Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16285

Adding semantics to tabular agrifood data

Top, J.[1]

[1] Agrotechnology & Food Sciences Group, Wageningen University & Research, Bornse Weilanden 9, 6708WG Wageningen, the Netherlands
Corresponding author's e-mail: jan.top@wur.nl

Problem context

Tabular data are abundant in the agrifood domain, as in many other fields. Tables in databases, spreadsheets and reports are common in business applications, research papers, policy documents, etc. They contain data with a repetitive structure, containing multiple rows of similar data. The header of each column explains to a certain extent the meaning of the numbers, text or graphics in the cells below it. However, in practice these annotations are often incomplete, not clear for outsiders and not machine readable. At the time of registration the data, usually not much attention is being given to providing all required details that are needed for future use. This downgrades the value of otherwise useful data. The FAIR[1] principles have been defined to address the issue of data reuse. In particular it defines the requirement of interoperability, meaning (among others) that (meta)data should use a formal, accessible, shared, and broadly applicable language for knowledge representation. It is clear that such a formalised approach will leverage data reuse, but on the other hand is also hard to realize in practice. Here we summarise work we have done to support semantic annotation of data in the agrifood domain, in a way that minimises the effort needed.

Approach

Semantics of data can be expressed using concepts and relations, defined in RDF-based vocabularies and ontologies[2]. The advantages of this approach are:

- Each concept used in the metadata gets a unique reference point (IRI). Users can use different words, but still refer to the same notion. For example, in different tables one header cell may say 'mass' and another 'weight'. By referring (hyperlinking) to the concept http://www.ontology-of-units-ofmeasure.org/resource/om-2/Mass their express their shared intention.
- We can express not only the basic field names, but also relations between the notions used in the field descriptions. For example, for the concept 'batch of apples' we can express that it has a GTINcode, a species and is produced by a specific grower. How these aspects are related is normally not expressed in table headers, but helps to understand and reason about the data.
- By using RDF (RDFS/OWL) the information becomes machine readable.
- If end-user software, such as Excel or dedicated applications, supports these shared vocabularies, awareness about the benefit of proper annotation grows as well as the readiness actually record this information.

Existing solutions for semantic tables typically view all columns in a table as the properties of a single phenomenon. However, the structure of tables is more subtle than that. Some columns are collectively used to identify one object (e.g., product) and other columns identify another object (e.g., producer). Still

[1] https://www.go-fair.org/fair-principles/

[2] https://www.w3.org/RDF/

other columns describe properties of these objects (e.g., hardness), or of the further unspecified environment (e.g., surrounding temperature).

With RDF Record Table [2], we have addressed the problem of annotating arbitrary table structures. We identify two types of columns in a table:

- Columns that identify phenomena (typically objects/artefacts, materials or events/processes)
- Columns that specify properties, often in terms of physical quantities and units, but also applying qualitative scales (nominal or ordinal).

For example, 'product', 'producer' and 'parcel' are typically phenomena, whereas 'price', 'harvesting time' and 'parcel area' would be properties. This distinction allows users to see to which vocabularies or ontologies they can link.

With RDF Record Table we have proposed an alternative to RDF Data Cube[3] , which is not optimal in handling irregular tables. In RDF Record Table we also have devised options to reduce the size of data files, which can easily grow due to the redundancy that comes with semantic annotation.

In order to show the feasibility of semantic annotation with RDF Record Table, we have implemented this model as an add-on to Excel called Rosanne. Rosanne allows users to link table cells to shared concepts. This helps user to annotate data, find data, perform unit conversion, select subsets and merge data from different tables.

Annotation sources

The next question is which vocabularies or ontologies one can use as references for expressing semantics of datasets. For this purpose, we have coined two solutions. First, we have created OM, Ontology of Measurement, to formalise a large amount of quantities and units in RDFS/OWL. OM has been rated as the most extensive an detailed ontology in this area [3], and is publicly available on the web[4].

Secondly, for annotating phenomena, often a more lightweight vocabulary is sufficient. Although the preferred practice is to reuse existing models, such as Agrovoc[5] in the agriculture domain, in practice always specific choices have to be made as to which terms to use and how to structure the them. Most of the existing tools require knowledge engineering expertise. Therefore, we have developed ROC+ [4] as a method and tool for non-IT domain experts. With ROC+, experts can collaboratively create a multilingual vocabulary. ROC+ sessions are conceived by experts as pleasant and effective to create consensus about the terms they use in daily practice.

One example of a vocabulary created with ROC+ is Valerie[6]. The thousands of terms related to agriculture and forestry in this vocabulary have been collected and organised by several domain experts in the field with the European Valerie project.

References

[1] Wilkinson, Mark D. et al, 'The FAIR Guiding Principles for scientific data management and stewardship', Scientific Data, vol. 3, article number: 160018, 2016

[2] Wigham, M., Rijgersberg, H., de Vos, M.G., Top, J.L. , 'Semantic Support for Tables using RDF Record Table', International Journal on Advances in Intelligent Systems, Vol. 8, No. 1&2. ISSN 1942-2679, 2016.

[3] Markus D. Steinberg, Sirko Schindler and Jan Martin Keil, 'Use Cases and Suitability Metrics for Unit Ontologies', in: OWL: Experiences and Directions – Reasoner Evaluation: OWLED 2016, and ORE 2016, Bologna, Italy, pp.40-54, 2016.

[4] Koenderink, N.J.J.P., Assem, M. van, Hulzebos, J.L., Broekstra, J., Top, J.L., 'ROC : a method for protoontology construction by domain experts', Lecture Notes in Computer Science 5367, p. 152 – 166. 2008

[3] https://dvcs.w3.org/hg/gld/raw-file/29a3dd6dc12c/data-cube/index.html
[4] http://www.foodvoc.org/page/om-2
[5] http://aims.fao.org/vest-registry/vocabularies/agrovoc
[6] http://www.foodvoc.org/page/Valerie-9