# Semi-Automated Data-Driven Methods to Support Ontology Development

## A Case Study on a Rehabilitation Therapy Ontology

Mohammad K. Halawani[1,3][0000−0003−0730−5870], Rob Forsyth[2][0000−0002−5657−4180], and Phillip Lord[1][0000−0002−4699−6769]

[1] School of Computing, Newcastle University, UK
[2] Institute of Neuroscience, Newcastle University, UK
[3] Department of Information Systems, Umm Al-Qura University, Saudi Arabia

Ontology development is expensive and requires significant efforts from both domain experts and ontologists. Automating the process usually produces unsatisfactory results and involves knowledge acquisition, which is intrinsically hard. In this abstract, we are investigating semi-automated techniques for bootstrapping and and supporting data-driven ontology development.

Rehabilitation therapies are hard to describe, measure and compare; unlike pharmacologic therapies, they are not precisely defined. This brings an interesting ontological challenge, because rehabilitation treatments are practice-based, diverse and involve interactions between a therapist, a patient and their environment. Therefore, we are using the domain of rehabilitation as a case study to build a rehabilitation therapy ontology (RTO).

Here, we are proposing a pipeline for building semantic knowledge structures to support developing ontologies from biomedical literature. The pipeline starts with an initial small set of articles provided by experts in the domain. This requires relatively little from the domain expert, beyond a set of references to appropriate papers, something that most researchers will have through their normal bibliography management facilities.

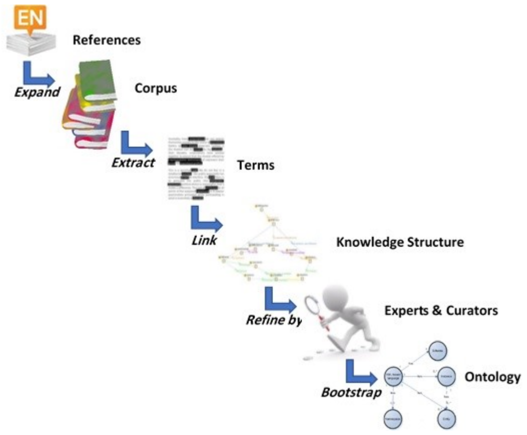

Fig. 1: Pipeline to support ontology development from literature.

The initial set of articles does not cover the domain; therefore, we expand this to a corpus of PubMed records that are relevant and cover the scope of the initial set using live PubMed's similar articles functionality and our pioneered relative similarity measure [1], that retrieves articles related to the whole initial set. In our case study , we were able to expand from initial set of 200 references, provided from two experts in the domain of rehabilitation, to around 28,000 references using this technique.

Full texts of the identified records of the corpus are then retrieved and pass through several text pre-processing and cleaning steps. For phrase detection, then, we apply *word2phrase* which is based on words' co-occurrences. Words and phrases in the text are the terms of the corpus, but they are not representative of the domain. To determine semantically meaningful and domain-related representative terminology, we apply the term frequency- inverse document frequency (*tf-idf*) technique. The result is a list of terms and phrases that are ranked according to their representation of the domain. Domain experts can arbitrarily threshold through the *tf-idf* scores to identify and extract top ranked representative terms.

The list of extracted terms can neither represent the semantics of the terms nor the relationships amongst them. Therefore, we develop a semantic knowledge structure that represents those. To develop the knowledge structure, we facilitate the list of extracted terms, their word embeddings from a trained *word2vec* [2] model, and a Directed Acyclic Graph (DAG) based on their lexical similarities, i.e. string-substring relationships. Semantic "subclass" relationships were found amongst the terms using the *word2vec* analogy technique. These were confirmed via the lexical DAG. Thus, we have a taxonomy-like knowledge structure based on *word2vec* semantic relationships. To add more relationships to the structure that are different from the "subclass" relationships, we can modify the *word2vec* analogy questions.

We hope that the final structure can be used to bootstrap an ontology by domain experts and curators rather than starting from scratch. This is similar to scaffolding the mitochondrial disease ontology [3]; nevertheless rather than using scaffolds from existing knowledge sources, here, we have generated the scaffolds in a data-driven method. These scaffolds are initially linked to easily discover semantic relations, and have a "todo" list ranked with their importance (i.e. the ranked list of terms) for curators to bootstrap the ontology in order.

## References

1. Halawani, M.K., Forsyth, R., Lord, P.: A literature based approach to define the scope of biomedical ontologies: A case study on a rehabilitation therapy ontology. arXiv preprint arXiv:1709.09450 (2017)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
3. Warrender, J.D., Lord, P.: Scaffolding the mitochondrial disease ontology from extant knowledge sources. arXiv preprint arXiv:1505.04114 (2015)