A universal and extensible data model for operational data of wind turbines

M. Pagitsch[*],
G. Jacobs, D. Bosse

Center for Wind Power Drives
Campus-Boulevard 61, 52074 Aachen, Deutschland

Contents

1 Abstract

Digitalization is becoming more and more important in the wind industry. All stakeholders in the value chain have realized that profound analyses of available datasets produce valuable information for increasing the efficiency both of processes and assets – which leads to an increase in compatibility. As the analysis of operational data is suited to cover as much as possible of available datasets, it needs to be conducted broadly automated. This again can only be realized meaningfully, if the data to be analyzed are as homogeneous as possible. The present paper describes a data model for data from wind turbines (WTs) and wind farms (WFs), which is one of the key enablers for a completely automated data evaluation process.

Keywords: Wind turbine SCADA data, database, data evaluation, data model, technical requirements, smart service platform

2 Introduction

Throughout the past years there have been intense activities related to collection and processing of operational data of WTs. Methods for analysing data from the supervisory control and data acquisition systems (SCADA) of WFs and the related WTs have been described in literature and are still in the focus of many research activities (see e.g. [BAC17, KUS10]). Mainly the big players in the wind industry are able to integrate evaluation of their assets' data in the processes of monitoring and maintenance. There are several activities aiming at enabling smaller operators to continuously monitor their turbines, detect anomalies in real-time, and thereby increase their competitiveness [JUN15, PAG18]: The possibility of quick reaction to underperformance of individual WTs leads to an increase in the assets' availability.

Why do we need a well-structured and universal data model?

The report on "Wind farm data collection and reliability assessment for operation and maintenance (O&M) optimization" issued by the International Energy Association (IEA) in May 2017 [IEA2017] states that there is a huge need for independent platforms and databases operated by neutral trustees. Especially collections of data from different assets allow for detailed and statistically profound analyses. Results of overarching analyses are beneficial for all stakeholders in the value chains of wind industry. [JUS17] shows the need for a standardised, independent and safe communication network for transferring operational data for the purpose of providing remote and predictive services in the energy supply networks. [PAG18] gives an overview of the most important design parameters for an independent smart services platform that collects and evaluates operational data from WTs.

Evaluating heterogeneous operational data at large scale (i.e. automated and for many WTs/WFs at a time) is only meaningful if algorithms do not have to be adjusted for the application to individual assets. On the one hand, the effort would be much too high; on the other hand, results from evaluations of data from different WTs would not be comparable. Basically this is an interface problem (between source data and evaluation algorithms) which can be resolved by converting the original data into a generic and universal data model before storing and evaluating them. Thereby, source data for algorithms are always equally structured. The quality of evaluation results is strongly dependent on the quality of source data, which can comparatively easily be ensured for homogeneous data sets. The Center for Wind Power Drives (CWD) aims at establishing a smart services platform for WTs together with its partners, which is capable of performing complex operations on heterogeneous datasets (i. e. by combining SCADA and maintenance data with the technical documentation). Therefore, a comprehensive and yet extensible data model based on existing guidelines and standards has been developed. A consistent data model comprising all kinds of data related to WTs (see sec. 3) is not yet known.

3 Development of the data model

In [IEA17], existing standards for operational data (not only wind-specific) have been analysed. None of those standards fulfils the requirements for a data model suitable for the collection of data in an overarching database. Some of the existing data models do not cover WTs as well as WFs, others are not consistently extensible by signals that have initially not been thought of. According to the report, WT data can be classed in four main groups:

- Equipment data: technical information on the WTs
- Operating data: data from the SCADA system and additional sensors
- Event data: data describing exceptional events, e.g. failures
- Maintenance data: documentation of maintenance activities

This structure provides the base for the CWD data model, which will be described in the following sections. A possible implementation in a database management system will be suggested. For reasons of brevity the paper focuses on the structure of and the systematic behind the model.

3.1 Base of the data model: IEC 61400-25-2

The standard DIN EN 61400-25-2 [DIN07] specifies the communication interfaces for monitoring and control of WTs and provides a substantial basis for the CWD data model. According to other standards as DIN EN 61850-7 [DIN04] it classes the signals from the SCADA system into so-called logical nodes (i. e. categories) guided by the logical structure of a WT. Each of the nodes contains signals (“attributes”) describing the state of the respective component (rotor, transmission, generator, etc.). The names of the signals in the nodes are constructed from abbreviations for assemblies, parts, and physical signals. Table 1 lists some examples.

Logical Node		Attribute	
Name	Explanation	Name	Explanation
WROT	Rotor	RotSpd	Rotor speed
		HubTmp	Temperature in rotor hub
		PtAngSpBl1	Setpoint of pitch angle in blade 1
WTRM	Transmission	BrkOpMod	Status of shaft break
		GbxOilLev	Gearbox oil sump level
		GsLev	Main bearing grease level
WGEN	Generator	Spd	Generator speed
		W	Generator active power
		GnTmpSta	Temperature of generator stator

Table 1: Examples of the taxonomy provided by [DIN07]

As can be seen in Table 1, there are several inconsistencies in the taxonomy which lead to ambiguities and need to be resolved.

- The logic of the signal names does not follow consistent rules (Oil sump level = WTRM:GbxOilLev whereas grease level in main bearing = WTRM:GsLev. The declaration “main bearing” is missing. This deteriorates the human readability of the signal names and facilitates errors.)
- In some cases the attribute’s name contains information on the superordinated logical node (“WGEN:GnTmpSta”), in most cases it does not.
- Symbols for physical quantities and the respective units are mixed up (generator speed = WGEN:Spd whereas active power = WGEN:W)

However, the taxonomy provides an easy and well-applicable method for structuring and labelling physical signals from WTs. Ideally, the result is both machine and human readable. The logic used in the CWD data model is closely related to the template provided by the standard.

Starting from the definition in the standard, the model has been extended to be capable of taking the data from different SCADA configurations. As sample datasets by some major companies have been taken into account, the model can be seen as “validated by design”. Furthermore, many data providers are able to make data available in the IEC format. Therefore it is to be expected, that only minor modifications/extensions of the data model will be necessary if new datasets are encountered.

Another approach for designing a data model could be to keep the original structures of given datasets and avoid conversion into a generic model. Aggregation could be taken care of not on database level, but rather in an additional software layer. However, this would increase complexity immensely. “Clean” data on database level allow for simple data retrieval by evaluation algorithms, thus the approach presented in this paper has been chosen.

3.2 Structure of the CWD data model

The CWD data model combines the structure postulated by [IEA17] (different classes of datasets) and the logic of [DIN07] (logical nodes, descriptive taxonomy) introduced above. It has been developed based on sample datasets provided by different operators and extends the amount of signals listed in [DIN07]. By avoiding inconsistencies as far as possible, it guarantees for high comprehensibility and extensibility. Parameters describing the WTs, signals from the SCADA system, alarm and status information, and data on maintenance activities are assigned to logical nodes. The composition of the signals' names follows the hierarchy of the system construction from standardised abbreviations: System – component – machine element – function. The name of each signal contains information on its origin as well as on its physical properties, which renders the names intuitively understandable. Complementing the list of abbreviations allows for easy extension of the data model without facing new inconsistencies.

Figure 1 gives an overview of the logical nodes contained in the data model. The light blue nodes contain time-dependent data from turbines. Respective data from WF controllers are assigned to the grey nodes. The dark blue nodes comprise time-independent data describing WFs and turbines (not contained in [DIN07]). Journal-like information (chronological status and alarm logs, documentation of maintenance activities) are contained in the petrol coloured nodes. The logical nodes' relation to the structure from [IEA17] and their content is described in the following paragraphs.

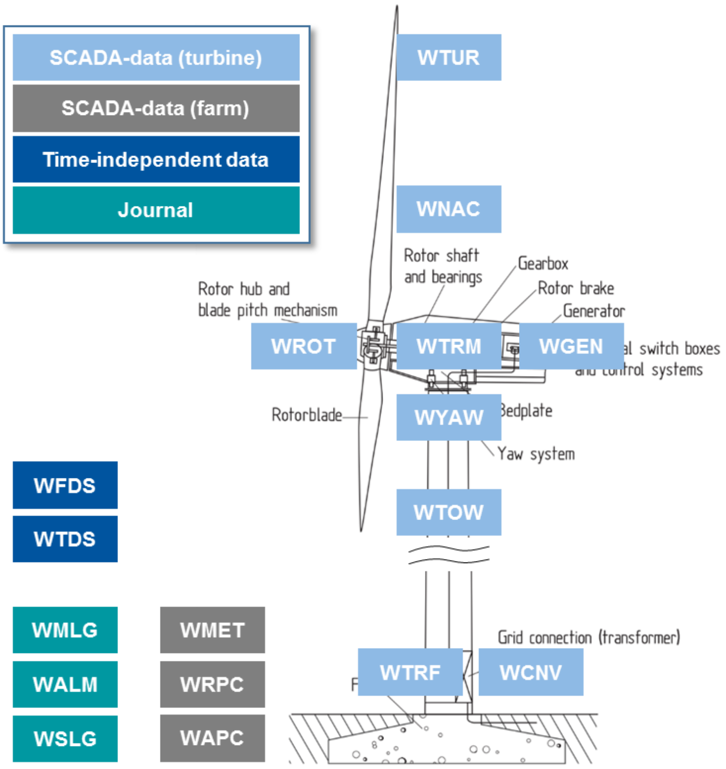

Figure 1: Logical nodes in the CWD data model

3.2.1 Equipment data

Those logical nodes comprise information on the WTs and WFs themselves, such as manufacturer, location, nominal power, rotor diameter, or nominal power curve.

- WFDS: Wind farm description; time-independent parameters of the wind farm

- WTDS: Wind turbine description; time-independent parameters of the WTs

3.2.2 Operating data

"Operating data" comprise status information from main components and the course of physical quantities such as rotational speed, active and reactive power production, temperatures, wind speed measured by the WT, or orientation.

- WTUR: Wind turbine general information; time-dependent data describing the general operation of the WT
- WNAC: Wind turbine nacelle information; time-dependent data describing the state of the nacelle
- WROT: Wind turbine rotor information; time-dependent data describing the operation of the rotor

The further logical nodes not mentioned in this list contain data describing state and operation of subsystems of the WTs accordingly.

3.2.3 Event data

"Event data" are lists of events as provided by the turbine and farm controllers. Furthermore, they can contain the history of commands received by the controllers.

- WALM: Wind power plant alarm information; alarm log of wind farm and turbines
- WSLG: Wind turbine state log information; condensed information on the operating state of WTs; history of commands and setpoints

3.2.4 Maintenance data

"Maintenance data" is a digital representation of the maintenance log.

- WMLG: Wind turbine maintenance log information; documentation of maintenance activities

3.3 The structure below using the example of WTRM

"WTRM" is the logical node containing all data of the WT transmission, i. e. all signals from the section of the mechanical drivetrain starting at the rotor flange and ending at the generator input shaft. Presumably this is the logical node with the highest heterogeneity in available source data: There are different drivetrain configurations and gearbox concepts which make it hard to standardise the signals. As a full description of the node would exceed the scope of this paper, the logic for temperature signals is explained in the following paragraphs.

"SftBrg1Tmp" and "SftBrg2Tmp" are temperature signals from the main bearing(s) carrying the main shaft. The numbering order of the measuring points is intended to follow the power flow increasingly. If there are more than two measuring points at the main shaft bearing(s), further numbers can be added accordingly.

The same system applies for the temperatures of the gearbox bearings (“GbxBrg1Tmp”, “GbxBrg2Tmp”, etc.), with one exception: Almost every SCADA dataset contains the temperature value of the gearbox bearing located next to the generator. As this bearing is subjected to failure more often than other bearings, its temperature is always being monitored. In the presented data model, the signal is named “GbxHSSBrgTmp”. The position of all other temperature signals cannot be exactly defined, as gearbox configurations and the availability of bearing temperatures vary widely among available datasets. However, the importance of those bearing temperatures is subordinate as compared to “GbxHSSBrgTmp”.

Usually, there are only few more temperature values in the transmission system: “Gbx-OilTmp” is describing the oil sump temperature of the gearbox, “SftBrkTmp” denotes the temperature of the shaft break.

3.4 Implementation in a database management system

As explained in previous publications [PAG18] the right choice of an appropriate database management system is built on Brewer’s CAP-theorem [BRE12]. According to that theorem, only two out of three main features (C – consistency, A – high availability, P – tolerance to network partitions) can be completely fulfilled by database management systems. In the present use case, reading processes (queries by evaluation algorithms) will constitute the prevailing load on the database, thus the establishment of consistency is not a problem. Availability and partition tolerance, on the other hand, guarantee for high scalability. Apache Cassandra is a widely used AP system and therefore recommended as the system of choice.

The data model described above can easily be implemented by use of only seven tables as shown in Figure 2. The figure describes how the data behind the individual logical nodes are combined in different tables. The primary keys of each table (unique identifiers for individual connected datasets) are denoted in brackets. Structurally equal parts of the dataset are stored in the same table, whereas parts differing in structure (e.g. by having timestamps that are not equidistant: SCADA datasets are usually recorded in intervals of ten minutes. Status or alarm logs, on the other hand, are not being collected at equidistant time intervals) are written to different tables in order to keep the structure of each table as simple as possible. Time-dependent values are stored with a primary key combined from the timestamp and the id of the respective asset for guaranteeing efficient queries of the database tables.

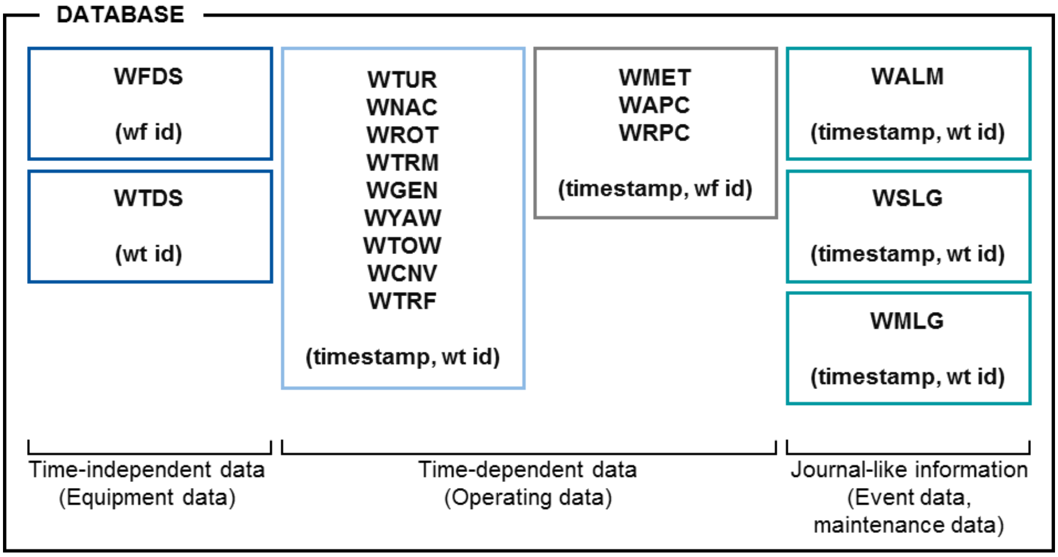

Figure 2: Table structure of the data model in the database

Some of the time-dependent SCADA data are often collected as a statistical distribution per time interval, denoted by mean value (avg), standard deviation (std), minimal (min), and maximal (max) value. In order to cope with that, most of the signal names in the SCADA tables exist four times with the abbreviations “avg”, “std”, “min”, and “max” as suffixes. If no information on the distribution is available, the signals are assigned to the respective “avg” name.

All physical quantities are stored in SI units. Values which are more but mere numbers, such as timestamps and information on geographical locations, are converted into formats that are easily processible by algorithms: For timestamps, the recommendations given in *Date and Time on the Internet: Timestamps* (RFC 3339) by *The Internet Society* [KLY02] are pursued. Locations are stored in the format defined in DIN EN ISO 6709 [DIN09]. Many programming languages have implementations of these standards and allow for easy conversion or basic operations on the data with next to no effort.

4 Benefits of the presented data model

The following sections give a short overview on the advantages yielded by application of the presented data model on database level.

4.1 Simplification of complex queries

By specifying and standardizing names and properties (such as physical units) of a huge number of signals, inhomogeneous datasets from different sources are homogenized and can easily be included in joint analyses. Therefore the data model forms the base for complex evaluations of huge volumes of data by relatively simple queries:

- Quantity (total amount of data) and quality (number of available signals) of datasets from different sources can easily be compared.

- All necessary data for detecting serial defects of individual machine elements in turbines of different manufacturers with a special gearbox configuration could be extracted.
- If event and/or maintenance data are available, the description of the events resp. maintenance activities can be used to generate annotated data for the training of artificial neural networks for failure prediction in a completely automated way.

The items in the list above are only meant to be examples for the impact generated by a well-structured pool of data.

4.2 Easy data retrieval and assessment by evaluation algorithms

All algorithms designed for working with the described data model, independently from their evaluation target, can be applied to any dataset *without modification*, provided that all required data are available. The quality of evaluation results calculated by specialised algorithms is highly dependent on the quality of the source data, therefore the algorithms deployed in the smart services platform need a certain awareness data quality. In order to facilitate the development of suchlike features, a guideline has been defined: Before performing the actual data evaluation, every algorithm assesses the quality of its source data and analyses the impact on the quality of its results. Three basic modes are distinguished:

- Source data ok, no restrictions
- Low quality and/or partial incompleteness in source data; algorithm operates with limitations. The quality of its result is scored with respect to the first case.
- Insufficient quality and/or incompleteness in source data; no calculations.

Assessing the quality of source data can be done with respect to different criteria such as quantity of available data, availability of statistical parameters, or plausibility. This process is dependent on the requirements of the regarded algorithms.

5 Summary

The pool of raw data is the core of a smart services platform. Good quality in the structure of the data is provided by a well-structured data model and guarantees for easy accessibility for evaluation algorithms and future viability due to extensibility. The data model used by the CWD is able to deal with heterogeneous datasets from turbines of different manufacturers and operators.

Database and algorithms are coordinated with respect to best quality of evaluation results or – as a fallback – the possibility of generating information on the loss of quality due to gaps in the input data.

By application of these principles – well-structured raw data and assessment of the quality of source data for algorithms – evaluations can be designed with high accuracy and confidentiality. This allows for providing valuable results, which contribute to gain deep insights in the WT’s operation and help to increase their effectivity.

6 Bibliography

[BAC17] Bach-Andersen, M., Rømer-Odgaard, B., Winther, O.: Deep learning for automated drivetrain fault detection.

In: Wind Energy, 1-13 (DOI: 10.1002/we.2142), 2017.

[BRE12] Brewer, E.: CAP Twelve Years Later: How the “Rules” Have Changed.

In: Computer 45, Issue: 2, Feb. 2012.

[DIN04] Deutsches Institut für Normung e.V.: Kommunikationsnetze und -systeme für die Automatisierung in der elektrischen Energieversorgung (DIN EN 61850-7:2004).

Beuth Verlag GmbH, Berlin, 2004.

[DIN07] Deutsches Institut für Normung e.V.: Kommunikation für die Überwachung und Steuerung von Windenergieanlagen – Informationsmodelle (DIN EN 61400-25-2:2007).

Beuth Verlag GmbH, Berlin, 2007.

[DIN09] Standarddarstellung für geographische Punkte durch Koordinaten (DIN EN ISO 6709:2009).

Beuth Verlag GmbH, Berlin, 2009.

[IEA17] International Energy Association: 17. Wind farm data collection and reliability assessment for O&M optimization.

Fraunhofer IWES, Kassel, 2017.

[JUN15] Jung, H. (Ed.): Erhöhung der Verfügbarkeit von Windenergieanlagen – EVW-Phase 2 (Final report)

https://wind-pool.iee.fraunhofer.de/opencms/opencms/wind_pool_de/Infocenter/ (2018-10-05)

[JUS17] Jussen, P., Nienke, S., Optehostert, F., Seelman, V., Birtel, F.: Connection of wind farms to an energy efficient and safe internet for energy communication network.

Proceedings of the Conference for Wind Power Drives, Aachen, 2017.

[KLY02] Klyne, G., Newman, C.: Date and Time on the Internet: Timestamps (RFC 3339).

The Internet Society, Network Working Group, 2002.

[KUS10] Kusiak, A., Li, W.: The prediction and diagnosis of wind turbine faults.

In: Renewable Energy 36, 16-32 (DOI: 10.1016/j.renene.2010.05.014), 2010.

[PAG18] Pagitsch, M., Jacobs, G., Bosse, D., Kock, S: Design of an independent smart service platform for wind turbines.

In: Journal of Physics: Conf. Series 1037 042026, 2018.