Creating controlled vocabularies for smart search at WUR

dr. JL (Jan) Top, B (Bengü) Öztürk PhD, JJ (Jim) Hoekstra MSc, dr. RJ (Rutger) Vlek

PUBLIC

WAGENINGEN
UNIVERSITY & RESEARCH

Creating controlled vocabularies for smart search at WUR

Authors: dr. JL (Jan) Top, B (Bengü) Öztürk PhD, JJ (Jim) Hoekstra MSc, dr. RJ (Rutger) Vlek

This study was carried out by Wageningen Food & Biobased Research and commissioned by Wageningen Food & Biobased Research.

Wageningen Food & Biobased Research
Wageningen, January 2023

Public

Report 2383
DOI 10.18174/583753

WFBR Project number: 3720021400
BAPS number: KB-35-005-001
Version: Final
Reviewer: G (Görkem) Simsek-Senel PhD
Approved by: dr.ir. H (Henk) Wensink
Carried out by: Wageningen Food & Biobased Research
Commissioned by: Wageningen Food & Biobased Research
This report is: Public

The client is entitled to disclose this report in full and make it available to third parties for review. Without prior written consent from Wageningen Food & Biobased Research, it is not permitted to:

a. partially publish this report created by Wageningen Food & Biobased Research or partially disclose it in any other way;
b. use this report for the purposes of making claims, conducting legal procedures, for (negative) publicity, and for recruitment in a more general sense;
c. use the name of Wageningen Food & Biobased Research in a different sense than as the author of this report.

The research that is documented in this report was conducted in an objective way by researchers who act impartial with respect to the client(s) and sponsor(s). This report can be downloaded for free at https://doi.org/10.18174/583753 or at www.wur.eu/wfbr (under publications).

© 20233 Wageningen Food & Biobased Research, institute within the legal entity Stichting Wageningen Research.

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher. The publisher does not accept any liability for inaccuracies in this report.

Contents

Summary **4**

1 Introduction **5**

2 Constructing controlled vocabularies **8**

2.1 TALK Tool 8
2.2 ROC+ 9
2.3 NLP for automated term extraction 10

3 Controlled vocabularies at WUR **12**

4 Conclusion **13**

Literature **14**

Summary

Searching text or documents in large unstructured and semi-structured data sources is not trivial. A search engine is supposed to make more search efficient and effective. It supports to build a query that can be applied automatically to extract the information that complies with the user's intention. Controlled vocabularies and ontologies help improving the search and make it domain-aware. In this document, we explain the notion of a controlled vocabulary, its construction methods and its use in smart search engines. Manual construction of controlled vocabularies and ontologies can be achieved using several existing tools, which require specific technical skills. Therefore, we refer to the ROC+ tool, developed within WFBR, which helps domain researchers build a controlled vocabulary in a faster and easier way. Another application, namely the TALK tool, was developed to start a discussion on a specific term in multidisciplinary teams. It proposes automatically generated associated terms, which can then be exported in a machine processible format as input for ROC+. We also briefly mention the use of NLP technology in text mining, where domain related concepts can be automatically extracted from pdf documents. Finally, some example of controlled vocabularies developed within WFBR are listed for further reference.

1 Introduction

The amount of data collected worldwide is growing rapidly. A substantial amount of the available data generated and collected by organizations is in unstructured format (e.g., web documents, videos, images, scientific publications). Searching for relevant items and extracting information from such resources is not straightforward. Therefore, we need tools to support users and systems to find specific information in unstructured or semi-structured data sources. We refer to these tools as *search engines.*

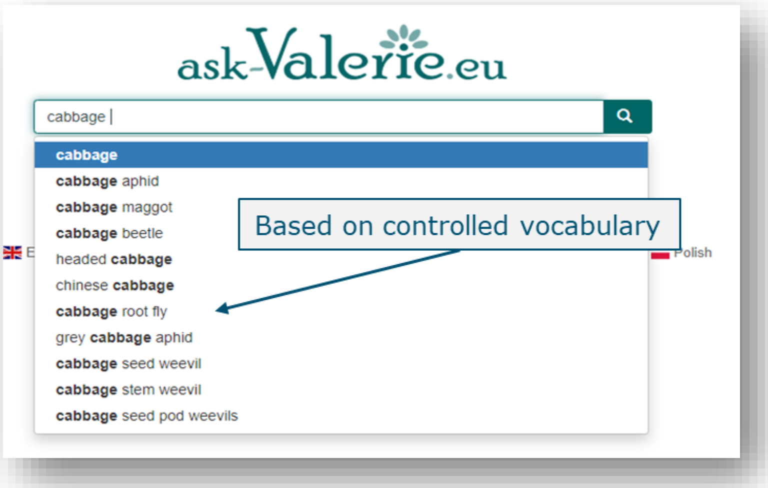
When searching for information, people normally start by stating a simple search phrase, based on their own knowledge of the domain and their expectation of what they can find in the sources they use. A search support tool is supposed to guide them from this initial phrase to a full, specific query. This query automatically extracts the information that complies with the user's intention. So, both the intention of the user and the content of the repository have to be 'discovered' during this process, looking for the closest and most specific match.[1]

This implies that a search engine plays two roles. First, it represents the domain experts that supplements the user's knowledge about the considered domains (physics, agriculture, genomics, etc.). In automated systems, part of this knowledge can be expressed in domain ontologies and controlled vocabularies. Ontologies typically define the concepts, relations and properties that can be used to build data models, used in *structured* repositories (databases). Controlled vocabularies, on the other hand, are lists of terms used in a domain, organized in a less formal and strict manner. Second, a search engine acts as the librarian who knows about the content of the repository and the items it contains. Therefore, she requires each item to be described by its *metadata,* in a way that can be understood by the future user. For automated systems, we require this metadata to be also machine-readable, such that it can be processed automatically by a search engine. In addition, the engine contains a *search index*. This is a list of terms that links a user's request to specific (elements of) items contained by the repository.

Ideally the metadata and index are linked to the above mentioned controlled vocabularies. In practice this allows search engines to execute certain support actions, i.e., act as domain experts and librarians. Examples of such functions are:

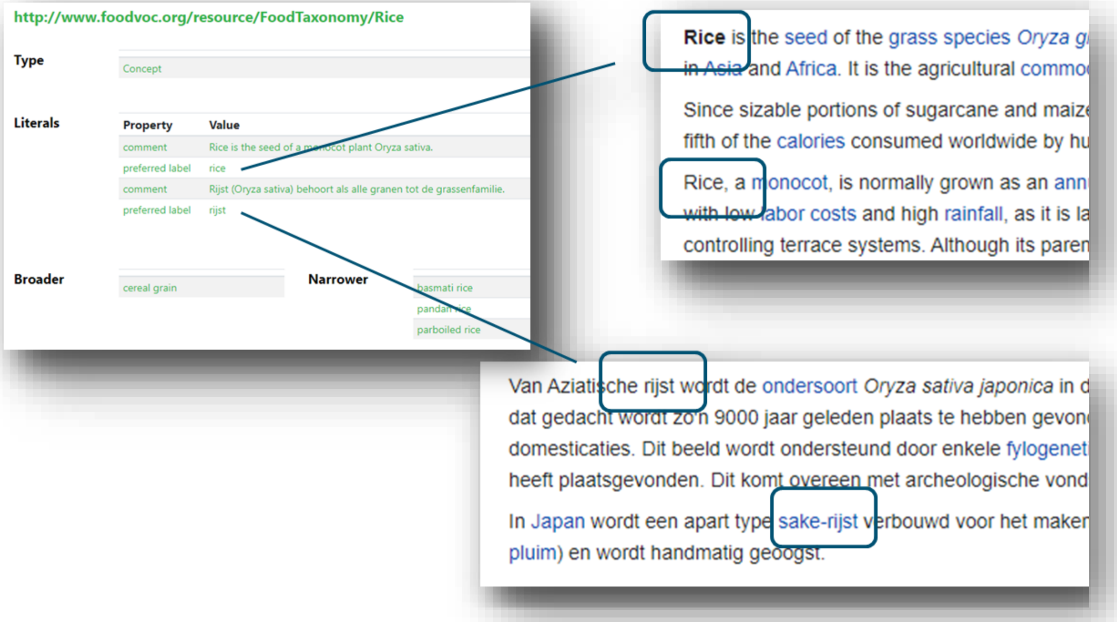
- Query support. When typing search terms in a search field, a user can get suggestions for terms to use. An example is shown in Figure 1. This helps him to make his intentions explicit, discover domain terms and identify relevant items (through their metadata).
- Query expansion. Once the user has formulated a search phrase, some of terms can be expanded automatically to make the search more effective (Figure 2). For this purpose for example the narrower terms and related terms can be added automatically. So, a search for 'bird' will also find an item that talks about 'sparrow', even if it doesn't not mention the word 'bird' explicitly. Search engines have to be careful in handling AND and OR operators when applying automated query expansion.
- Term editing. A search engine can also allow the user to manually adapt terms in a search phrase by walking through the vocabulary (Figure 3). So rather than leaving 'bird' in the example search string, he can decide to replace it by the narrower term 'sparrow'.
- Multilanguage support. Controlled vocabularies can contain translations of terms, allowing the search engine to find items using a different language that is used by the user.
- Fragment selection. In textual items (reports, documents, papers, etc.) vocabularies can assist in selecting specific fragments of a text, in support of the search index.

[1] Although in this document we focus on unstructured data, user also need support to find information in structured repositories (databases). However, in those cases the knowledge needed to extract information is (partly) expressed in the repository itself. This asks for a different approach, such as faceted search and database querying.



cabbage
cabbage aphid
cabbage maggot
cabbage beetle
headed cabbage
chinese cabbage
cabbage root fly
grey cabbage aphid
cabbage seed weevil
cabbage stem weevil
cabbage seed pod weevils

Figure 1 An example of query support using a controlled vocabulary


http://www.foodvoc.org/resource/FoodTaxonomy/Rice

Type: Concept

Literals

Property	Value
comment	Rice is the seed of a monocot plant Oryza sativa.
preferred label	rice
comment	Rijst (Oryza sativa) behoort als alle granen tot de grassenfamilie.
preferred label	rijst

Broader: cereal grain

Narrower: basmati rice
pandan rice
parboiled rice

Rice is the seed of the grass species *Oryza g*
in Asia and Africa. It is the agricultural commo

Since sizable portions of sugarcane and maiz
fifth of the calories consumed worldwide by hu

Rice, a monocot, is normally grown as an ann
with low labor costs and high rainfall, as it is la
controlling terrace systems. Although its paren

Van Aziatische rijst wordt de ondersoort *Oryza sativa japonica* in d
dat gedacht wordt zo'n 9000 jaar geleden plaats te hebben gevon
domesticaties. Dit beeld wordt ondersteund door enkele fylogenet
heeft plaatsgevonden. Dit komt overeen met archeologische vond

In Japan wordt een apart type sake-rijst verbouwd voor het maken
pluim) en wordt handmatig geoogst.

Figure 2 An example of query expansion, leading to the retrieval of two documents about rice, in different languages

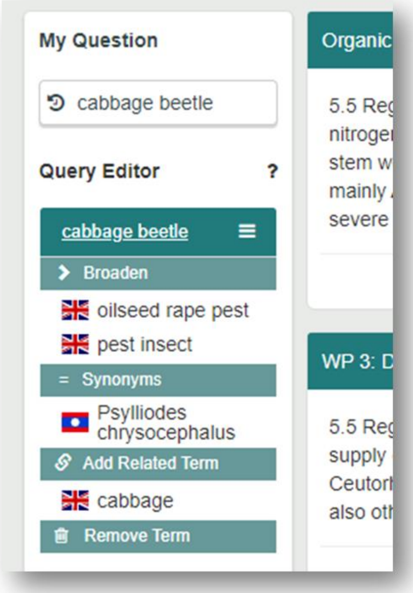


Figure 3 A query-editor that uses a controlled vocabulary to make a query more precise

The above examples of these different functions are taken from Ask-Valerie (www.ask-valerie.eu), which can be accessed for further exploration. This engine was developed to disclose expertise in the field of agriculture and forestry.

In this document, we focus on the development of controlled vocabularies. For an introductory text on ontologies for WUR, refer to this page[2] on WUR Intranet.

The rest of the document is organized as follows: Section 2 is about constructing controlled vocabularies using TALK and ROC+ tools, Section 3 presents some examples of controlled vocabularies developed within WUR projects, and Section 4 provides our final remarks and future suggestions.

[2] Using ontologies within Wageningen University & Research; https://intranet.wur.nl/Project/KnowledgeEngineeringGroup/Pages/UDvV3J9l4Ee9IiIgktwCtg

2 Constructing controlled vocabularies

A so-called controlled vocabulary (CV) is a hierarchical list of terms (or *noun phrases*, consisting of multiple terms, such as 'molecular biology') that is maintained by domain experts. A typical example application is the taxonomy of biological organisms, represented as a tree of names of these organisms. Controlled vocabularies can be used in metadata for subsequent retrieval of relevant items, as described in the previous section.

If such a controlled vocabulary is expressed in a standardized format, search can be partly automated. SKOS (*Simple Knowledge Organization System*, www.w3.org/2004/02/skos/) is such a format. SKOS basically contains the taxonomic relations 'narrower term' and its inverse 'broader term', 'related term'. Each term can have multiple labels to express the human readable interpretation. This allows us to use multiple synonyms or translations for a single concept. As SKOS is based on the RDF-format (https://www.w3.org/RDF/), each term has a unique identifier, and the relations between terms are expressed as triples: subject-predicate-object. This allows software systems to recognize terms unambiguously, and move from one triple to another. For example: (1) Peter – has broader term – writer, (2) Peter – relates to – Peter's biography. The expressive power of controlled vocabularies is less than that of ontologies, but on the other hand they are more flexible.
A well-known controlled vocabulary in the agricultural domain is AGROVOC (https://www.fao.org/agrovoc/), which was developed by FAO. It contains about 1M terms. Interestingly, WUR has assisted FAO in the early stage of the development of Agrovoc. This vocabulary is rather generic, and to get a collection of more specific terms, the Valerie vocabulary was developed by WUR in the Ask-Valerie project (http://www.foodvoc.org/page/Valerie-9, about 10k terms).

Until now, the construction of controlled vocabularies is a manual process in which a number of domain experts collectively set a shared target, collect and organize terms, and express them in the SKOS format. See for example https://boxesandarrows.com/creating-a-controlled-vocabulary/ for a description of this process. There are several tools for constructing SKOS-based vocabularies, such as Protégé (https://protege.stanford.edu/), TopBraid (https://www.topquadrant.com/resources/skos-xl-taxonomies-in-topbraid-edg), Swoop (http://www.mindswap.org/2004/SWOOP/), VocBench (http://vocbench.uniroma2.it), Semantic Turkey (http://semanticturkey.uniroma2.it/), PoolParty, (http://www.poolparty.biz/), TemaTres, (http://www.vocabularyserver.com/) [1,2]. However, these tools require some technical skills. Therefore, we have developed the ROC+ tool. We will describe ROC+ further in section 2.2.

Furthermore, recent developments in NLP-technology (Natural Language Processing, https://en.wikipedia.org/wiki/Natural_language_processing) can assist human experts by proposing terms, and even the relations, by extracting these from domain-specific documents. As a first step, before proceeding manually in ROC+, the TALK-tool can be used for this purpose, even though it was originally developed for a slightly different purpose. We will explain this in the next section.

2.1 TALK Tool

Given the diversity of experts and stakeholders in research areas such as that of Wageningen University and Research, confusion and misunderstanding of the terminology can easily arise. To assist researchers in creating a common understanding of the terms they use when working together in a project, we have developed the TALK-tool (see https://talk-tool.containers.wur.nl/) in the KB 35 programme on Food Security and Valuing Water – using the Food System Approach. TALK stands for *Team Associations for Linking Knowledge*. One of the objectives of the TALK-tool is to automatically suggest related terms when the user enters a base term, as shown in Figure 4. It is a simple game-like tool that facilitates the discussion between researchers in an early phase of their collaboration. The tool inspires the participants and react to the automatically generated associations rather than having to come up with their own related terms from scratch. It can also broaden the scope of the discussion by suggesting terms that the participants would not have considered themselves, or assist in setting the focus of a project.

TALK uses NLP (Natural Language Processing) methods to generate term associations. We have trained a neural network model to learn vector representations of words (Word2Vec model, https://en.wikipedia.org/wiki/Word2vec). This model takes a single word as input and transforms it into a n-dimensional vector of continuous numbers. The model is trained in such a way that two words that often occur in close proximity to each other in the training data are transformed into vectors that have a high cosine similarity. It is shown that when using this method, a high cosine similarity corresponds to a certain extent to the words being semantically similar.

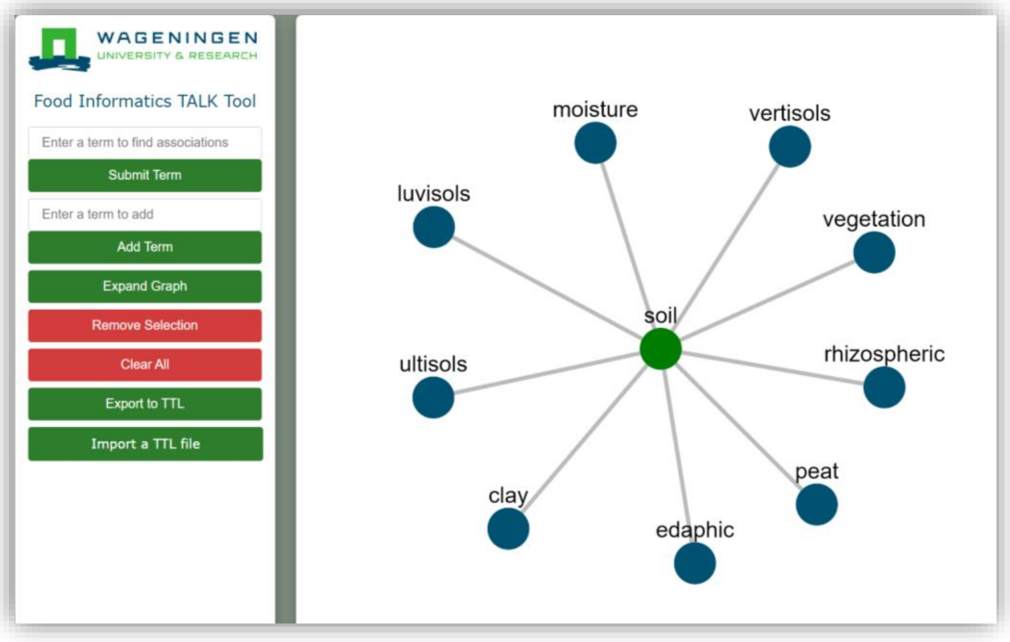


Figure 4 The TALK-tool with example associations for 'soil'

In the TALK-tool, we initially used existing Word2Vec models from Google and Twitter. When evaluating the tool with WUR researchers from different disciplines we learnt that the suggested terms were often to general or clearly out of scope for WUR-type of research. Although it seems to be attractive to use a more domain-specific set of documents as a source of terms, we deliberately want to avoid ‘tunnelling’ in the early phase of getting to know each other and determining the project focus. Moreover, for training a neural network, the set of documents should be sufficiently large. We decide that a proper intermediate solution would be to use the set of publications that cover the entire WUR domain. For this we found a large corpus of documents from approximately 5M abstracts from scientific articles of WUR research domains, provided by the WUR library. The term associations that the trained model will predict are therefore likely to be relevant for the WUR researchers using the TALK tool. However, it might be the case that a user of the TALK tool enters a word that does not occur (frequently enough) in the training corpus. In that case we use the pre-trained Word2Vec model that was trained using generic texts from Twitter as a secondary source.

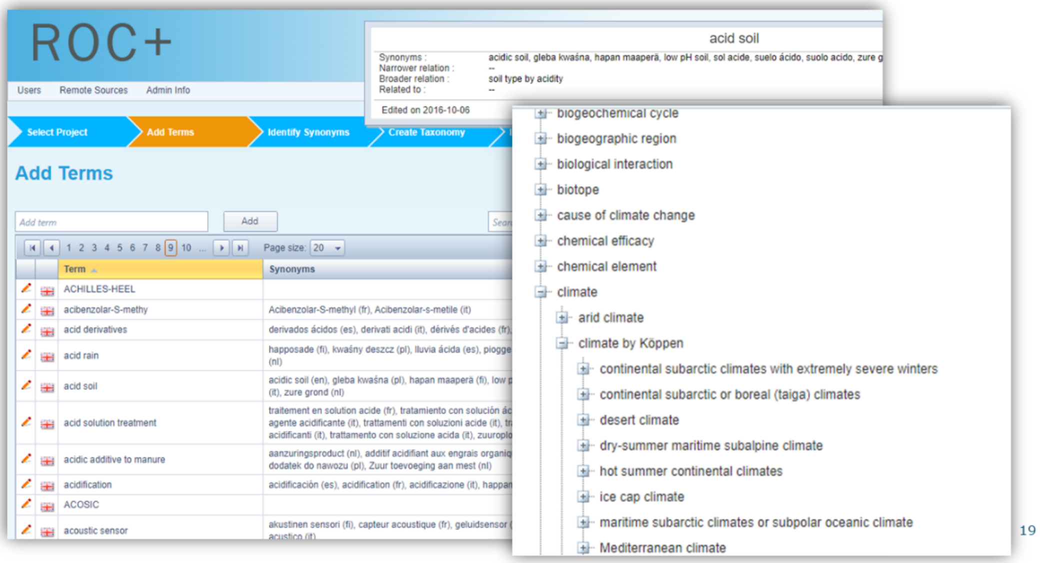
As TALK helps experts to define terminology by automatically suggesting associations, it can also be considered as a first step in creating in controlled vocabulary. It provides functionality to export the constructed association cloud in .ttl format, using the SKOS for representing related terms.

2.2 ROC+

ROC+ (Rapid Ontology Creation) is a vocabulary development tool that relies on the manual input of domain experts. It was developed as an alternative to the more technical tools mentioned above, to facilitate domain experts without data modelling expertise. The ROC+ tool supports collaborative vocabulary development and generates (and accepts) vocabularies in SKOS format in multiple languages.

In ROC+, a team of domain experts starts by setting the scope and depth of the domain knowledge to be modelled, taking into account the applications it should cover. After defining a new ROC+ project, they list relevant terms in a specific language, inspiring each other and discussing their inputs. The initial list is then extended by defining synonyms and, if needed, translations. Synonyms can be terms that were already added or new terms. Once a substantial set of terms has been created, they are organised in a tree structure (taxonomy), from general to specific. It is possible to have more than one broader term for each concept. Finally, terms that are related in any other sense can be connected by defining 'related terms'. This process is run in several cycles, until the domain and applications are sufficiently covered. At any point the vocabulary can be exported as SKOS file. Figure 5 gives an impression of some ROC+ screens, including a concept tree on the right hand side.

With the TALK tool, the first step of this process becomes easier: it automatically generates suggestions for terms. So TALK can be seen as an even more lightweight entry point for ROC+.



acid soil

Synonyms : acidic soil, gleba kwaśna, hapan maaperä, low pH soil, sol acide, suelo ácido, suolo acido, zure g
Narrower relation : --
Broader relation : soil type by acidity
Related to : --

Edited on 2016-10-06

Add Terms

Add term

1 2 3 4 5 6 7 8 9 10 ... Page size: 20

Term	Synonyms
ACHILLES-HEEL	
acibenzolar-S-methy	Acibenzolar-S-methyl (fr), Acibenzolar-s-metile (it)
acid derivatives	derivados ácidos (es), derivati acidi (it), dérivés d'acides (fr),
acid rain	happosade (fi), kwaśny deszcz (pl), lluvia ácida (es), piogge (nl)
acid soil	acidic soil (en), gleba kwaśna (pl), hapan maaperä (fi), low p (it), zure grond (nl)
acid solution treatment	traitement en solution acide (fr), tratamiento con solución ác agente acidificante (it), trattamenti con soluzioni acide (it), tr acidificanti (it), trattamento con soluzione acida (it), zuuroplo
acidic additive to manure	aanzuringsproduct (nl), additif acidifiant aux engrais organiq dodatek do nawozu (pl), Zuur toevoeging aan mest (nl)
acidification	acidificación (es), acidification (fr), acidificazione (it), happan
ACOSIC	
acoustic sensor	akustinen sensori (fi), capteur acoustique (fr), geluidsensor acustico (it)

- biogeochemical cycle
- biogeographic region
- biological interaction
- biotope
- cause of climate change
- chemical efficacy
- chemical element
- climate
  - arid climate
  - climate by Köppen
    - continental subarctic climates with extremely severe winters
    - continental subarctic or boreal (taiga) climates
    - desert climate
    - dry-summer maritime subalpine climate
    - hot summer continental climates
    - ice cap climate
    - maritime subarctic climates or subpolar oceanic climate
    - Mediterranean climate

Figure 5 Impression of some ROC+ screens

2.3 NLP for automated term extraction

The terms and concepts in controlled vocabulary and ontologies can also be generated through knowledge extraction by text mining techniques. One of artificial intelligence techniques (AI) called Natural Language Processing (NLP) can be used in text mining applications. NLP is an application area of AI that made rapid progress in the last decade, driven by advances in deep learning [3], delivering specialized techniques with large performance benefits on NLP tasks, such as embeddings (e.g. word2vec [4], bloom embeddings, floret vectors) and transformer [5] models (e.g. BERT, GPT).

Some tasks performed by NLP are, e.g., named entity recognition (NER), relationship extraction, information extraction (IE) and entity linkage (EL) [6]. The algorithms to perform these tasks help to identify entities with specific meanings in the text, such as names of people, places, institutions, proper nouns, etc. [7] and they are used for mapping mentions in text to corresponding entities in knowledge base (KB) [8]. Recently, machine learning algorithms were shown to be capable of supporting the task of 'entity linking' (e.g. see Spacy example of Wikipedia entity linker [9]), making it much more efficient to link sources of data and knowledge on a larger scale.

Automated knowledge extraction using NLP generates an output that contains entities and co-occurrences of those entities in a specific domain or subdomain. Those entities, namely concepts, can then be used to construct controlled vocabulary or ontologies for this particular domain.

3 Controlled vocabularies at WUR

Sometimes existing vocabularies do not cover specific applications. As listed below, there is a number of controlled vocabularies that were developed at Wageningen University and Research.

Valerie: The EU funded Valerie project aimed to improve the accessibility and availability of new knowledge for innovation in agriculture and forestry (www.valerie.eu). A smart search engine, ask-Valerie.eu (https://www.ask-valerie.eu/) and repository of structured information was developed to interactively provide information to farmers, agricultural organisations and researchers. The controlled vocabulary for ask-Valerie was built by several domain experts using ROC+ software. It contains 10391 concepts including 'broader', 'narrower' and 'related to' terms and translations in 9 different languages.

FAIRshare: The FAIRshare project aims to support farm advisors to support a more productive and sustainable agriculture (https://www.h2020fairshare.eu/about-fairshare/). In this project, a search engine was developed to give access to collections of descriptions of Digital Advisory Tools and Services (DATS) (https://fairshare-pnf.eu/). A query expander was built using terms listed in a dedicated controlled vocabulary. This query expander is expected to improve search by finding more relevant DATS. The controlled vocabulary currently contains approximately 600 concepts on farming and will be extended in 2023.

The Wageningen Model Gallery (https://modelgallery.wurnet.nl/): The search engine of this repository of models created by WUR researchers is using a controlled vocabulary consisting of 945 concepts in Dutch and English, covering in all domains of Wageningen University and Research.

Shared Research Facilities (www.sharedfacilityfinder.com): This search engine applies a controlled vocabulary that consists of about 1500 concepts in English regarding research equipment, analysis methods, research techniques and analytes.

4 Conclusion

Searching text or documents in large unstructured and semi-structured data sources is not trivial. Controlled vocabularies can be used to represent domain knowledge and guide users in expressing their need for information. Several tools are available for developing such vocabularies. We have highlighted the TALK-tool and ROC+ as easy-to-use solutions for this purpose, not requiring domain experts to have specific technical skills. In addition, NLP techniques can give further assistance in creating vocabularies but also factual information from textual documents. Applying NLP for all kinds of purposes should be intensified at WUR. Finally, with the appearance of more publicly available and well-maintained vocabularies and ontologies on the web, we should also start reusing those common resources.

Literature

1) Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Van Gemert, W., Dechandon, D., ... & Keizer, J. (2020). VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. Semantic Web, 11(5), 855-881.

2) Conway, M., Khojoyan, A., Fana, F. et al. Developing a web-based SKOS editor. J Biomed Semant 7, 5 (2016). https://doi.org/10.1186/s13326-015-0043-z.

3) Bokka, K. R., Hora, S., Jain, T., & Wambugu, M. (2019). Deep Learning for Natural Language Processing: Solve your natural language processing problems with smart deep neural networks. Packt Publishing Ltd.

4) Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

5) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

6) Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

7) Wen, Y., Fan, C., Chen, G., Chen, X., Chen, M. (2020). A Survey on Named Entity Recognition. In: Liang, Q., Wang, W., Liu, X., Na, Z., Jia, M., Zhang, B. (eds) Communications, Signal Processing, and Systems. CSPS 2019. Lecture Notes in Electrical Engineering, vol 571. Springer, Singapore. https://doi.org/10.1007/978-981-13-9409-6_218.

8) Fang, Z., Cao, Y., Li, R., Zhang, Z., Liu, Y., & Wang, S. (2020, April). High quality candidate generation and sequential graph attention network for entity linking. In Proceedings of The Web Conference 2020 (pp. 640-650).

9) See https://en.wikipedia.org/wiki/SpaCy and https://github.com/egerber/spaCy-entity-linker.

To explore the potential of nature to improve the quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
www.wur.eu/wfbr
E info.wfbr@wur.nl

Report 2383

Public

The mission of Wageningen University and Research is “To explore the potential of nature to improve the quality of life”. Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 5,000 employees and 10,000 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.