A Use Case of Data Integration in Food Production

Liliana Ibanescu[1]*, Thomas Allard[2], Stéphane Dervaux[1], Juliette Dibie[1], Elisabeth Guichard[2], Caroline Pénicaud[3], and Joe Raad[1]

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France
[2] CSGA, AgroSupDijon, CNRS, INRA, Université Bourgogne Franche-Comté, 21000, Dijon, France
[3] UMR GMPA, AgroParisTech, INRA, Université Paris-Saclay, 78850, Thiverval-Grignon, France

Abstract. This paper presents a use case about knowledge representation and integration of data from different domains in food science. An ontology named PO^{2}DG, the Process and Observation Ontology for the production of Dairy Gels, has been designed in order to provide a shared vocabulary for domain experts. The available data have been semantically structured using PO^{2}DG and are stored in an RDF repository named PO^{2}DG_dataset. This use case identifies some of the challenges when dealing with a multi domain representation problem, gives some hints about possible solutions and suggests some further work.

Keywords: process representation, experimental observations representation, food science, ontology based data integration

1 Introduction

One of the most significant current discussions in food production is how to formulate food products with high nutritional and sensory values and low environmental impact. This supposes being able to build a decision support tool combining data and knowledge from different domains in food science (e.g. nutrition, sensory and perception, eco-design, microbiology, biochemistry, process engineering) with data and knowledge in environmental analysis.

The aim of the ongoing NutriSensAl project involving experts from INRA, the French National Institute for Agricultural Research, is to propose a proof of concept for a decision support tool allowing the formulation of well-balanced products in terms of nutritional requirements (e.g. less fat, sugar and salt) with acceptable sensory qualities for the consumer while using eco-friendly production processes. While a lot of data have been collected in different collaborative

* This work was supported by the Qualiment Carnot Institute of the French National Research Agency through the NutriSensAl project (grant number 16CARN002601) and by the Saclay Data Center through the LIONES project.

research projects, a major problem is the lack of cross domain studies fully combining nutritional and sensory properties with eco-design processes.

The knowledge representation task of the NutriSensAl project focuses on the production process of French hard cheeses, while studying consumer sensory perception. In order to address data and knowledge integration, a relevant solution is the use of an ontology. Therefore, PO^{2}DG, the Process and Observation Ontology for the production of Dairy Gels, has been designed in order to provide a shared vocabulary for the different domain experts involved in the project. The ontology has been developed using Scenario 6 of the NeON methodology [1], i.e. reusing, merging and re-engineering ontological resources. PO^{2}DG was implemented in OWL 2, is available on AgroPortal and some of its concepts are aligned with concepts from foundational ontologies or concepts from available ontologies on the LOD (Linked Open Data) cloud [2]. An RDF repository PO^{2}DG_dataset stores the available data of the NutriSensAl project. The definition of a common vocabulary and the use of Semantic Web technologies should facilitate access to other similar data in order to reuse and integrate them in the future NutriSensAl decision support tool.

The paper is organized as follows. In Section 2 the context of the NutriSensAl project and the available data are presented. Section 3 gives details about the data and knowledge representation task. In Section 4 we give some comments about our experience during this project, and, finally, we conclude in Section 5 and present our further work.

2 NutriSensAl Project

The NutriSensAl project involves experts from INRA, and its goal is to propose a proof of concept for a decision support tool allowing the formulation of well-balanced products in terms of nutritional requirements with acceptable sensory qualities for the consumer while using eco-friendly production processes.

Available data are mainly research data, collected during projects studying real or model cheeses and focussing on specific parameters. Those projects involved domain experts from different domains in food science (see [3] for details):

1. the production process of French hard cheeses. Different parameters were measured during the production steps, e.g. the composition of the milk (i.e. lipid, protein, lactose and water content) or the pH of the product during each step.
2. the sensory perception during in-mouth food breakdown. The main attributes for texture are Springiness, Firmness, Granularity, Hardness and Moisture, while the attributes for taste are Taste intensity, Salty, Sour and Sweet.
3. the rheological properties of cheeses, e.g. The Young modulus.
4. the life cycle assessment for the production of stabilized micro-organisms.

3 Data and Knowledge Representation Task

Each domain presented in the previous section has its specific concepts and terms, and the domain experts from these different domains need to share the same vocabulary in order to integrate available data of the NutriSensAl project. Therefore an ontology named PO^{2}DG has been developed using Scenario 6 of the NeON methodology [1], i.e. reusing, merging and re-engineering ontological resources. It extends PO^2 [4], a core ontology for process and observation, which reuses BFO [5], IAO [6], OM [7], and naRyQ ontology [8]. PO^{2}DG reuses concepts from GACS [9], AgroVoc [10] and NALT [11].

The core ontology PO^2 and the domain ontology PO^{2}DG are both implemented in OWL 2 [12]. The core ontology PO^2 contains 90 classes, 122 properties and 15 individuals. The PO^{2}DG ontology contains 3475 classes, 122 properties and 6760 individuals.

Both ontologies are available on the the AgroPortal repository, `http://agroportal.lirmm.fr/ontologies/PO2` and `http://agroportal.lirmm.fr/ontologies/PO2_DG`, and both are under the licence Creative Commons Attribution International 4.0 International (CC BY 4.0) [13].

A tool, PO^2`_VocabularyManager`, was developed in order to help build PO^{2}DG by first performing a syntactic search in a set of existing resources, then by assisting the user in the concept creation task. This new tool reduces the time necessary for building the domain ontology and allows a more efficient contribution of the domain experts in this phase of building the domain ontology.

Available data concerning the ongoing NutriSensAl project is now under the process of integration into the PO^{2}DG`_dataset` RDF repository available on `http://sonorus.agroparistech.fr:7200/`.

The following steps were needed for data integration:

1. the knowledge engineer defined a set of EXCEL files structured according to the PO^{2}DG ontology;
2. the domain experts were trained how to fill in these EXCEL files and a guideline was written for the domain experts;
3. a script was written by the computer scientists in order to lift data from EXCEL files into the RDF database.

Our experience shows that the second step is very difficult, error-prone and time-consuming because of the complexity of the data. A second tool, PO^2`_DataManager`, is under development and will be available soon. It will assist the domain experts in integrating data in a more user-friendly way.

Fourteen projects are now stored in the PO^{2}DG`_dataset` repository with 358 processes belonging to four distinct process types. 180 cheese samples are stored in PO^{2}DG`_dataset`.

4 Discussion

In [3] five competency questions concerning the NutriSensAl project are presented along with the corresponding SPARQL queries. Trying to answer cross-

domain questions shows that there was not enough data. When trying to estimate missing data using the available one, very important domain questions arise: i) what method can be used for the estimation? ii) what available data should be used?, and iii) what are 'similar' data?. Domain experts need to find their answers to these questions and knowledge engineers can help.

The question *what are 'similar' data?* is also very challenging in computer science and is investigated more and more in the context of linked data [14, 15]. The need to define 'similar' products and 'similar' processes in NutriSensAl project inspired in [16] the definition of a contextual identity link and an algorithm for its detection in a knowledge base. Preliminary results on the `PO²DG_dataset` show how to use the contextual identity link in order to predict missing observation measures.

5 Conclusion

This paper presents the data representation task of the NutriSensAl project consisting of the design of a domain ontology and the integration of available data into an RDF repository using Web Semantic technologies. The preliminary results illustrate how this semantic approach can be useful to estimate missing data. Further work is to integrate more data into the `PO²DG_dataset` repository and one of the next tasks is to explore how industrial data could be integrated and if open data available on the LOD could be reused.

Another issue we would like to explore is to evaluate the FAIRness [17] of the domain ontology and of the data stored in the RDF repository.

The data and knowledge representation and integration task of the NutriSensAl project represents the first step and the backbone of the decision support tool to be designed and implemented handling multi-criteria indicators. The aim of the NutriSensAl project is to better understand and control the food production process in order to formulate well-balanced products with a low environmental impact.

References

1. Suárez-Figueroa, M.d.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn methodology for ontology engineering. In del Carmen Suárez-Figueroa, M., Gómez-Pérez, A., Motta, E., Gangemi, A., eds.: Ontology Engineering in a Networked World. Springer (2012) 9–34
2. : Linked Open Data. `https://lod-cloud.net/`
3. Pénicaud, C., Ibanescu, L., Allard, T., Fonseca, F., Dervaux, S., Perret, B., Guillemin, H., Buchin, S., Salles, C., Dibie, J., Guichard, E.: Relating transformation process, eco-design, composition and sensory quality in cheeses using PO2 ontology. International Dairy Journal (accepted for publication 2018)
4. Ibanescu, L., Dibie, J., Dervaux, S., Guichard, E., Raad, J.: PO² - A Process and Observation Ontology in Food Science. Application to Dairy Gels. In Garoufallou, E., Coll, I.S., Stellato, A., Greenberg, J., eds.: Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25,

2016, Proceedings. Volume 672 of Communications in Computer and Information Science. (2016) 155–165
5. : Basic Formal Ontology (BFO). https://github.com/BFO-ontology/BFO
6. : Information Artifact Ontology (IAO). https://bioportal.bioontology.org/ontologies/IAO
7. : Ontology of units of Measure (OM). https://github.com/HajoRijgersberg/OM
8. Buche, P., Dibie, J., Ibanescu, L., Soler, L.: Fuzzy web data tables integration guided by an ontological and terminological resource. IEEE Trans. Knowl. Data Eng. **25**(4) (2013) 805–819
9. : Global Agricultural Concept Scheme (GACS). http://aims.fao.org/global_agricultural_concept_scheme_gacs
10. : AGROVOC. http://aims.fao.org/vest-registry/vocabularies/agrovoc
11. : National Agricultural Library's Agricultural Thesaurus (NALT). https://agclass.nal.usda.gov/agt.shtml
12. : Web Ontology Language (OWL). https://www.w3.org/2001/sw/wiki/OWL
13. : Creative Commons Attribution International 4.0 International (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/
14. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isn't the same: An analysis of identity in linked data. In Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B., eds.: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. Volume 6496 of Lecture Notes in Computer Science., Springer (2010) 305–320
15. Beek, W., Raad, J., Wielemaker, J., van Harmelen, F.: sameas.cc: The closure of 500m owl: sameas statements. In Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M., eds.: The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings. Volume 10843 of Lecture Notes in Computer Science., Springer (2018) 65–80
16. Raad, J., Pernelle, N., Saïs, F.: Detection of contextual identity links in a knowledge base. In Corcho, Ó., Janowicz, K., Rizzo, G., Tiddi, I., Garijo, D., eds.: Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017, ACM (2017) 8:1–8:8
17. : The FAIR data principles. https://www.force11.org/group/fairgroup/fairprinciples