Semantic Web for interoperable food safety legislation data: A case study

Carlos Enrique Pintor[1], Carlos Francisco Ragout[1], Diego Torres[1,2], and Alejandro Fernandez[1,3]

[1] LIFIA, Facultad de Informática, UNLP, Calle 50 y 120, La Plata, Argentina.
[2] Depto CyT, Universidad Nacional de Quilmes. Roque Saenz Peña 352, Bernal, Argentina.
[3] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, La Plata, Argentina.

Abstract. Food safety legislation plays a central role in regulating the levels of chemicals used in agriculture practices in order to prevent potential risks to consumers' health within a certain region or country. Public Health organizations publish these regulations as recommendations on allowed quantities of chemicals residues for different types of crops. These documents pose a major challenge for automatic processing as their format is not normalized nor the terminology used is uniform in any way. Semantic Web technology tools offer a solution as these documents may be published as linked data which would allow computers to process them automatically, so that further analysis and interoperability would be possible. In this paper we introduce MRL-O, an ontology for describing data on allowed levels of residues present in commodities of agricultural origin. MRL-O serves as a standardized framework for sharing interoperable data and to provide tracking metadata about its sources and transformation processes. We also describe a step-by-step procedure to obtain MRL-O linked data from real non-normalized documents. Also, we applied this procedure on data published by Argentina and Brazil with promising results. Consequently, we argue that the proposed ontology is sufficient to model the domain of MRL regulation and serves as the basis for tools that support interoperability in this domain.

Keywords: Maximum Residue Limits, Agriculture, Health, Regulation, Semantic Web, Linked Open Data

1 Introduction

Agrochemical substances and its derivatives are used throughout agricultural processes to prevent and control the presence of pests. As a result the products obtained from these practices may contain certain levels of chemical residues potentially harmful to human health. A mechanism to monitor and control the maximum concentration of pesticide residues (MRL) in food commodities is therefore required [7]. Governments and Health Organizations determine and publish recommended values of MRL periodically, as these values have a significant impact in human health [6] and in international food trade [5][8].

Given the lack of official or standardized guidelines regarding how this data should be produced and published, a wide range of methods and supporting media for publishing documents on MRL are used, involving different formats (e.g. pdf, xml, csv, etc.), content types (tables, graphics, lists, etc.) and language. There is no formal curation process on the data itself to prevent inaccurate terms, syntax errors, omissions, synonyms and proprietary data structures.

The diversity in publication formats makes it difficult to process and analyze the datasets by using computers due to incompatibility issues among documents from different sources, or even between versions of the same document. We believe the Semantic Web [1] offers an alternative to address this interoperability challenge.

In this paper we apply Semantic Web technologies and tools to design and create MRL-O (Section 3), a specific ontology to represent MRL-related data. We propose a semantic pipeline (Section 4) to transform non-normalized data into MRL-O semantic datasets ready to be consumed and processed by computers without any regards about formats of origin.

2 Background and related work

The concept of Semantic Web was introduced by Berners-Lee [1] to encompass a set of technologies that provides a better knowledge representation with the use of ontologies, software agents, and logic rules. It is an extension to the World Wide Web where the information is described in a machine-readable format. Data in the Semantic Web is modeled using RDF (Resource Description Framework) [9, Chapter 2]. RDF models are built around web resources and triples.

This work builds upon several existing developments, ontologies and vocabularies.

AGROVOC [3] is a multilingual open dataset about agriculture concepts and relationships which is used to identify resources covering all areas of interest of United Nations FAO (Food and Agriculture Organization).

ChEBI [2] is a database and ontology specialized in small chemical compounds of biological interest developed by the European Bioinformatics Institute. ChEBI has been widely adopted by numerous bioinformatics projects and as ontological reference in several semantic-web projects.

The Units Ontology [4] is also part of the ontology network of MRL-O. It is used to express quantities and proportions of agrochemical components under standard terms.

3 MRL-O

MRL-O (Maximum Residue Limit Ontology) is an ontology that models the domain of MRL regulation. Following the best practices of the Semantic Web, MRL-O borrows elements from other existing ontologies. In particular MRL-O relies on AGROVOC, ChEBI, and the Units Ontology to express the information contained in a single record of an MRL-O dataset. Similarly, it relies on PROV-O, Dublin Core, and Wikidata to describe the process of applying the transformation pipeline (presented in the following section) to datasets published by a given organization.

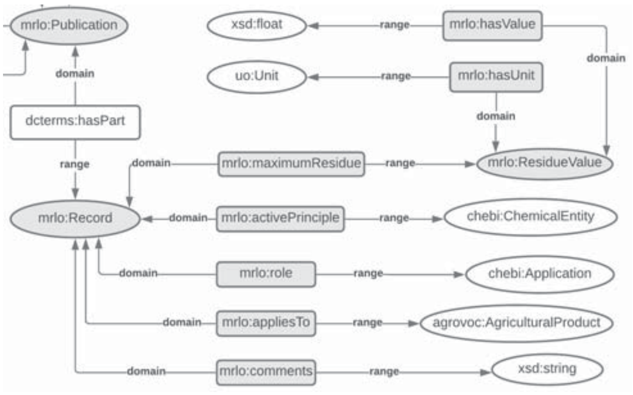

Fig. 1. Part of the MRL-O ontology that represents MRL records

Figure 1 uses the graph notation of RDF to provide an overview of the terms used to express the information contained in a record. The elements with a gray background are those introduced by MRL-O, whereas elements with a white background are adopted from other vocabularies.

4 Transformation pipeline

At the heart of our proposal lies the transformation pipeline. Figure 2 provides an overview showing its main activities namely, Clean, Align, and Transform. Following, we discuss each of these activities with more detail.

Clean

The “Clean” activity takes data files as inputs and produces a single file of clean data rows representing a unique statement involving one crop, one active principle, and one application. Then, normalization takes place to trim blanks, collapse consecutive blanks, and standardize capitalization. The output of this first activity is a clean table, with one row per MRL record, and four columns:

- Active principle: Name of a chemical substance (e.g., 2,4-D)
- Role: Role or usage of the chemical substance (e.g., herbicide)
- Product: Crop or agricultural product or commodity (e.g., tomato)
- MRL: Maximum residue level in mg/Kg (e.g., 0.05)

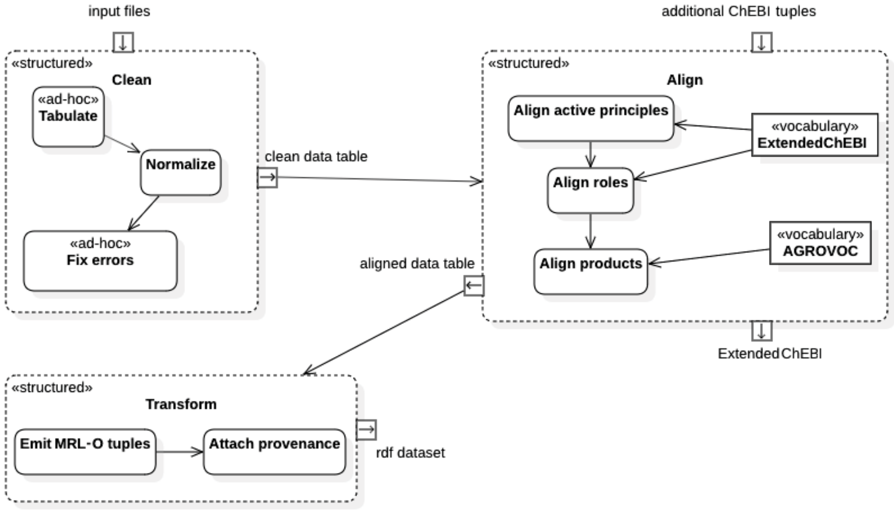

Fig. 2. Semantic transformation pipeline

– Comments: Any comments regarding this record (possibly empty)

Align

The goal of the "Align" activity is to replace all references to chemical substances and to agricultural products or crops in the data with the corresponding resources in the Semantic Web using the previously mentioned reference ontologies.

Transform

The final activity transforms the clean table into an RDF dataset. After all rows in the table have been processed, the process adds provenance information triples for the publication resource. The final result is an MRL-O based dataset.

5 Case study

As a proof of concept we applied our vocabulary and the pipeline on two real world datasets from Argentina and Brazil. These documents were published in 2020, and the pipeline was implemented using OpenRefine.

Regarding the Argentinean case study, we found that the reference document on MRL is published by SENASA as an excel worksheet. On the other hand, the Brazilian government through its national sanitary agency (ANVISA) publishes information regarding phytosanitary legislation on their official website from which a csv file can be downloaded.

5.1 Inter-operable queries

The most interesting part of generating semantic datasets for us was creating usage scenarios where useful information could be extracted. For example, we created a set of SPARQL queries to answer some common questions comparing food legislation in Argentina and Brazil. It is worth mentioning that this exercise, although plausible, would have been extremely complex to achieve without the support of the Semantic Web tools we applied.

6 Conclusions

The two case studies in this article show that MRL-O is rich enough to cover the basic requirements to express meaning within the MRL domain. The results obtained from the pipeline execution proved to be effective as well. The set of sample SPARQL queries shows how simple it is to extract meaningful information from MRL-O data, and that more complex combinations between records are also possible. The idea of having a food safety legislation based upon the Semantic Web is feasible and valuable.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5), 34–43 (2001). Publisher: JSTOR
2. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic acids research **36**(suppl_1), D344–D350 (2007)
3. FAO: AGROVOC. FAO (2021). DOI 10.4060/cb2838en
4. Gkoutos, G.V., Schofield, P.N., Hoehndorf, R.: The Units Ontology: a tool for integrating units of measurement in science. Database **2012**(0), bas033–bas033 (2012). DOI 10.1093/database/bas033
5. Li, Y., Xiong, B., Beghin, J.C.: The Political Economy of Food Standard Determination: International Evidence from Maximum Residue Limits. In: Nontariff Measures and International Trade, *World Scientific Studies in International Economics*, vol. Volume 56, pp. 239–267. World Scientific (2016). DOI 10.1142/9789813144415_0014
6. Li, Z.: Evaluation of regulatory variation and theoretical health risk for pesticide maximum residue limits in food. Journal of Environmental Management **219**, 153–167 (2018). DOI 10.1016/j.jenvman.2018.04.067
7. WHO: Principles and Methods for the Risk Assessment of Chemicals in Food. World Health Organization (2009)
8. Xiong, B., Beghin, J.C.: Stringent maximum residue limits, protectionism, and competitiveness: The cases of the us and canada. In: Nontariff Measures and International Trade, pp. 193–207. World Scientific (2017)
9. Yu, L.: A developer's guide to the semantic web. Springer, Berlin (2011). OCLC: 700066210