, but they are not immediately described in these international standards.

The lack of widely accepted standards prevents the interoperability between smart devices, applications, smart meters, and renewable sources [47]. The Institute of Electrical and Electronics Engineers (IEEE), and NIST have recommended a list of standards that should be considered while developing smart grids [1, 38]. These standards have been developed by different working groups, leading to a lack of harmonizations. Although these standards describe different parts of smart grids, they share a common feature, i.e., the smart grid concepts. The question here is how to structure the domain concepts such that semantics is exploited effectively, knowledge sharing and data exchange are eased, and new concepts are updated in knowledge bases timely.

3.2 Relations between Concepts

Ontologies can be used to support data integration in terms of facilitating knowledge sharing and data exchange between participants in a domain. Ontologies describe the relations between concepts and their properties. These relations are metadata since relations can lead to computability of derived data. This opens several possible paths for calculation and gives users the possibility of selecting the most suitable one. However, there is a lack of a formal description of such relations in ontologies. One important question in managing metadata in ontologies is how to handle relations so that the selection of data (independent of type of data: base or derived data) can be done at runtime depending on the actual situation.

3.3 Data Quality

It is normal to use more than one sensor to measure, e.g., pressure or temperature at a particular point. The quality of each sensor is different from the others and depends on the conditions. In offshore wind energy, a couple of sensors are deployed on a windmill and they frequently measure and deliver the data to the users and applications by means of services. As sensors are prone to failures their results might be inaccurate, incomplete, and inconsistent [50]. Therefore, the data quality should be handled in such a way that users and applications can specify the desired quality level of the data. Only when the data source has the requested quality descriptions it would be used for further processing. One of the issues related to data quality is the handling of data quality at user level in enterprise applications where there is a potentially large number of data sources with quality information. Another issue is that sometimes none of the available data sources has the required quality information.

In this case, how a system should respond to such a request should be considered and a way to provide requested data to users should be investigated.

4 Solutions

This section presents three solutions and approaches to overcoming the challenges described in Sect. 3. Smart grids involve vast amounts of data from consumers, generators, billing, and management. Here, we use a case in which offshore wind energy plays a role as a renewable energy source generator to demonstrate our points.

4.1 Semantic-enhanced Concept Modeling

This section discusses the solution to the challenge described in Sect. 3.1. We look into how the semantic technologies can help us to solve the challenges with taking into consideration the requirements for the developed metadata models presented in Sect. 1.

4.1.1 The Information Model

An information model plays an important role in building a smart grid. It not only provides a common basis for understanding the general behavior of smart grid communication, but also facilitates the collaboration process between smart grid stakeholders due to shared concepts with a common semantics. An example of sharing common concepts between partners of offshore wind energy is illustrated in Fig. 2.

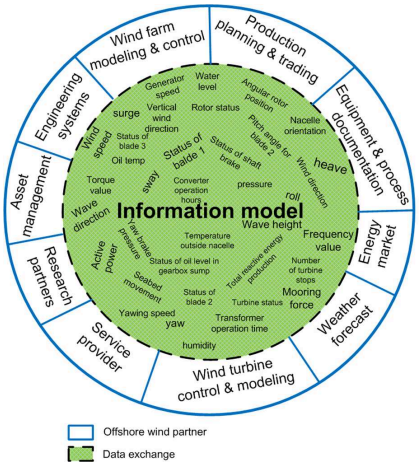

Fig. 2 Data pie chart for the offshore wind industry [36]

Availability and reliability of data are significant for any systems and partners. Offshore wind partners can efficiently perform their work using the available data. For example, wind speed information is the input to a wind speed prediction program. The output from the program can be used with the generator speed to predict the availability of wind power in the next few hours. In order to optimize wind farm efficiency, wind farm operations information regarding wind direction, active power, status of blades, etc. is needed. The weather forecast and energy market information is used to manage wind power production as well as maintenance for wind turbines (e.g., a wind turbine can be stopped when consumer demand is low). An information model is developed based on the IEC 61400-25 standard [24] to keep pace with the continual introduction of new technologies. More details about the information model can be found in [37].

4.1.2 An Offshore Wind Ontology

An information model represents the knowledge concerning specific domain communication. In particular, the purpose of creating an offshore wind information model is to facilitate the process of agreement on data exchange models as well as collaborations among offshore wind partners. We use the developed information model to build an offshore wind ontology (OWO) as depicted in Fig. 3. The idea of creating OWO from the terminologies is to share, reuse knowledge, and reason about behaviors across a domain and task. It is also a key instrument in developing the semantic web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [9]. An ontology helps to make an abstract model of a phenomenon by identifying the relevant concepts of that phenomenon [53].

Suppose several different sources/data storages contain wind turbine information. If these sources share and publish the same underlying ontology of the terms they all use, then computer agents can extract and aggregate information from these different sources. The agents can use this aggregated information to answer user queries or to provide input data to other applications. For example, a SQWRL query over OWO that is used to get oil pressure and pitch angle set point of the wind power plant which has ID is “2300249”, is expressed as follows:

$WF(?p) \wedge hasID(?p, \text{“2300249”}) \wedge hasWPPComponent(?p, ?comp) \wedge hasOilPressure(?comp, ?pres)$
$\wedge\ hasPitchAngleSetPoint(?comp, ?pitchAngle) \text{ -}\!> sqwrl : select(?p, ?pres, ?pitchAngle)$

4.1.3 Semantic Sensor Network Ontology

As the number of devices and appliances grows, the number of sensors embedded in such devices will also grow. Ontologies are an adequate way to model sensors and their capabilities [35]. Sensor metadata are used for selecting sensor sources

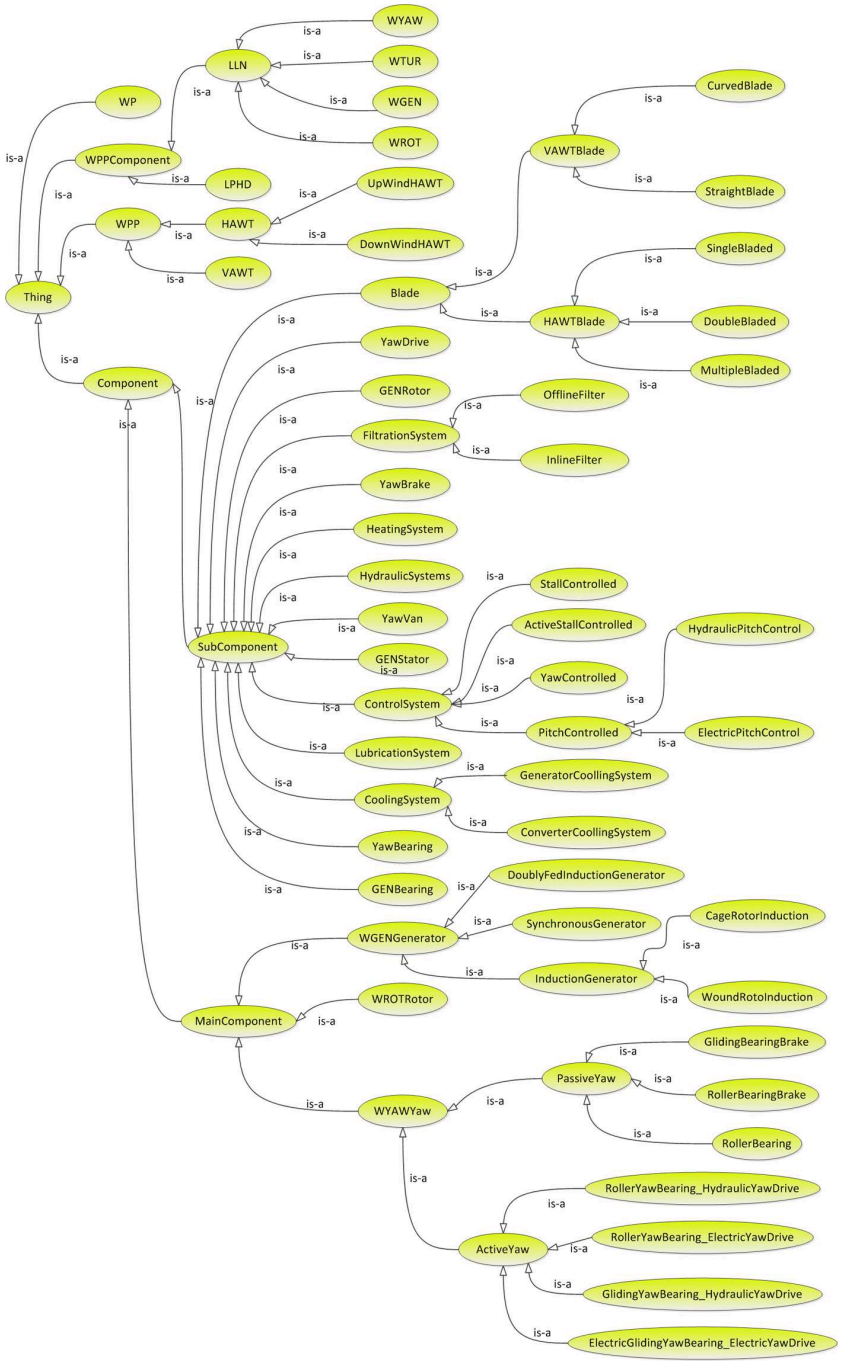

Fig. 3 OWO visualization

and for integrating with other data sources [28]. Thus sensor metadata are important and needs to be exploited. However, sensor metadata alone cannot make a grid become smart. These metadata must be associated with metadata from devices and appliances that are participated in the grid.

The W3C semantic sensor network incubator group has introduced a semantic sensor network (SSN) ontology[1] to describe sensors, observations, and measurements. The ontology describes sensors and their properties such as accuracy, precision, resolution, measurement range, and capabilities. The ontology includes models for describing changes or states in an environment that a sensor can detect and the resulting observation [14]. An example of the alignment of the SSN ontology to the developed OWO is depicted in Fig. 4.

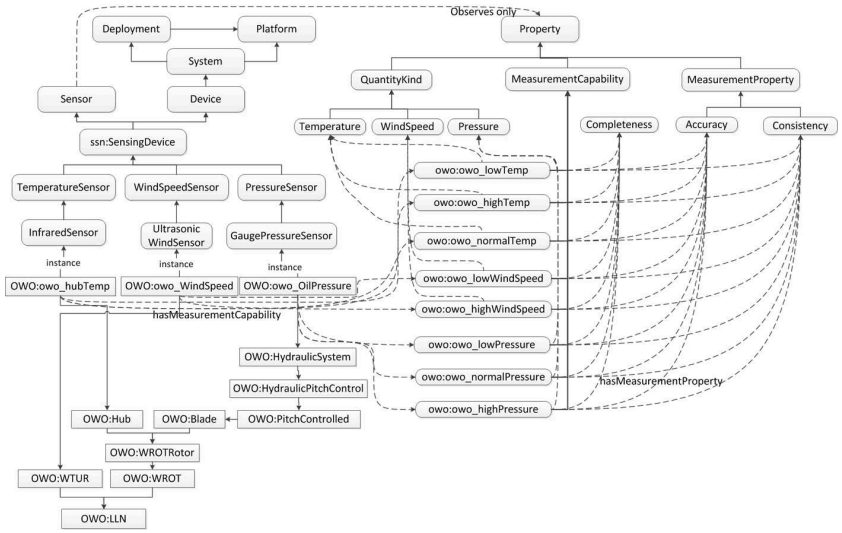

Fig. 4 An example of the alignment of the SSN ontology to OWO

The developed OWO can be connected to SSN to share common information such as measurement values from sensors embedded on a wind power plant. At the same time, OWO can still guarantee the complete description of a wind power plant data model. These two ontologies should be maintained separately since the number of concepts in these ontologies might grow as new technologies are introduced.

[1] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn

4.2 Relations between Concepts

Missing data can be caused by network disconnection, device faults, and software bugs. In some cases, where monitoring of devices or components is extremely important, a single missing value of a data point could lead to wrong predictions or damage of components. In the wind energy domain, many prediction and monitoring applications are employed, for example, power output prediction, wind turbine blade monitoring. The performance of these applications relies very much on data collected from the wind turbines. Missing of a single data item in the set of input data to these applications can make the applications produce wrong output or no output at all. In this case, the missing data item needs to be derived from other available data items. Derivation of data also plays a significant role in decision support systems [43]. For instance, in time-series data analysis, missing data that are located in the middle of a time-series have a high influence on the efficiency of algorithms that are used to reveal hidden temporal patterns such as vector autoregression and exponential smoothing [62]. This section describes a way to model possible paths to deriving missing data from relations between the concepts.

4.2.1 Derived Data Modeling

Data are classified into two categories: base data and derived data [20]. Base data are those data obtained from data sources. Derived data are those data obtained by combining or computing from base data. The combination and computation of base data are based on relations between domain concepts.

Derived data are described by derived classes and derived attributes. A derived attribute is an attribute that is derived from other attributes in the same class or from different classes that have relationships with the class that contains the attribute. If all attributes of a class are derived, the class is called derived class [5].

Derived data give an advantage of storing data since there is no need to store derived data in a database. Another advantage is that the structure of the data storage is undisclosed to users, derived attributes are accessed via user interface.

Guaranteeing the correctness of derived data is an important task because applications that use the data might produce wrong results if they receive insufficient input. Therefore, derived data need to be handled in such a way that its correctness is ensured. Formally modeling of derived data can help us to figure out different aspects of handling the data, and hence guaranteeing the correctness.

We use UML [18] to model the concepts in the wind domain. UML is based on object-oriented design concepts and is independent of any specific programming language. We also use Object Constraint Language (OCL) to express constraints in UML models [59]. OCL is a complement of UML. It makes models precise, consistent, and complete. In this work, we add OCL constraints to our models to tackle the derived issue mentioned in Sect. 3.2. We analyze two wind energy related cases where derived data play an significant role. We use the ontology introduced in Sect. 4.1.2 to demonstrate the cases.

4.2.2 Derived Data within One Concept

Temperature measurement can be presented in different units such as Fahrenheit (F) or Celsius (C). The relation between F and C is as follows.

$$F = \frac{9}{5} * C + 32 \tag{1}$$

or

$$C = \frac{5}{9} * (F - 32) \tag{2}$$

The derivation can be obtained during execution time, for example, the authors of [11] use SWRL to define the transformation between temperature measurement units. However, such an approach will limit the possibility of expressing complex equations. A better approach is to attach formulas directly to properties in ontologies such as [23]. Let us consider a simple ontology describing the wind turbine generator (WGEN) concept and temperature as one of its properties. Figure 5 illustrates a formal model of temperature conversion using UML and OCL.

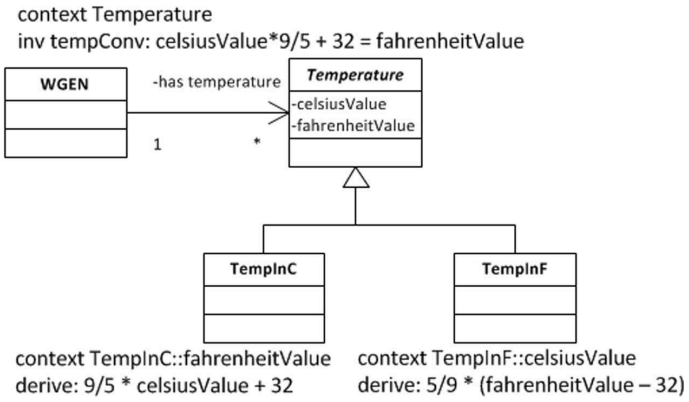

Fig. 5 Temperature conversion

WGEN denotes the wind turbine generator class as described in [24]. *Temperature* is an abstract class that contains two attributes: the *celsiusValue* and *fahrenheitValue*. The two classes *TempInC* and *TempInF* contain rules to convert temperature unit from C to F and from F to C, respectively.

4.2.3 Derived Data between Two Concepts

Let us consider an offshore wind farm scenario where many sensors are located on a wind turbine to capture information. What if one of them loses the connection? In-

formation related to that one will be lost. How can we utilize other devices to derive that information so that the monitoring of the wind turbine is still ensured? Figure 6 shows how to make use of derived data from two parameters within the wind domain. The basic mathematical relation between wind speed and power output is expressed in Eq. (3) [33].

$$P_{avail} = \frac{1}{2}\rho\pi r^2 v^3 C_p \tag{3}$$

where P_{avail} denotes the available power output (W), ρ denotes air density (kg/m^3), r denotes blade length (m), v is the wind speed (m/s), and C_p denotes the power coefficient. Please note that the power coefficient is not constant; it depends on other factors such as rotational speed of the turbine, pitch angle, and angle of attack [34].

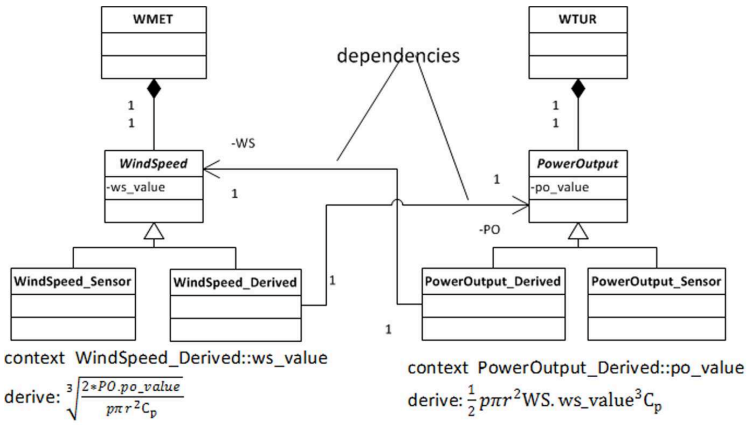

Fig. 6 Derivative relationships between two concepts

4.2.4 Derived Data with More than Two Concepts

What happens if one more parameter is added to the system? As an extension of the two concept model, we can have a model for three parameters as shown in Fig. 7.

Equation (3) can be rewritten as follows:

$$P_{avail} = \frac{1}{2}\rho\pi r^2 C_p \left(\frac{r}{TSR}\right)^3 \omega^3 \tag{4}$$

where TSR is tip speed ratio, ω (rpm) is the rotational speed of the blade. The TSR value can be obtained from the blade manufacturer, otherwise let TSR equal 7 since

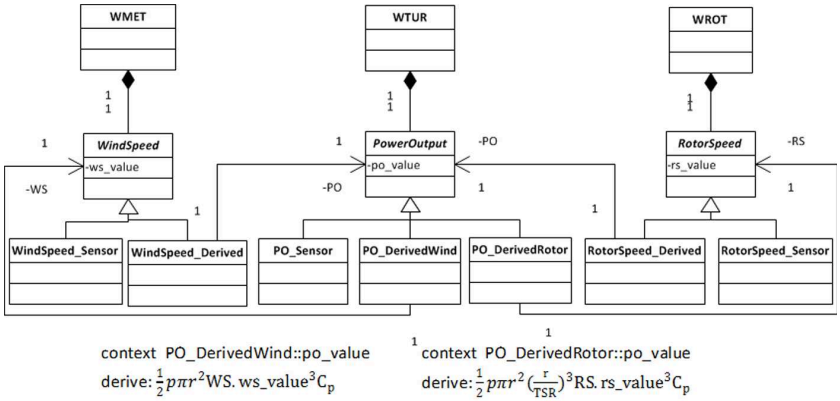

Fig. 7 Derivative relationships between three concepts

it is the most widely reported value in three bladed wind turbines [42]. We can then easily obtain *PO_DerivedRotor* as shown in Fig. 7.

A simple path, which is extracted from the model described in Fig. 7, is shown in Fig. 8a where *WindSpeed_Derived* can be derived from *PO_DerivedRotor* which can be derived from *RotorSpeed_Sensor*.

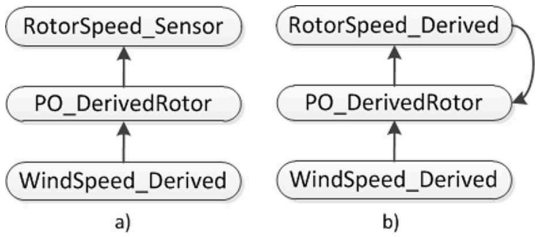

Fig. 8 WindSpeed is derived from PowerOutput and RotorSpeed

If we choose *RotorSpeed_Derived* instead of *RotorSpeed_Sensor*, this leads to a cyclic dependency as shown in Fig. 8b. Cyclic dependencies have to be avoided, as they cannot be computed.

Fig. 9 depicts a model which is the extension of the model illustrated in Fig 7. In order to solve the derivation cycle issue, the transitive closure of the dependency *dependsOn* should not be reflexive. The transitive closure of *dependsOn* is expressed in OCL as follows:

---

contextProperty

inv cycleRestriction : **not** $self.dependsOn.closure() -> include(self)$

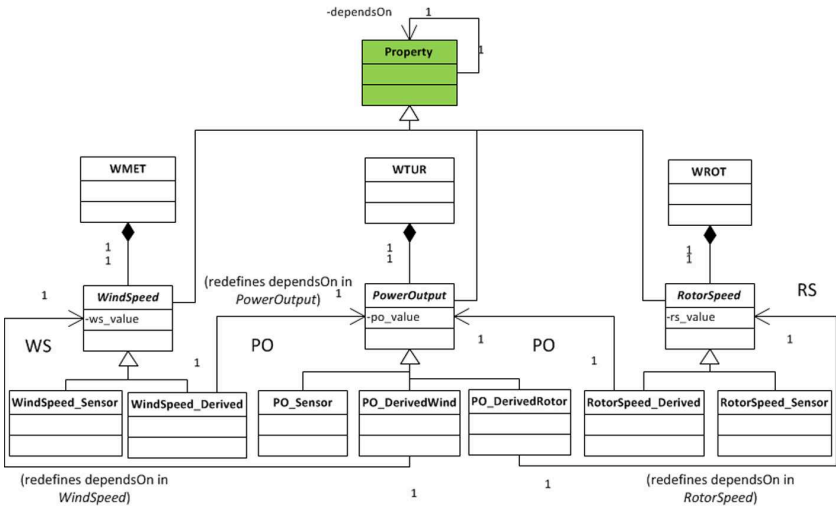

Fig. 9 Solving the cyclic derivation issue in derivative relationships between three parameters

4.3 Data Quality

Data quality can influence the decisions made by organizations. Indeed, wrong decisions can be made because of poor quality data [52, 22]. Data quality describes the characteristics of data and hence gives users a better view on data they want to request for. We consider data quality as metadata. Data quality has several dimensions which are criteria for selecting the most suitable data source according to users' requests. This section presents a solution to the challenge posed in Sect. 3.3.

4.3.1 Data Quality Dimensions

There are more than 17 data quality dimensions which have been mentioned in literature, e.g., accuracy, completeness, timeliness, consistency, access security, data volume, confidence, and understandability [58, 29, 7, 17]. The most commonly used quality dimensions are *accuracy*, *completeness*, and *timeliness* [44]. The other dimensions such as *confidence*, *value-added*, and *coverage* are only suggested by a couple of studies because these dimensions can be either derived from the other dimensions or applicable only in a few domains. There is no unique definition for each data quality dimension, so we describe the dimensions based on existing definitions and our understanding. Table 1 shows the notation that we use in our definitions.

Accuracy is defined as how close the observed data are to reality. According to the ISO 5725 standard [26], accuracy consists of precision and trueness.

We assume that the sensors are calibrated, meaning that the trueness is very close to zero. Therefore, we only consider precision as the accuracy in our system. A

Table 1 Table of notation

Symbol	Explanation
D	Data source
R	Reference data source (reality)
N_D	total number of data points in D
N_R	total number of data points in R
d_i	a single data point in D
r_i	real value corresponding to d_i
x_i	d_i - r_i
$t(r_i)$	the moment when the data point i is due
$t(d_i)$	the moment when the data point i is available

statistical measure of the precision for a series of repetitive measurements is the standard deviation. Let μ denote the trueness ($\mu = 0$). Thus, the accuracy of data source D can be obtained using Eq. (5).

$$Acc(D) = \sqrt{\frac{1}{N_D}\sum_{i=1}^{N_D}(x_i - \mu)^2} = \sqrt{\frac{1}{N_D}\sum_{i=1}^{N_D}(d_i - r_i)^2} \tag{5}$$

Completeness is defined as the ratio of the number of successful received data points to the number of expected data points. The completeness of the data source D can be calculated using Eq. (6).

$$Compl(D) = \frac{N_D}{N_R} \tag{6}$$

Timeliness is the average time difference between the moment a data point has been successfully received and the moment it is produced. The timeliness of data source D is calculated using Eq. (7).

$$Time(D) = \frac{\sum_{i=0}^{N_D}(t(d_i) - t(r_i))}{N_D} \tag{7}$$

4.3.2 Combination and Computation of Data Quality

By combining existing data sources, it is possible to improve the quality of data to meet the user defined requirement. The combination of data sources is defined as the process of constructing a data source from existing data sources. We present three simple methods to combine data quality: D1 (E) D2, D1 $\bigoplus$ D2, and D1 (A) D2.

- D1 (A) D2: taking a conventional average of the data sources D1 and D2.
- D1 $\bigoplus$ D2: use data points from data source D1 if available, otherwise use D2.
- D1 (E) D2: pick up the earliest received data point from either D1 or D2.

Table 2 gives an overview of all combination methods with data quality dimensions. These methods are used to generate the virtual data source from the real data

sources. P(D1) denotes the probability of the event *D1 having data available* and P(D2) denotes the probability of the event *D2 having data available*. $Acc(D1)$ and $Acc(D2)$ are the accuracy (precision) of $D1$ and $D2$, respectively. α is the probability of the event a data point $D1_i$ arrives before a data point $D2_i$.

Table 2 Combination Results

Method	Completeness	Accuracy	Timeliness
D1 (A) D2	$\overline{\overline{P(D1)} \cdot \overline{P(D2)}}$	$\sqrt{\frac{Acc(D1)^2+Acc(D2)^2}{4}}$	$\approx \frac{3}{2}Time(D1)$
D1 $\bigoplus$ D2	$\overline{\overline{P(D1)} \cdot \overline{P(D2)}}$	$\frac{P(D1)*Acc(D1)+\overline{P(D1)}*P(D2)*Acc(D2)}{P(D1)+\overline{P(D1)}*P(D2)}$	$\frac{Compl(D1)*Time(D1)+\overline{P(D1)}*P(D2)*Time(D2)}{P(D1)+\overline{P(D1)}*P(D2)}$
D1 (E) D2	$\overline{\overline{P(D1)} \cdot \overline{P(D2)}}$	$\alpha Acc(D1) + \overline{\alpha} Acc(D2)$	$\frac{Time(D1)*Time(D2)}{Time(D1)+Time(D2)}$

The following assumptions are made in order to obtain Table 2. (1) Data sources $D1$ and $D2$ are independent and normally distributed; (2) timeliness $Time(D1)$ and $Time(D2)$ of $D1$ and $D2$ are two independent distributed exponential random variables.

The combination methods have different effects on the data quality dimensions. A quality dimension can increase or decrease depending on a combination method. Table 3 shows relation the between the combination operations and the data quality dimensions, where (✓) indicates that it can be better than both of D1 and D2, (−) means it varies from case to case, and (×) means it is worse than both of D1 and D2.

Table 3 Quality Combination Relations

Combination method	Completeness	Accuracy	Timeliness
D1 (A) D2	✓	✓	×
D1 $\bigoplus$ D2	✓	−	−
D1 (E) D2	✓	−	✓

According to this table, all three methods can increase the completeness. By using the average method, the combined data source would have better accuracy. However, it makes the timeliness become worse. For the $\bigoplus$ method, both the accuracy and timeliness of the combined data source varies from case to case. The (E) method helps to increase the completeness and timeliness, but not the accuracy. If the timeliness is the critical choice, the (E) method is recommended to use.

4.3.3 A Data Quality-based Framework for Data Source Selection

We have developed a framework for data source selection based on data quality dimensions. An overview of the framework is shown in Fig. 10. The framework offers ways to manage data sources, to insert a new data source, and to provide the

best suited data source to users. Due to limitation of space, we cannot describe the prototype in detail. More information about the prototype implementation can be found in [44].

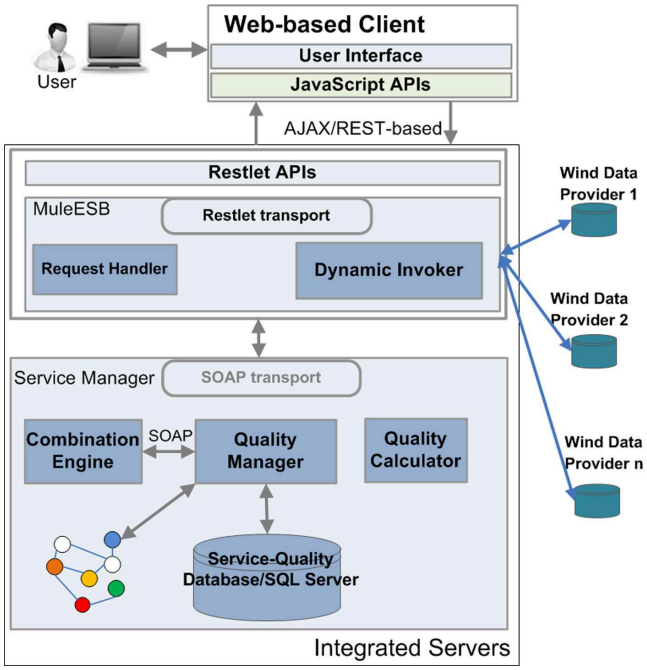

Fig. 10 An overview of a quality-based data source handling framework

The prototype contains three main parts: web-based client application, integrated servers (IS), and data provider services. The web-based client application receives requests from users and forwards them to the IS. The client is in charge of data visualization in terms of graphs. The IS is responsible for data quality handling and communicating with data providers. The data providers store the data and provide addresses to access those data. The IS consists of an open source enterprise service bus, MuleESB and the *Service Manager* which contains the *Combination Engine*, the *Quality Manager*, the *Quality Calculator*, and the *Service-quality Database*.

5 Discussion and Future Directions

One reason of having ontologies is to share an understanding of domain concepts between partners who are working in different domains. We have proven the useful-

ness of having ontologies in smart grids where energy generator, energy providers, consumers need to share the common view on domain concepts.

Many technologies (smart meter, semantic technologies, etc.) are mature enough to be used in building smart grids. But bringing these technologies together to enable smart grids is still a challenging task.

Information and communication security always has a significant role in any information systems and it is not an exception for smart grid systems. The power industry needs to manage not only the power system infrastructure, but also the information infrastructure. The reason is that the power industry increasingly relies on information to operate power systems and many manual operations are being replaced by automation. It is obvious that better decisions can be made by humans or intelligent systems based on available information. However, information needs to be made accessible in a secure way. One way of doing it is to lower the risk by granting access to metadata to only trusted partners.

Metadata provide information about data that are stored in a database without having accessed it [32]. Quality of metadata guarantees that proper sensing resources and data sources are found and data are used properly. The quality of metadata definitely affects the use of data and decisions that are based upon the data. There are several metadata quality criteria that must be taken into consideration such as correctness, completeness, accuracy, consistency, value-added, interpretability. Among these criteria accuracy, completeness, and consistency are the most common criteria for measuring metadata quality in literature [40]. The challenge is among those metadata quality dimensions which ones are the most important and how to check their quality. Another challenge that has not been addressed in this work is tracking provenance of metadata when it comes to metadata combination and enhancement. Besides management of metadata, agreement on the definition of concepts is also an important task since without understanding the definitions, metadata may be misinterpreted or misused. We plan to tackle these challenges in future work.

6 Conclusions

Technologies bring us closer to our vision for living in smart environments. Even though there are available technologies for us, it is still not an easy task to bring all the technologies together. A smart grid is an example of a smart environment. In smart grids, a huge number of smart meters, sensors, smart appliances, and other smart devices are employed and connected to Internet. This leads to issues in handling and processing vast amounts of data, and integrating these devices in a network so that they can communicate with each other through intelligent systems and applications. In this chapter, we have discussed issues related to management of big data metadata in smart grids. Three problems were addressed: concept modeling for knowledge sharing and data exchange, formal description of derived data from concept relations, and data quality handling. We have also proposed solutions and

approaches to solving these problems. Some concrete examples within the offshore wind energy were used to demonstrate our points.

This work shows that the semantic technologies are mature enough to be used in the development of smart grids in particular and smart environments in general. The work also proves that data quality can be improved in some cases by combining different data sources that provide measurements about the same physical phenomenon. Relations between concepts not only describe real-world objects/phenomena, but also open several possible paths for calculation and give users the possibility of selecting the most suitable one.

Acknowledgements This work has been (partially) funded by the Norwegian Centre for Offshore Wind Energy (NORCOWE) under grant 193821/S60 from the Research Council of Norway (RCN). NORCOWE is a consortium with partners from industry and science, hosted by Christian Michelsen Research.

References

1. IEEE Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation with the Electric Power System (EPS), End-Use Applications, and Loads. IEEE Std 2030-2011 pp. 1–126 (2011)
2. Ackoff, R.L.: From data to wisdom. Journal of Applied Systems Analysis **16**, 3–9 (2010)
3. Antoniou, G., Harmelen, F.v.: Web ontology language: OWL. In: Handbook on Ontologies, International Handbooks on Information Systems, pp. 91–110. Springer Berlin Heidelberg (2009)
4. Baclawski, K., Kokar, M., Waldinger, R., Kogut, P.: Consistency checking of semantic web ontologies. The Semantic Web ISWC 2002 pp. 454–459 (2002)
5. Balsters, H.: Modelling database views with derived classes in the UML/OCL-framework. In: UML 2003-The Unified Modeling Language. Modeling Languages and Applications, pp. 295–309. Springer (2003)
6. Barnaghi, P., Wang, W., Henson, C., Taylor, K.: Semantics for the internet of things: early progress and back to the future. International Journal on Semantic Web and Information Systems (IJSWIS) **8**(1), 1–21 (2012)
7. Baumgartner, N., Gottesheim, W., Mitsch, S., Retschitzegger, W., Schwinger, W.: Improving situation awareness in traffic management. In: Proc. Intl. Conf. on Very Large Data Bases (2010)
8. Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., Stein, L., et al.: OWL web ontology language reference. W3C recommendation **10** (2004)
9. Berners-Lee, T., Fischetti, M.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. HarperInformation (2000)
10. Bredillet, P., Lambert, E., Schultz, E.: CIM, 61850, COSEM standards used in a model driven integration approach to build the smart grid service oriented architecture. In: Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference, pp. 467 –471 (2010)
11. Bröring, A., Maué, P., Janowicz, K., Nüst, D., Malewski, C.: Semantically-enabled sensor plug & play for the sensor web. Sensors **11**(8), 7568–7605 (2011)
12. Camacho, E.F., Samad, T., Garcia-Sanz, M., Hiskens, I.: Control for renewable energy and smart grids. The Impact of Control Technology, Control Systems Society pp. 69–88 (2011)
13. Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., Zhou, X.: Big data challenge: a data management perspective. Frontiers of Computer Science **7**(2), 157–164 (2013)

14. Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al.: The SSN ontology of the W3C semantic sensor network incubator group. Web Semantics: Science, Services and Agents on the World Wide Web **17**(0), 25 – 32 (2012)
15. Datastax Corporation: Big Data: Beyond the Hype (2013). White paper
16. ETP: Smart Grids - Strategic Deployment Document for Europe's Electricity Networks of the Future (2010)
17. Geisler, S., Weber, S., Quix, C.: Onotology-based data quality framewrok for data stream applications. In: 16th International Conference on Information Quality, November 2011, Adelaide, AUS (2011)
18. Ghazel, M., Toguyéni, A., Bigand, M.: An UML approach for the metamodelling of automated production systems for monitoring purpose. Computers in Industry **55**(3), 283–299 (2004)
19. Gruber, T.R., et al.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human Computer Studies **43**(5), 907–928 (1995)
20. Hachem, N.I., Qiu, K., Serrao, N., Gennert, M.A.: GaeaPN: A Petri Net model for the management of data and metadata derivations in scientific experiments. Worcester Polytechnic Institute Computer Science Department Technical Report WPI-CS-TR-94 **1** (1994)
21. Horrocks, I., Patel-Schneider, P., Van Harmelen, F.: From SHIQ and RDF to OWL: The making of a web ontology language. Web semantics: science, services and agents on the World Wide Web **1**(1), 7–26 (2003)
22. Huang, K.T., Lee, Y.W., Wang, R.Y.: Quality Information and Knowledge. Prentice Hall PTR (1998)
23. Iannone, L., Rector, A.L.: Calculations in OWL. In: OWLED (2008)
24. IEC: IEC 61400 wind turbines - part 25: Communications for monitoring and control of wind power plants (2006)
25. Informatica Corporation: Metadata Management for Holistic Data Governance (2013). White Paper
26. ISO: ISO 5725-2: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 2: Methods for the Determination of Repeatability and Reproductibility. International Organization for Standardization (1994)
27. Kollia, I., Glimm, B., Horrocks, I.: SPARQL query answering over OWL ontologies. In: The Semantic Web: Research and Applications, pp. 382–396. Springer (2011)
28. Le-Phuoc, D., Nguyen-Mau, H.Q., Parreira, J.X., Hauswirth, M.: A middleware framework for scalable management of linked streams. Web Semantics: Science, Services and Agents on the World Wide Web **16**, 42–51 (2012)
29. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: A methodology for information quality assessment. Information and Management **40**(2), 133–146 (2002)
30. Lenzerini, M., Milano, D., Poggi, A.: Ontology representation & reasoning. Universit di Roma La Sapienza, Roma, Italy, Tech. Rep. NoE InterOp (IST-508011) (2004)
31. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity. Tech. rep., McKinsey Global Institute (2011)
32. Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., Manitsaris, A.: A conceptual framework for metadata quality assessment. Universitätsverlag Göttingen p. 104 (2008)
33. Muljadi, E., Pierce, K., Migliore, P.: Control strategy for variable-speed, stall-regulated wind turbines. In: American Control Conference, 1998. Proceedings of the 1998, vol. 3, pp. 1710–1714. IEEE (1998)
34. Muyeen, S., Tamura, J., Murata, T.: Wind turbine modeling. Stability Augmentation of a Grid-connected Wind Farm pp. 23–65 (2009)
35. Neuhaus, H., Compton, M.: The semantic sensor network ontology. In: AGILE Workshop on Challenges in Geospatial Data Harmonisation, Hannover, Germany, pp. 1–33 (2009)
36. Nguyen, T.H., Prinz, A., Friiso, T., Nossum, R.: Smart grid for offshore wind farms: Towards an information model based on the iec 61400-25 standard. In: Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES, pp. 1–6 (2012). DOI 10.1109/ISGT.2012.6175686

37. Nguyen, T.H., Prinz, A., Friisø, T., Nossum, R., Tyapin, I.: A framework for data integration of offshore wind farms. Renewable Energy **60**(0), 150 – 161 (2013)
38. NIST: NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 2.0, NIST Special Publication 1108R2 edn. (2012)
39. O'Connor, M., Das, A.: SQWRL: a query language for OWL. In: Proc. of 6th OWL: Experiences and Directions Workshop OWLED2009 (2009)
40. Park, J.R.: Metadata quality in digital repositories: A survey of the current state of the art. Cataloging & Classification Quarterly **47**(3-4), 213–228 (2009)
41. Patel-Schneider, P.F., Hayes, P., Horrocks, I., et al.: OWL web ontology language semantics and abstract syntax. W3C recommendation **10** (2004)
42. Ragheb, M., Ragheb, A.M.: Wind turbines theory-the betz equation and optimal rotor tip speed ratio. R. Carriveau, Fundamental and Advanced Topics in Wind Power pp. 19–37 (2011)
43. Ramirez, R.G., Kulkarni, U.R., Moser, K.A.: Derived data for decision support systems. Decision Support Systems **17**(2), 119–140 (1996)
44. Rasta, K., Nguyen, T.H., Prinz, A.: A framework for data quality handling in enterprise service bus. In: Innovative Computing Technology (INTECH), 2013 Third International Conference on, pp. 491–497 (2013)
45. Rossouw, L., Re, G.: Big data–big opportunities. RISK **16**(2) (2012)
46. Sheth, A., Anantharam, P., Henson, C.: Physical-cyber-social computing: An early 21st century approach. Intelligent Systems, IEEE **28**(1), 78–82 (2013). DOI 10.1109/MIS.2013.20
47. Singh, A.: Standards for Smart Grid. International Journal of Engineering Research and Applications (IJERA) (2012)
48. Sirin, E., Parsia, B.: SPARQL-DL: SPARQL query for OWL-DL. In: OWLED, vol. 258 (2007)
49. Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semantics: science, services and agents on the World Wide Web **5**(2), 51–53 (2007)
50. Snyder, B., Kaiser, M.J.: Ecological and economic cost-benefit analysis of offshore wind energy. Renewable Energy **34**(6), 1567–1578 (2009)
51. Solntseff, N., Yezerski, A.: A survey of extensible programming languages. Annual Review in Automatic Programming **7**, 267–307 (1974)
52. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data duality in context. Communications of the ACM **40**(5), 103–110 (1997)
53. Studer, R., Benjamins, V., Fensel, D.: Knowledge engineering: principles and methods. Data & Knowledge Engineering **25**(1-2), 161–197 (1998)
54. Tambouris, E., Manouselis, N., Costopoulou, C.: Metadata for digital collections of e-government resources. Electronic Library, The **25**(2), 176–192 (2007)
55. Tannenbaum, A.: Metadata solutions: using metamodels, repositories, XML, and enterprise portals to generate information on demand. Addison-Wesley Longman Publishing Co., Inc. (2001)
56. Vermesan, O., Friess, P., Guillemin, P., Gusmeroli, S., Sundmaeker, H., Bassi, A., Jubert, I.S., Mazura, M., Harrison, M., Eisenhauer, M., et al.: Internet of things strategic research roadmap. O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, et al., Internet of Things: Global Technological and Societal Trends pp. 9–52 (2011)
57. Wagner, A., Speiser, S., Harth, A.: Semantic web technologies for a smart energy grid: Requirements and challenges. In: ISWC Posters&Demos (2010)
58. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems pp. 5–33 (1996)
59. Warmer, J., Kleppe, A.: The object constraint language: getting your models ready for MDA. Addison-Wesley Longman Publishing Co., Inc. (2003)
60. Xu, H.: Critical success factors for accounting information systems data quality. Ph.D. thesis, University of Southern Queensland (2009)
61. Zikopoulos, P.C., Eaton, C., DeRoos, D., Deutsch, T., Lapis, G.: Understanding big data. The McGraw-Hill Companies (2012)
62. Zubcoff, J., Pardillo, J., Trujillo, J.: A UML profile for the conceptual modelling of data-mining with time-series in data warehouses. Information and Software Technology **51**(6), 977–992 (2009)