How Agricultural Digital Innovation Can Benefit from Semantics: The Case of the AGROVOC Multilingual Thesaurus [†]

Esther Mietzsch ^{1,*}, Daniel Martini ¹, Kristin Kolshus ^{2,‡} , Andrea Turbati ^{2,‡} and Imma Subirats ^{2,‡}

¹ Association for Technology and Structures in Agriculture (KTBL), Bartningstr. 49, 64289 Darmstadt, Germany; d.martini@ktbl.de

² Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla, 00153 Rome, Italy; Kristin.Kolshus@fao.org (K.K.); turbati@info.uniroma2.it (A.T.); Imma.Subirats@fao.org (I.S.)

* Correspondence: e.mietzsch@ktbl.de; Tel.: +49-6151-7001-230

[†] Presented at the 13th EFITA International Conference, online, 25–26 May 2021.

[‡] Group Email of Food and Agriculture Organization of the United Nations: agrovoc@fao.org.

Abstract: AGROVOC is the multilingual thesaurus managed and published by the Food and Agriculture Organization of the United Nations (FAO). Its content is available in more than 40 languages and covers all the FAO's areas of interest. The structural basis is a resource description framework (RDF) and simple knowledge organization system (SKOS). More than 39,000 concepts identified by a uniform resource identifier (URI) and 800,000 terms are related through a hierarchical system and aligned to knowledge organization systems. This paper aims to illustrate the recent developments in the context of AGROVOC and to present use cases where it has contributed to enhancing the interoperability of data shared by different information systems.

Keywords: AGROVOC; linked open data; FAIR principles; thesaurus; controlled vocabulary; food and agriculture; collaborative editing; multilinguality

Citation: Mietzsch, E.; Martini, D.; Kolshus, K.; Turbati, A.; Subirats, I. How Agricultural Digital Innovation Can Benefit from Semantics: The Case of the AGROVOC Multilingual Thesaurus. *Eng. Proc.* **2021**, *9*, 17. <https://doi.org/10.3390/engproc2021009017>

Academic Editors: Dimitrios Aidonis and Aristotelis Christos Tagarakis

Published: 24 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Since the 1980s, the Food and Agriculture Organization of the United Nations (FAO) has managed and published the AGROVOC Multilingual Thesaurus. It covers all of the FAO's areas of interest, mainly food, nutrition, agriculture, forestry, fisheries, and environment, as well as scientific and common names of organisms, biological notions, techniques of plant cultivation and environmental research, and other related subjects in more than 40 languages. Currently, around 25 national and international organizations worldwide support AGROVOC by contributing to the editorial community. These organizations provide language coverage by supplying terms in a specific language, and support thematic coverage by sharing their expertise in a specific discipline. AGROVOC also incorporates three specialized subsets: LandVoc, maintained by the LandPortal Foundation for concepts related to land governance; ASFA for aquatic sciences and fisheries; and FAOLEX for legislative and policy concepts in FAO areas.

2. AGROVOC as Linked Open Data

AGROVOC is available online as a linked open data set. The structural basis is the RDF [1] and the SKOS [2]. AGROVOC has more than 39,000 concepts formalized as SKOS concepts, identified by dereferenceable URIs. Each concept has at least one preferred term in a language. Optionally, a concept can have alternative or non-preferred terms. AGROVOC uses the SKOS extension for labels: SKOS-XL, which treats labels as full resources, thus allowing to assign to them further properties alongside the pure label text string. The predicates used are as follows: skosxl:prefLabel, used for preferred terms, and skosxl:altLabel, used for alternative or non-preferred terms.

The concepts are related through a hierarchical system based on skos:broader and skos:narrower and aligned to other vocabularies and thesauri mainly by skos:exactMatch and skos:closeMatch. Using this basis, AGROVOC is the key element to produce FAIR data in food and agricultural sciences. The FAIR principles state that data must be findable, accessible, interoperable, and reusable [3] and formulate measures to be taken to achieve these goals. AGROVOC fulfills these recommendations in itself, and thus its use in data sets and services contributes to the fulfillment of principle I2: (meta)data use vocabularies that follow FAIR principles. This is to support knowledge discovery and innovation, data and knowledge integration, and promote sharing and reuse of data, which focuses more on data intensive research and sharing data across the data value chain [4].

3. AGROVOC Ecosystem

The AGROVOC technical infrastructure is based on a comprehensive ecosystem of tools for users and editors to provide access to the data for both humans and machines. AGROVOC is a stable and reliable resource, which is continuously expanded by the activity of the curators and the editorial community. Improvements in the underlying technology have led to improvements in content representation. The Skosmos search interface allows users to search for terms and to browse through the hierarchy tree [5]. The data are also available in machine-readable formats like RDF/XML, Turtle, and JSON-LD. For targeted querying, selection and extraction of subsets, a public query endpoint is available [6]. It allows use of the SPARQL query language for semantic data [7] on the AGROVOC data set and includes sample queries. The results of the queries can be displayed in tables and can be downloaded as files in a number of formats, including comma-separated value (CSV).

Access for editors (who edit terms, add definitions, add concepts, and so on) is provided via a dedicated web access tool called VocBench [8]. Apart from providing a form-based frontend for data entry, VocBench supports the editorial workflow by offering features such as history, automatic capturing of modification metadata based on Dublin Core [9], and role-based validation of new entries and changes. Editorial rules and guidelines have been defined and are continuously evaluated and revised to facilitate the work delivered by the multilingual and distributed community of editors [10]. Currently, AGROVOC publishes new releases on a monthly basis.

4. AGROVOC Use Cases

AGROVOC can be used to enable agricultural digital innovation in different ways; i.e., linking AGROVOC concept URIs from data sets or other resources like bibliographic records and annotations of research data and text corpora allows users to unambiguously define concepts. Reliance on this common URI set implicitly links all these resources to each other, effectively integrating them into a global, interoperable data space. Apart from data integration, concept labels available in multiple languages support the internationalization of applications and information systems. Possible areas of application include the following:

- Organization of knowledge for subsequent data retrieval;
- Metadata annotation of agricultural (research) data;
- Standardization of agricultural information data and services;
- Indexing of literature;
- Auto-tagging/annotation of text corpora and web sites;
- Thesaurus for international cooperation (translation purposes);
- Multilingual search engine discovery.

A current use case is the project HortiSem [11], where an agricultural advisory system focusing on horticulture is enhanced by semantic web technology. Data relevant for planning of pest control measures and developing plant disease management strategies are currently spread across different heterogeneous data sources. Currently, three of them have been prioritized to be dealt with in the project and to be semantically annotated, enriched, and integrated into a knowledge graph. One is the German Federal Office of

Consumer Protection and Food Safety (BVL) registered pesticides relational database, which contains data on active ingredients, crops, pests, and pesticides, alongside their most important attributes like temporal and spatial application restrictions and relations, such as which pesticide can be used on which combination of crop and pest (so-called indications). Another source is the EU Pesticides Database on maximum residue levels (MRLs). The MRL data are available in XML format. Finally, a large text corpus comprising agricultural advisory alerting newsletters is used as a large sample of free-text unstructured data.

Each of these sources undergoes a preparatory processing step to convert (part of) the data into RDF and load them into a Fuseki Triple Store to create an integrated knowledge graph. Structured data sources are converted to RDF using specific tools, namely db2triples for the database and an XML2RDF converter for the XML files. Unstructured data (texts) are processed using the SpaCy toolkit for named entity recognition/linking. Using that, automatically generated entity annotations are produced and integrated into the graph using the Web Annotation Vocabulary [12]. Generating these annotations is currently a work in progress, but, depending on concept availability, they point on the one hand to entity occurrences in the documents, and on the other hand to AGROVOC concepts or additional concepts for pest control products derived from the BVL database data. Approaches to prepare the required knowledge base for the named entity linker with subsets of AGROVOC are currently being developed. All of the data sets processed are then linked and mapped among the data sources, as well as to AGROVOC as a central concept hub (see Figure 1). This merged data set is then accessible by several applications, among them three horticultural information portals. AGROVOC here is both a set of concepts and a hub to the global data space built by a multitude of vocabularies and knowledge systems, which are linked by alignments in AGROVOC.

Figure 1. HortiSem overall architectural model.

5. Outlook on Further Development

In 2010, a set of properties describing agricultural and biological concept relations to be used in AGROVOC were devised and specified in the Agrontology. While that approach could have led to a semantic enrichment of AGROVOC, in practice, only some of these properties have been considered useful and others have been rarely or inconsistently used. An ongoing activity is thus revising the Agrontology, deprecating superfluous or unclear properties, better documenting the remaining ones, and giving guidance for editors on how to use it. Apart from using ontology relations in AGROVOC, approaches to using AGROVOC concepts in external ontologies are also addressed in initiatives with partners like CGIAR. Further work includes improvements to data quality and coverage

like adding more language, enhancing language coverage, and closing thematic gaps in close collaboration with specific expert communities like fisheries or land governance.

Author Contributions: Writing—original draft preparation, E.M.; writing—review and editing, D.M., K.K., A.T. and I.S.; supervision, I.S.; project administration, I.S.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: HortiSem is supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme (funding number: 2818508B18).

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: <https://agrovoc.fao.org/browse/agrovoc/en/> and <https://hortisem.de/> (accessed on 1 June 2021).

Acknowledgments: The authors acknowledge the work of the 22 institutions in the AGROVOC Editorial Community.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cyganiak, R.; Wood, D.; Lanthaler, M. RDF 1.1 Concepts and Abstract Syntax. World Wide Web Consortium (W3C). 2014. Available online: <https://www.w3.org/TR/rdf11-concepts/> (accessed on 7 June 2021).
2. Miles, S. Bechhofer. SKOS Simple Knowledge Organization System Reference. World Wide Web Consortium (W3C). 2009. Available online: <https://www.w3.org/TR/skos-reference/> (accessed on 8 June 2021).
3. The Fair Data Principles. Available online: <https://www.force11.org/group/fairgroup/fairprinciples> (accessed on 1 June 2021).
4. FAO. AGROVOC—Semantic Data Interoperability on Food and Agriculture. Rome. 2021. Available online: <https://doi.org/10.4060/cb2838en> (accessed on 1 June 2021).
5. AGROVOC: AGROVOC Multilingual Thesaurus. Available online: <https://agrovoc.fao.org/browse/agrovoc/en/> (accessed on 1 June 2021).
6. AGROVOC SPARQL Endpoint. Available online: <https://agrovoc.fao.org/sparql> (accessed on 1 June 2021).
7. Harris, S.; Seaborne, A. SPARQL 1.1 Query Language. World Wide Web Consortium (W3C). 2013. Available online: <https://www.w3.org/TR/sparql11-query/> (accessed on 8 June 2021).
8. Stellato, A.; Fiorelli, M.; Turbati, A.; Lorenzetti, T.; Van Gemert, W.; Dechandon, D.; Laaboudi-Spoiden, C.; Gerencsér, A.; Waniart, A.; Costetchi, E. VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. *Semant. Web* **2020**, *5*, 855–881. [CrossRef]
9. DCMI Usage Board. DCMI Metadata Terms. 2020. Available online: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/> (accessed on 8 June 2021).
10. FAO. The AGROVOC Editorial Guidelines 2020. Rome. 2020. Available online: <http://www.fao.org/documents/card/en/c/cb2328en> (accessed on 1 June 2021).
11. HortiSem. Available online: <https://hortisem.de/> (accessed on 1 June 2021).
12. Sanderson, R.; Ciccarese, P.; Young, B. Web Annotation Vocabulary. World Wide Web Consortium (W3C). 2017. Available online: <https://www.w3.org/TR/annotation-vocab/> (accessed on 8 June 2021).