Semi-automatic taxonomy development for research data collections : the case of wind energy

Haakon Lund[1] & Anna Maria Sempreviva[2]

[1] *hl@hum.ku.dk*
University of Copenhagen, Department of Information Science, Njalsgade 76, DK-2300 Copenhagen (Denmark)

[2] *anse@dtu.dk*
Technical University of Denmark, Department of Wind Energy, Frederiksborgvej 399, DK-4000 Roskilde (Denmark)

Introduction

A metadata scheme and related controlled vocabularies, taxonomies, for the wind energy sector, have been proposed by the FP7 Project Integrated Research Programme in Wind Energy, IRPWind (Sempreviva et al., 2017). The goal was twofold: on one hand to comply with the principles of Findable, Accessible, Interoperable and Re-usable data (FAIR) (Wilkinson et al., 2016) introduced by the European Commission to support the open data policy for EU funded research projects. On the other hand, to answer to the growing concern within the research communities on how to identify and locate the vast amount of already available and future data from the ongoing digital transformation for research data management purposes. Research data management is increasingly adopted by funding agencies at national level as well. The faceted IRPWind taxonomies were developed by expert elicitation where a group of domain experts collaborated to establish e.g. a hierarchy of terms describing the WE topics. This process does demand extensive use of human resources and in a future perspective is not sustainable. Here, we propose using alternative methodology to create a semi-dynamic taxonomy, updated in time with new research trends, that relies on the analysis of keywords provided by authors of articles in domain journals. To test the method, we sat the goal of reproducing the IRPWind taxonomy of the topics in the wind energy sectors. For this purpose, we use keywords provided by authors to tag papers in the Wind Energy journal (ISSN 1099-1824) ISI Journal Citation Reports © Ranking: 2017:43/97 (Energy & Fuels) 2017:22/128 (Engineering, Mechanical). Impact factor 2.938. The Wind Energy journal does in its scope align with the topics covered by the IRPWind taxonomy.

Methodology

The co-occurrence analysis of author keywords from research papers has long been established as a viable way of identifying new trends in research and the development of a scientific domain (Woon and Madnick, 2009) and within the community of bibliometrics (Romo-Fernandez et al., 2013). This based on the assumptions that author provided keywords do express recent trends in research and therefore can provide valuable input to necessary taxonomy updates. The identification of research trends in a specific research domain is closely related to the identification of new terms to include in a domain specific taxonomy. Woon and Madnick (2009) suggest the use of keyword co-occurrence of author generated keywords for automated taxonomy construction.

We extracted 5717 keywords from 1159 papers published in Wind Energy journal covering the period from 1998 to 2018. Due to different forms of terms, equivalence and incoherencies in the choice of keywords by authors e.g. Speed /velocity, blade/turbine blade; wind turbine/Energy conversion system etc., only 2917 unique keywords were retained of which 356 occurred 3 times or more.

First, we clustered the filtered keywords based on the analysis of their co-occurrence and visualized the clusters using the integrated bibliometric tool, VOS – viewer (Van Eck & Waltman, 2010). Then, we used the resulting clustering maps to identify the core themes in the temporal development of wind energy (see figure 1). Last, a growth indicator (Woon, Henschel & Madnick, 2009), as a proxy for trends, was calculated based on term frequency expressed as weighted average publication year. The actual growth indicator was calculated as:

$$\theta_I = \frac{\sum_{t\in[firstyear,endyear]} t.TF_i\,[t]}{\sum_{t\in[firstyear,endyear]} TF_i\,[t]}$$

Where θ_I is the growth potential for keyword *I* and $TF_i[t]$ is the term frequency for term *I* and year *t.* A recent year suggests more prevalence of the topic.

Results

To evaluate the viability of using author provided keywords as candidate terms for taxonomy updates

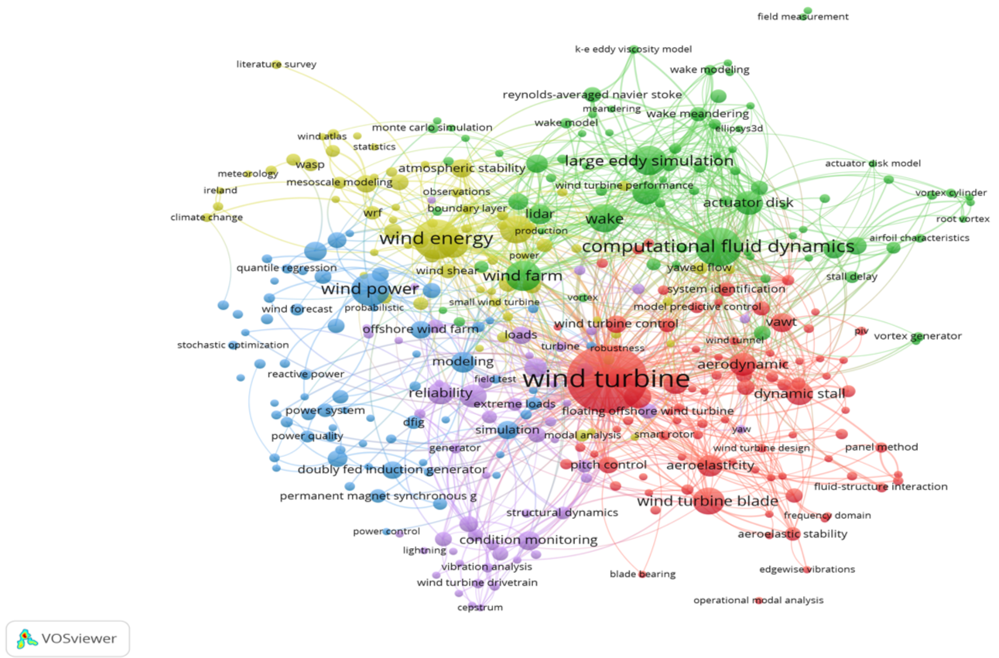

Figure 1. Keyword co-occurrence 1998 – 2018. 356 keywords occurring 3 or more times where analysed. A cluster size of 50 and a resolution of 0.5 were applied in VOS - viewer.

we calculated the overlap of existing IRPWind taxonomy terms with terms found by co-occurrence analysis.

Table 1. Overlap of IRPWind taxonomy terms and author keywords. L1 to L4 indicates the hierarchical levels in the IRPWind taxonomy for topics.

IRPWind level	L1	L2	L3	L4	Sum
#IRPWind terms (topics)	5	28	36	6	75
#author keywords (%)	3 (60%)	10 (36%)	13 (36%)	1 (17%)	27 (36%)

Conclusions

The resulting clusters were comparable to the IRPWind taxonomy of the WE topics. In research fields lacking metadata schemes and taxonomies, the use of author keywords instead of expert elicitation to arrange suitable taxonomies has pros and cons. An advantage is that using uncontrolled vocabularies allows detecting trends in scientific disciplines. Also, the procedure does not demand extensive use of human resources, as experts only will supervise the automatic procedures.

A shortcoming is that authors might use different words to identify the same activity, topic, instrument or variable depending on their field of activity e.g. electrical engineers use mostly wind power plant instead of wind farm. Also, the cumulated amount of author keywords will be a mix of terms identifying different categories e.g. activities, variable, topics, instruments etc., that must be semantically filtered in a number of acknowledged categories and meaningfully clustered.

References

Romo-Fernandez et al. (2013). Co-word based thematic analysis of renewable energy (1990–2010) *Scientometrics*. *97*(3), 743-765

Sempreviva, A.M. et al. (2017). *Taxonomy and meta data for wind energy R&D. Work Package 2-Deliverable D2.3.* IRPWind http://doi.org/10.5281/zenodo.1199489

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538

Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data,* 3:160018 doi: 10.1038/sdata.2016.18

Woon, W.L., Henschel, A. and Madnick, S. (2009) A framework for technology forecasting and visualization. *2009 International Conference on Innovations in Information Technology*. IEEE, 155-159. http://dx.doi.org/10.1109/IIT.2009.5413768

Woon, W.L. and Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, *21*(1), 91-111.