Wind energy open data web portal: Metadata and Taxonomy for data search

Anna Maria Sempreviva*, David Robert Verelst*, Gregor Giebel*, Lars Pilgaard Mikkelsen*, Christian Bak*, Hilmar Kjartansson Danielsen*, Nikola Vasiljevic*, Allan Vesth*, Mattias Anderson*, Stephan Barth, Javier Sanz Rodrigo***, Pawel Gancarski***, Tor Inge Reigstad ****, Hans Christian Bolstad ****, Jan Willem Wagenaar *****, Koen Hermans*****

*Technical University of Denmark, DTU Wind Energy, Roskilde, Denmark, anse@dtu.dk
** ForWind, University of Oldenburg, Oldenburg, Germany, stephan.barth@forwind.de
*** CENER, Sarriguren, Navarra, Spain, jsrodrigo@cener.com
**** SINTEF, Trondheim, Norway, toringe.reigstad@sintef.no
*****ECN, Petten, The Netherlands, j.wagenaar@ecn.nl

Abstract

We present the conceptual scheme for a Wind Energy data portal intended to make data Findable, Accessible, Inter-operable and Re-usable, FAIR, adhering to the Open Data strategy of the European Commission H2020 Programme. As a first step, metadata have been suggested and taxonomies for of the wind Energy topics and related data have been developed to be used as a common vocabulary for tagging data in the metadata card describing datasets. This effort is within the Open Data roadmap of the European Energy Research Alliance, the Joint Programme on Wind Energy, EERA JP Wind Energy.

Keywords: *Open data, Data portal, Taxonomy, Metadata*

Introduction

The Open Access, OA, to knowledge is a principle established by the European Commission underlying the H2020 EU Framework Programme for Research and Innovation. OA aims at optimizing the impact of publically funded projects, by making information openly available and reusable to all in Europe. The Open Data, OD, policy is part of the OA strategy and is widely acknowledged as a fundamental step to support a fast track from research to innovation. Although there is a general acknowledgment for the need of OD, a mindset similar to the "not-in-my-backyard" holds back the scientific and industrial communities to implement a joint OD policy. This is partly due to the fear that sensitive and proprietary data could be misused. To overcome this problem, the European Commission posed an important milestone by declaring that data must be at the same time "Findable, Accessible, Interoperable and Re-usable, FAIR" [1] , and "as much open as possible, and as closed as necessary". The European Energy Research Alliance, Joint Programme on Wind Energy (EERA JPWind)[i], as the largest public European scientific community in the Wind Energy (WE) sector, has widely recognized the necessity of implementing an OD plan and set the goal to create a data portal. The data portal will collect information on data from "cloud distributed" data centers, catalogue the collected information and provide end-users with tools to search and locate data for their needs. Here, we focus on the process for the design of the information system architecture of the data portal within the European Commission funded project Integrated Research Programme in Wind Energy, (IRPWind)[ii]

Method

The first phase, the process relates to make data "Findable" helping end-users to accurately retrieve the needed data. There are two components for this task (i) metadata and (ii) taxonomy for the WE sector topics and related data. (i) Metadata. To accurately locate specific datasets, they should be tagged with a series of information, metadata used for both documenting data for a future re-use, and for indexing datasets to refine their findability. Metadata are classified into three categories: descriptive, administrative and structural. Descriptive metadata provide information on e.g. what (associated topic, type of variables, etc.), where data were collected (external conditions or geographical location, etc.) or how data were collected (instruments, activity type). Administrative metadata provide information on e.g. who collected

the data (data owner), access rights, links to data, etc. Structural metadata provide information on e.g. data format. In this task, we use standard core metadata defined in the Dublin Core[iii] metadata element set. (ii) Taxonomy. Taxonomy is the descriptive type of metadata containing terms that assign textual information to the data. In a broad sense, it is any means of organizing concepts of knowledge (e.g. Environment, Climate, Agriculture, Engineering etc.). In a narrow sense, it is a hierarchical classification as we know e.g. from the Linnaean classification of species. Taxonomy of wind energy topics is used to put data into the correct context by defining and hierarchically classifying the WE research areas and organize data within topics. A good taxonomy enables users to immediately grasp the overall structure of the knowledge domain and the associated data. Practically, taxonomy terms are used by data owners as a controlled wind energy vocabulary for tagging data in the metadata card and by end-users as "facets" to filter content progressively by a "faceted search".

We created 7 taxonomies: the taxonomy of the topics distinctive of the WE sector; the taxonomy of the variables relevant in different topics, and other 5 facets for filtering data. The first step was to choose the number of hierarchical levels with top topics and subtopics. To keep the topic taxonomy structure simple , we stopped the development of taxonomy levels as soon as the next lower level reached the "data" dimension. The following case is given as an example: the topic "Siting" includes, amongst others, "wind mapping" for prospective sites. The "wind mapping" activity needs time series of wind speed and direction, and terrain roughness and orography data. In this view, we do not need to go beyond "wind mapping" as research topic and data needed for wind mapping is listed under the taxonomy for "Data Type".

Facets created for describing data are: External Conditions, Activities, Instruments, Models, Materials and Data type. The following case is given as an example: to perform the resource assessment offshore in Denmark, wind speed and directions from long-term observations using a wind lidar are needed. The search would be: Siting (Topic), Wind Resources (Subtopic), offshore, (External conditions), long-term monitoring (Activity type), wind lidar (Instrument) and wind speed and direction (Data type).

Conclusions

With metadata cards, describing available data, exposed by each organization, data can be searched through a data portal containing a metadata catalogue updated by a program continuously harvesting metadata cards (web crawler). The data physically resides on the data owner domain and any security and data management issues remains in the hand of the data owner.

A user will access the portal to submit a query containing key words from the established taxonomy and metadata vocabulary. The system will return an optimized list of available data. Data can be accessed either directly via provided download links, or by contacting data owners.

This approach has a two-fold purpose: to make data owners feel more comfortable in sharing data by maintaining the control on data access and data use, while end users will access information on datasets needed for a specific goal optimising time and funding. Both data owners and end-user will have the opportunity to start or re-inforce collaboration activities.

References

[1] M. D. Wilkinson, M. Dumontier, et al, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, 2016.

[i] **http://dublincore.org/documents/dces/**
[ii] **www.IRPWind.eu**
[iii] http://dublincore.org/documents/dces/