Data-driven Update of AGROVOC Using Agricultural Text Corpora

Hercules Panoutsopoulos [1] and Christopher Brewster [1,2]

[1] *Maastricht University, Institute of Data Science, Paul-Henri Spaaklaan 1 (PHS1), Maastricht, 6229 EN, The Netherlands*
[2] *TNO, Data Science Group, Kampweg 55, Soesterberg, 3769 DE, The Netherlands*

Abstract

AGROVOC is a well-known multilingual controlled vocabulary covering the fields of agriculture, forestry, fisheries, and food. It is used for dataset annotation, indexing of literature, and automated text tagging, and its effective use depends on its continuous update. Currently, updates are done manually by a dispersed community of editors. In this paper, we present work towards automated update recommendations using large corpora of agricultural text (such as the AGRIS database). The work is based on the extraction of agricultural concept mentions from text through the deployment of custom trained Named Entity Recognition models and the exploitation of Graph Neural Networks to recommend concept and relation additions towards predicting future AGROVOC states. The research questions and methodology are presented together with the results of an initial experiment. The next steps and future research directions are outlined. This work forms part of a PhD research on monitoring and predicting changes in knowledge graphs utilising textual data.

Keywords

AGROVOC, knowledge graph, update, Named Entity Recognition, Graph Neural Networks

1. Motivation

AGROVOC is a multilingual, structured vocabulary of more than 40K agricultural concepts, concept definitions and relations, and concept labels. It is structured as a directed acyclic graph using the SKOS standard[2] and represents associations between concepts by means of hierarchical and non-hierarchical relations. Utilising semantic web technology standards, AGROVOC provides knowledge organisation affordances enabling data retrieval. It allows standardised indexing via the unambiguous identification of resources, thus making search operations more efficient [1]. AGROVOC is curated by FAO experts in collaboration with editors from affiliated organisations. However, the pace at which new information and data become available, through the various kinds of publications, poses challenges to keeping it up to date. Advances in Natural Language Processing and Machine Learning hold the promise of providing technological support to the manual work involved in AGROVOC's maintenance and curation. In this context, the aim of this paper is to present work on the provision of automated recommendations for AGROVOC updates based on agricultural text corpora (such as the abstracts in the AGRIS database[3]). The goal is to identify concepts absent in AGROVOC but present in text to recommend for addition to an updated vocabulary version. Such recommendations include identifying where in the graph the new concepts should be added also specifying links to existing concepts. This work will eventually lead to methods for predicting future AGROVOC states based on the computation of diachronic changes.

Proceedings of HAICTA 2022, September 22–25, 2022, Athens, Greece
EMAIL: herculespanoutsopoulos@gmail.com (A. 1); christopher.brewster@maastrichtuniversity.nl (A. 2)
ORCID: 0000-0002-8060-9750 (A. 1); 0000-0001-6594-9178 (A. 2)


[2] https://www.w3.org/2004/02/skos/
[3] https://agris.fao.org/agris-search/index.do

2. Background and Related Work

There is a growing body of research on the development of knowledge graphs utilising unstructured or structured data sources (cf. [2] for a review of literature on automated knowledge graph construction). However, less research has been undertaken on automated knowledge graph update [3]. A method based on the combined use of Relational Graph Convolutional Networks (R-GCNs) [4], capturing an entity's context in a graph with bi-directional Gated Recurrent Units (bi-GRUs), having the capacity to identify the context of a word's appearance in text, is proposed in [5]. In that work, graph update is approached as a task of adding or deleting relations, assuming fixed sets of entities and relation types, to codify the information in the text. Fundamentally, research on automated graph update methods has taken the form of link prediction (e.g., [4, 6, 7, 8]). However, in such a context, important aspects, such as new concept addition, are overlooked. Apart from that, there is also interest in temporal node and graph embeddings [9, 10, 11]. Within this context, there has been work in time-aware relational Graph Neural Networks (GNNs) predicting new relations based on diachronic changes in the graph [12].

3. Description of Proposed Research

AGROVOC provides affordances for annotation of agricultural data, information retrieval, literature indexing, and automated text tagging [13]. Given the pace at which new information becomes available, it is important to timely capture domain developments, taking these from food- and agriculture-related publications, and integrate them into AGROVOC, to ensure an up-to-date knowledge representation enabling accurate resource identification. AGROVOC has grown over the years following changes in the domain as shown in Figures 1 and 2[4].

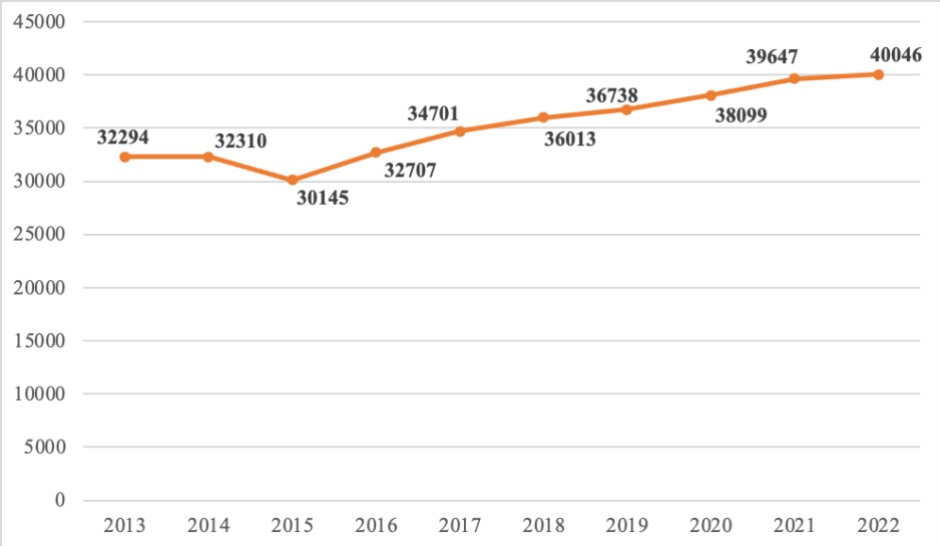

Figure 1: Number of concepts in AGROVOC per year

The number of concepts in AGROVOC (Figure 1) has increased over time, which is to be expected given the developments in the fields of food and agriculture. However, changes in the number of relation types (Figure 2) have not followed a similar pattern, with the observed drops in the recorded numbers requiring further explanation. To acquire further insights into how AGROVOC is updated, the creation dates and temporal distribution of concept occurrences in literature (abstracts from the AGRIS database) were computed for a random sample of concepts from the 2022 AGROVOC version (Table 1).

[4] Figures 1 and 2 have been created using data from SPARQL queries submitted to the AGROVOC versions from 2013 and 2022. The queries are available in the paper's GitHub repository.

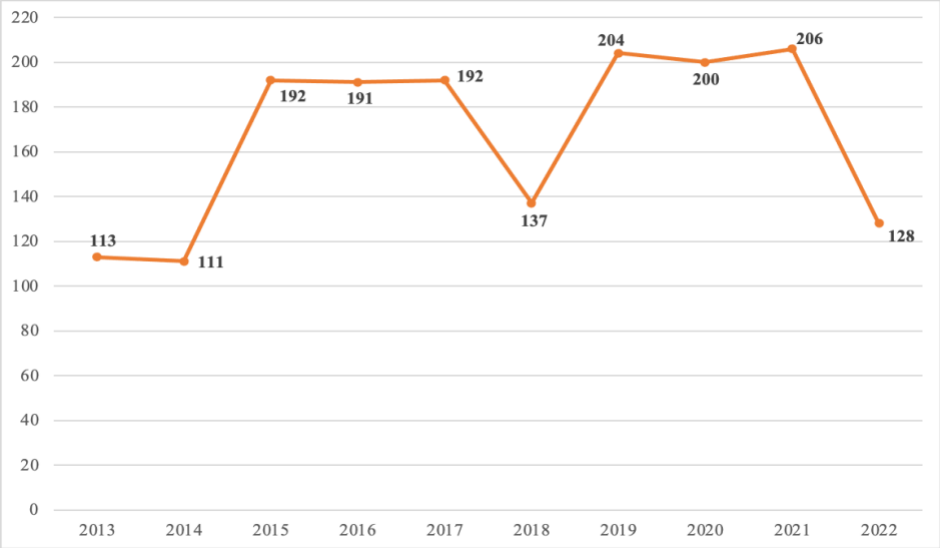

Figure 2: Number of relation types in AGROVOC per year

Table 1

Dates of addition of a sample of concepts in AGROVOC and numbers of their occurrence in literature before and after their addition to AGROVOC

Concept	Date of addition to AGROVOC	Occurrences in literature before creation	Occurrences in literature after creation
c_5903	2011	422	119
c_59e0f842	2019	1668	34
c_25740	2011	143	29
c_27140	2011	125	19
c_786c0cff	2019	2843	464
c_33193	1990	0	494
c_62e403a1	2019	209	327
c_41ce07e7	2017	231	307

Despite the small sample size, it is evident that in many cases the number of concept occurrences in literature before their addition to AGROVOC is greater than the number of their occurrences after being added to AGROVOC. It can be concluded that the addition of new concepts to AGROVOC is not based on their frequency of occurrences in literature. This is further supported by the temporal distribution of new concept additions illustrated in Figure 3[5]. A high peak in the number of concepts added in 2011 is observed (26,667 concepts) with the average number of concept additions per year being much lower before 2011 (≅ 66 concepts) and after 2011 (≅ 800 concepts). Based on these findings and considering the rapid pace of advances in agriculture, we propose that manual updates appear to not be sufficient for the timely capture and representation of new knowledge.

The proposed PhD research aims to develop, test, and evaluate methods recommending automated AGROVOC updates based on text. This forms part of a broader effort on the monitoring and predicting of changes in knowledge graphs utilising textual data. To this end, we have posed the following research questions:

1. How can we extract agricultural concepts from text, absent in AGROVOC, and identify which ones to propose as new concepts to be integrated into AGROVOC?
2. Given a new concept to be integrated into AGROVOC what existing relations need also to be added to link the new concept to existing concepts?

[5] The code used to obtain the data shown in Table 1 and Figure 3 is available in the paper's GitHub repository.

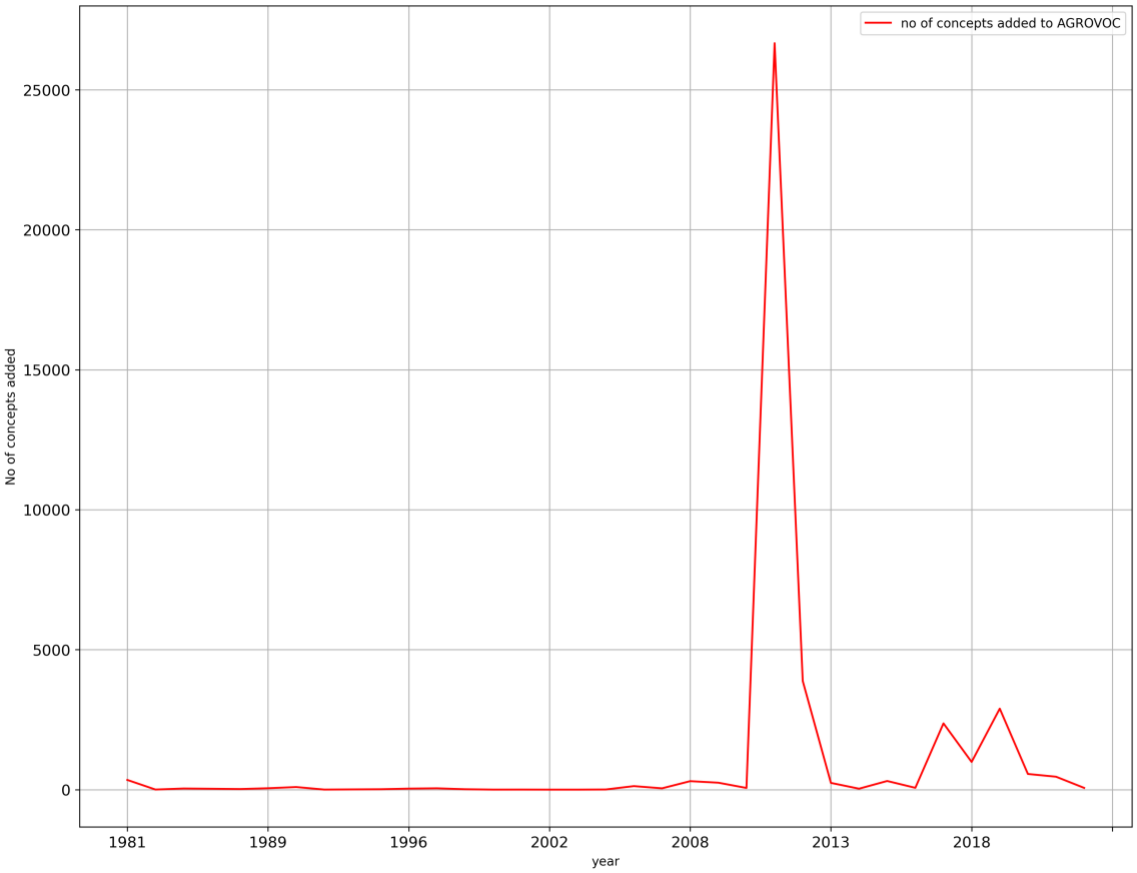

Figure 3: Temporal distribution of new concept additions in AGROVOC

4. Research Methodology and Experiments

The research methodology, depicted in Figure 4, has two phases: (i) Extraction of novel agricultural concepts from text; and (ii) Generation of recommendations for automated AGROVOC updates. Each phase involves the implementation of an experiment. The experiments are described below.

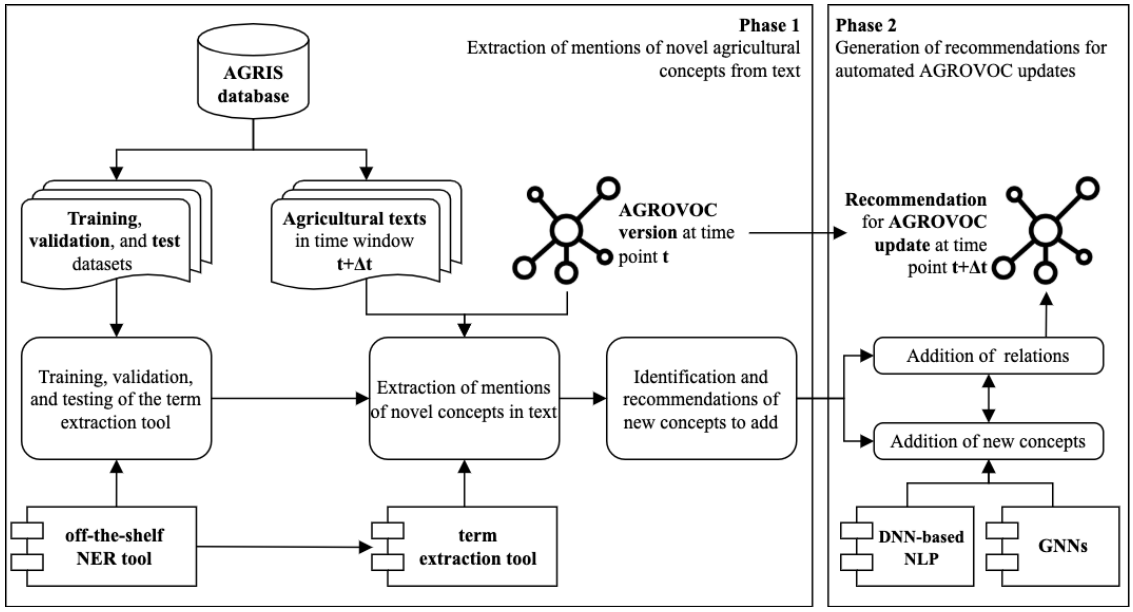

Figure 4: PhD research methodology

Extraction of mentions of novel agricultural concepts from text: The focus is on the development of an agricultural term extraction tool to identify mentions of novel concepts (not seen in AGROVOC) in the corpus of texts. Given a version of AGROVOC available at a time point t and a corpus spanning across a time frame t+Δt, the goal is to identify new concept mentions and recommend them to be added to the vocabulary. The term extraction tool is based on off-the-shelf Named Entity Recognition (NER) models. Abstracts of AGRIS publications are used as the tool's training, validation, and test datasets. An initial version of the tool was built based on the spaCy library's Tok2Vec[6] and NER[7] components, using their default architectures (spacy.Tok2Vec.v2 and spacy.TransitionBasedParser.v2 respectively) and the language models shipped with spaCy (en_core_web_sm and en_core_web_lg). Training was made on a set of 617 AGRIS abstracts annotated manually with labels of agricultural concepts appearing in them. Table 2 lists the best precision, recall and F1-score achieved in the initial experiment and the tool configurations giving those results. The results reveal the challenges related to the classification of a string as an agricultural term, when manually annotating text with agricultural terms, which has a high degree of vagueness, and hence subjectivity, leaving room for different interpretations by humans and impacting performance. Optimisation of the term extraction tool based on the use of transformer-based architectures and agriculture-related vocabularies and ontologies to unambiguously annotate text is currently in progress.

Table 2
Best precision, recall, and F1-score and configurations of the term extraction tool giving those results

Model configuration (language model - batch size - learning rate)	Precision	Recall	F1-score
"en_core_web_lg" - 128 - 0.01	**50.73%**	47.34%	48.97%
"en_core_web_sm" - 64 - 0.0001	46.08%	**54.52%**	49.95%
"en_core_web_sm" - 64 - 0.0001	50.70%	52.96%	**51.81%**

Generation of recommendations for automated AGROVOC updates: This experiment focuses on the generation of automated updates of AGROVOC drawing upon recommendations for adding new concepts and relations (from a set of existing relation types) to link the new conceps to concepts already in AGROVOC. The method will be based on Deep Neural Network-based Natural Language Processing (DNN-based NLP), capturing the context of agricultural concept mentions in text, and Graph Neural Networks (GNNs) capable of capturing a concept's context in the graph, thereby allowing to identify where in the graph the new concept should be added and how it should be linked to existing concepts. The available AGROVOC versions will be used as ground truth to evaluate the method's performance.

5. Discussion

AGROVOC is an agriculture-related graph knowledge representation structure that can be used in various application scenarios. To facilitate an accurate identification of resources, based on its use, it is important to keep AGROVOC up to date. However, the rate at which new information and data become available together with the issues emerging from the AGROVOC's update methods currently in practice (appearing not to follow the pace of domain developments as made evident from the relevant literature) necessitate the adoption of automated update solutions based on means of technological support. In this context, this paper has presented a PhD research on automated AGROVOC updates based on the extraction of novel concept mentions from text. Further work is currently in progress related to the development of the tool for extracting agricultural terms from text towards improving its performance. To this end, domain ontologies and vocabularies are intended to be used to annotate text automatically and unambiguously for obtaining the tool's training, validation, and test datasets. Moreover, drawing upon transformer-based architectures will help to get better performance results. Future research will

[6] https://spacy.io/api/tok2vec
[7] https://spacy.io/api/entityrecognizer

be concerned with the deployment of time aware GNNs predicting future states of AGROVOC solely based on the computation of changes that have diachronically occurred in it.

6. Acknowledgements

The authors would like to thank FAO's support facility for providing previous AGROVOC versions. This work has been partly supported by the H2020 EUREKA project, contract number 862790.

7. References

[1] I. Subirats-Coll, K. Kolshus, A.Turbati, A. Stellato, E. Mietzsch, D. Martini, and M. Zeng. AGROVOC: The linked data concept hub for food and agriculture. Computers and Electronics in Agriculture 196 (2022) p. 105965. doi: 10.1016/j.compag.2020.105965.

[2] M. Masoud, B. Pereira, J. McCrae, and P. Buitelaar. Automatic Construction of Knowledge Graphs from Text and Structured Data: A Preliminary Literature Review, in D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, B. Heinisch (Eds.), Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021), Informatics Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, 2021, Article No. 19; pp. 19:1–19:9. doi:10.4230/OASIcs.LDK.2021.19.

[3] G. Weikum, X.L. Dong, S. Razniewski, and F. Suchanek. Machine knowledge: Creation and curation of comprehensive knowledge bases. Foundations and Trends in Databases 10 (2021) 108-490. doi: arXiv:2009.11564v2.

[4] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. Modeling Relational Data with Graph Convolutional Networks. arXiv preprint (2017). doi: arXiv:1703.0610 3v4.

[5] J. Tang, Y. Feng, and D. Zhao. Learning to Update Knowledge Graphs by Reading News, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics pages, Hong Kong, China, 2019, pp. 2632–2641. doi: 10.18653/v1/D19-1265.

[6] A. Grover, and J. Leskovec. node2vec: Scalable feature learning for networks, in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2016, pp. 855-864. doi: 10.1145/2939672.2939754.

[7] M. Zhang, and Y. Chen. Link prediction based on graph neural networks. arXiv preprint (2018). doi: arXiv:1802.09691v3.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S.Y. Philip. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32 (1) (2020) 4-24. doi: 10.1109/TNNLS.2020.2978386.

[9] O. Michail. An introduction to temporal graphs: An algorithmic perspective. arXiv preprint (2015). doi: arXiv:1503.00278v1.

[10] U. Singer, I. Guy, and K. Radinsky. Node embedding over temporal graphs. arXiv preprint (2019). doi: arXiv:1903.08889v3.

[11] A. Taheri, and T. Berger-Wolf. Predictive temporal embedding of dynamic graphs, in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 57-64. doi: 10.1145/3341161.3342872.

[12] A. Pareja, G. Domeniconi, J. Chen, T. Ma, T. Suzumura, H. Kanezashi, T. Kaler, T. Schardl, and C. Leiserson, C. Evolvegcn: Evolving graph convolutional networks for dynamic graphs, in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 5363-5370. doi: arXiv: 1902.10191v3.

[13] E. Mietzsch, D. Martini, K. Kolshus, A. Turbati, and I. Subirats-Coll. How Agricultural Digital Innovation Can Benefit from Semantics: The Case of the AGROVOC Multilingual Thesaurus. Engineering Proceedings 9 (1) (2020) 17. doi: 10.3390/engproc2021009017.