Towards the Integration of Agricultural Data from Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies

Shufan Jiang[1,2(✉)], Rafael Angarita[1], Raja Chiky[1], Stéphane Cormier[2], and Francis Rousseaux[2]

[1] Institut Supérieur d'Electronique de Paris, LISITE,
28 Rue Notre Dame des Champs, 75006 Paris, France
{shufan.jiang,rafael.angarita,raja.chiky}@isep.fr
[2] Université de Reims Champagne Ardenne, CReSTIC EA 3804,
CReSTIC - UFR Sciences Exactes et Naturelles - Moulin de la Housse - BP 1039,
51687 Reims CEDEX 2, France
{stephane.cormier,francis.rousseaux}@univ-reims.fr

Abstract. Sustainable agriculture is crucial to society since it aims at supporting the world's current food needs without compromising future generations. Recent developments in Smart Agriculture and Internet of Things have made possible the collection of unprecedented amounts of agricultural data with the goal of making agricultural processes better and more efficient, and thus supporting sustainable agriculture. These data coming from different types of IoT devices can also be combined with relevant information published in online social networks and on the Web in the form of textual documents. Our objective is to integrate such heterogeneous data into knowledge bases that can support farmers in their activities, and to present global, real-time and comprehensive information to researchers. Semantic technologies and linked data provide a possibility for data integration and for automatic information extraction. This paper aims to give a brief review on the current semantic web technology applications for agricultural corpus, then to discuss the limits and potentials in construction and maintenance of existing ontologies in agricultural domain.

Keywords: Ontology · Smart agriculture · Internet of Things (IoT) · Semantics · Data integration

1 Introduction

Recent advances in Information and Communication Technology (ICT) aim at tackling some of the most important challenges in agriculture we face today [5]. Supporting the world's current food needs without compromising future generations through sustainable agriculture is of great challenge. Indeed, among

S. Dupuy-Chessa and H. A. Proper (Eds.): CAiSE 2020 Workshops, LNBIP 382, pp. 89–94, 2020.
https://doi.org/10.1007/978-3-030-49165-9_8

all the topics around sustainable agriculture, how to reduce the usage, and the impact of pesticide without losing the quantity or quality in the yield to fulfill the requirement of the growing population has an increasingly important place [6].

Researchers have applied a wide range of technologies to tackle some specific goals. Among these goals: climate prediction in agriculture using simulation models [7], making the production of certain types of grains more efficient and effective with computer vision and Artificial Intelligence [11], soil assessment with drones [14], and the IoT paradigm when connected devices such as sensors capture real-time data at the field level and that, combined with Cloud Computing, can be used to monitor agricultural components such as soil, plants, animals and weather and other environmental conditions [16]. The usage of such ICTs to improve farming processes is known as *smart farming* [18].

In the context of smart farming, **IoT devices** themselves are both data producers and data consumers and they produce *highly-structured data*; however these devices and the technologies we presented above are far from being the only data sources. Indeed, important information related to agriculture can also come from different sources such as official periodic reports and journals like the French Plants Health Bulletins (BSV, for its name in French *Bulletin de Santé du Végétal*)[1], social media such as Twitter and farmers experiences. The goal of the BSV is to: i), present a report of crop health, including their stages of development, observations of pests and diseases, and the presence of symptoms related to them; and ii), provide an evaluation of the phytosanitary risk, according to the periods of crop sensitivity and the pest and disease thresholds. The BSV and other formal reports are *semi-structured data*.

In the agricultural context, **Twitter** -or any other social media- can be used as a platform for knowledge exchange about sustainable soil management [10] and it can also help the public to understand agricultural issues and support risk and crisis communication in agriculture [1]. **Farmer experiences** (aka Old farming practices or ancestral knowledge) may be collected through interviews and participatory processes. Social media posts and farmer experiences are *non-structured data*.

Figure 1 illustrates how this heterogeneous data coming from different sources may look like for farmers: information is not always explicit or timely. Our objective is to integrate such heterogeneous data into knowledge bases that can support farmers in their activities, and to present global, real-time and comprehensive information to researchers and interested parties. We present related work in Sect. 2, our initial approach in Sect. 3 and conclusions and perspectives in Sect. 4.

2 Previous Works

We classify existing works into two categories: information access and management in plant health domain, and data integration in agriculture. In the information access and management in plant health domain category, the *semantic annotation in BSV* focuses on extracting information for the traditional BSV.

[1] https://agriculture.gouv.fr/bulletins-de-sante-du-vegetal.

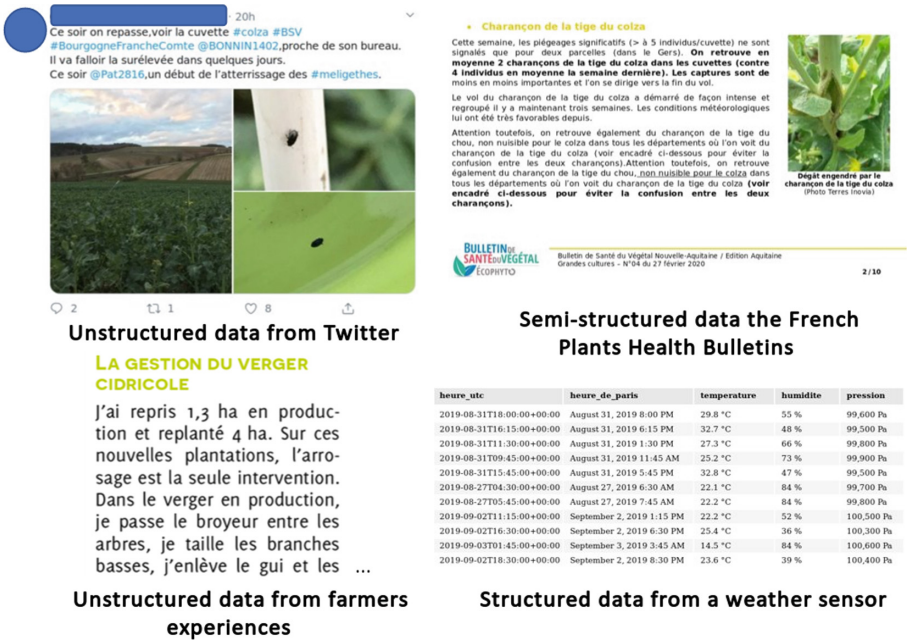

Fig. 1. Heterogeneous sources of agricultural data: *non-structured data* from Twitter and from farmers experiences www.bio-centre.org; *semi-structured data* from The French Plants Health Bulletins; and *structured data* from a weather sensor from www.data.gouv.fr.

Indeed, for more than 50 years, printed plant health bulletins have been diffused by regions and by crops in France, giving information about the arrival and the evolution of pests, pathogens, and weeds, and advises for preventive actions. These bulletins serve not only as agricultural alerts for farmers but also documentation for those who want to study the historical data. The French National Institute For Agricultural Research (INRA) has been working towards the publishing of the bulletins as Linked Open Data [12], where BSV from different regions are centralized, tagged with crop type, region, date and published on the Internet. To organize the bulletins by crop usage in France, an ontology with 272 concepts was manually constructed. With the volume of concepts and relations augmenting, manual construction of ontologies will become too expensive [3]. Thus, ontology learning methods to automatically extract concepts and relationships should be studied.

INRA has also introduced a method to modulate an ontology for crop observation [13]. The process is the following: 1) collect competency questions from researchers in agronomy; 2) construct the ontology corresponding to requirements in competency questions; 3) ask semantic experts who have not participated in the conception of the ontology to translate the competency questions into SPARQL queries to validate the ontology design. In this exercise, a model to describe the appearance of pests was given but not instantiated, nevertheless it could be a reference to our future crop-pest ontology conception.

Finally, *Pest observer* (http://www.pestobserver.eu/) is a web portal [15] which enables users to explore BSV with a combination of the following filters: crop, disease and pest; however, crop-pest relationships are not included. It relies on text-mining techniques to index BSV documents.

Regarding data integration in agriculture, *AGRIS*[2], the International System for Agricultural Science Technology states that many initiatives are developed to return more meaningful data to users [4]. Some of these initiatives are: extracting keywords by crawling the Web to build the AGROVOC vocabulary, which covers all areas of interest of the Food and Agriculture Organization of the United Nations; and SemaGrow [9], which is an open-source infrastructure for linked open data (LOD) integration that federates SPARQL endpoints from different providers. To extract pest and insecticide related relations, SemaGrow uses Computer-aided Ontology Development Architecture (CODA) for RDF triplification of Unstructured Information Management Architecture (UIMA) results from analysis of unstructured content.

Though INRA kick-started categorizing the french crop bulletins using linked open data, and that project SemaGrow shed light upon heterogeneous data integration using ontologies, both projects focused on processing formal and technical documents. Moreover, in CODA application case, IsPestOf rule was defined but not instantiated. Therefore, a global knowledge base, that covers the crops, the natural hazards including pests, diseases, and climate variations, and the relations between them, is still missing. There is also an increasing necessity to a comprehensive and an automatic approach to integrate knowledge from an ampler variety of heterogeneous sources.

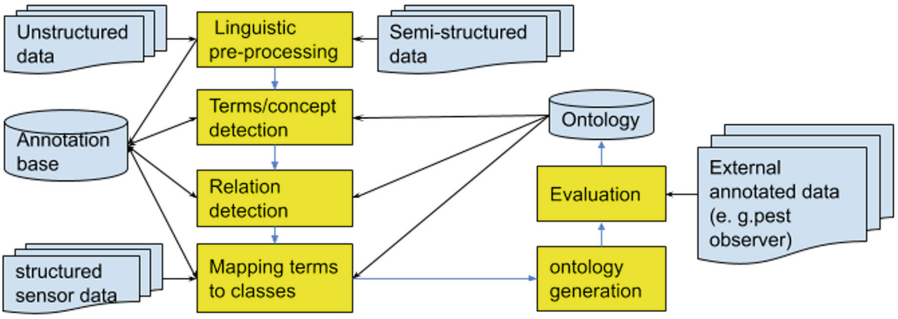

Fig. 2. Our approach for building a phytosanitary knowledge

3 Proposed Design

Figure 2 illustrates our initial design to manage the phytosanitary knowledge from heterogeneous data sources. It consists of a first phase based on ontology learning and a second phase based on ontology-based information extraction:

[2] http://agris.fao.org.

- *Linguistic preprocessing*: Unstructured and semi-structured textual data are passed through a linguistic prepossessing pipeline (Sentence segmentation, Tokenization, Part-of-Speech (POS) tagging, Lemmatization) with existing natural language processing (NLP) tools such as Stanford NLP (https://nlp.stanford.edu/), GATE (https://gate.ac.uk/) and UIMA (https://uima.apache.org/).
- *Terms/concept detection*: At the best of our knowledge and from the state of the art study, there is no ontology in french that modulates the natural hazards and their relations with crops. Existing french thesaurus like french crop usage and Agrovoc can be applied to filter collected data and served as gazetteer. Linguistic rules represented by regular expressions can be used to extract temporal data. Recurrent neural network (RNN), conditional random field (CRF) model and bidirectional long-short term memory (BiLSTM) were applied for health-related name entity recognition from twitter messages and gave a remarkable result [2]. Once the ontology is populated, it could provide knowledge and constraints to the extraction of terms [17].
- *Relation detection*: Similar to term/concept detection, initially there's no ontology. A basic strategy could be using self-supervised methods like Modified Open Information Extraction (MOIE): i) use wordnet-based semantic similarity and frequency distribution to identify related terms among detected terms from previous step ii) slicing the textual patterns between related terms [8]. Once the ontology is populated, it could contribute to calculate semantic similarities between detected terms in phase i).
- *Ontology generation*: Ontology generation with CODA and Pearl, as in the SemaGrow project presented in Sect. 2.
- *Evaluation*: This architecture presents a mutual application-based evaluation design: ideally the learned ontology should improve the information extraction. Besides, Pest observer web portal can be served to validate phytosanitary information extraction from plant health bulletins.

4 Conclusions and Perspectives

New digital technologies allow farmers to predict the yield of their fields, to optimize their resources and to avoid or protect their fields from natural hazards whether they are due to the weather, pests or diseases. This is a recent area where research is constantly evolving. We have introduced in this paper work relevant to our problem, namely: the integration of several data sources to extract information related to the natural hazards in agriculture. We then proposed an architecture based on ontology learning and ontology-based information extraction. We plan in a first phase build an ontology from twitter data that contains vocabulary in the existing thesaurus. To evaluate the constructed ontology, we will extract crops and pests from the learnt ontology, and compare it with tags in pest observer. In the following iterations, we will work on ontology alignment strategies to update the ontology with data from other sources. To go further, multilingual ontology management with keeping tempo-spacial contexts should be investigated.

References

1. Allen, K., Abrams, K., Meyers, C., Shultz, A.: A little birdie told me about agriculture: best practices and future uses of Twitter in agricultural communications. J. Appl. Commun. **94**(3), 6–21 (2010)
2. Batbaatar, E., Ryu, K.H.: Ontology-based healthcare named entity recognition from Twitter messages using a recurrent neural network approach. Int. J. Environ. Res. Public Health **19**, 3628 (2019)
3. Becka, H.W., Kima, S., Haganb, D.: A crop-pest ontology for extension publications. In: EFITA/WCCA Joint Congress on IT in Agriculture (2005)
4. Celli, F., Keizer, J., Jaques, Y., Konstantopoulos, S., Vudragović, D.: Discovering, indexing and interlinking information resources. F1000Research **4**, 432 (2015)
5. Cox, S.: Information technology: the global key to precision agriculture and sustainability. Comput. Electron. Agric. **36**(2–3), 93–111 (2002)
6. Ecophyto: Appel à projets - durabilité des systèmes de productions agricoles alternatifs (2019). Accessed 7 Mar 2020
7. Hammer, G., et al.: Advances in application of climate prediction in agriculture. Agric. Syst. **70**(2–3), 515–553 (2001)
8. Kaushik, N., Chatterjee, N.: Automatic relationship extraction from agricultural text for ontology construction. Inf. Process. Agric. **5**(1), 60–73 (2018)
9. Lokers, R., Konstantopoulos, S., Stellato, A., Knapen, M., Janssen, S.: Designing innovative linked open data and semantic technologies in agro-environmental modelling. In: 7th International Congress on Environmental Modelling and Software, Conference Date: 15–19 June 2014 (2014)
10. Mills, J., Reed, M., Skaalsveen, K., Ingram, J.: The use of Twitter for knowledge exchange on sustainable soil management. Soil Use Manag. **35**(1), 195–203 (2019)
11. Patrício, D.I., Rieder, R.: Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. Comput. Electron. Agric. **153**, 69–81 (2018)
12. Roussey, C., et al.: A methodology for the publication of agricultural alert bulletins as LOD. Comput. Electron. Agric. **142**, 632–650 (2017)
13. Roussey, C., Ghorfi, T.A.: Annotation sémantique pour une interrogation experte des Bulletins de Santé du Végétal. In: Ranwez, S. (ed.) 29es Journées Francophones d'Ingénierie des Connaissances, IC 2018, AFIA, Nancy, France, pp. 37–52, July 2018
14. Tripicchio, P., Satler, M., Dabisias, G., Ruffaldi, E., Avizzano, C.A.: Towards smart farming and sustainable agriculture with drones. In: 2015 International Conference on Intelligent Environments, pp. 140–143. IEEE (2015)
15. Turenne, N., Andro, M., Corbière, R., Phan, T.T.: Open data platform for knowledge access in plant health domain: VESPA mining. CoRR abs/1504.06077 (2015)
16. Patil, V.C., Al-Gaadi, K.A., Biradar, D.P., Rangaswamy, M.: Internet of things (IoT) and cloud computing for agriculture: an overview (2012)
17. Wimalasuriya, D., Dou, D.: Ontology-based information extraction: an introduction and a survey of current approaches. J. Inf. Sci. **36**, 306–323 (2010)
18. Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J.: Big data in smart farming-a review. Agric. Syst. **153**, 69–80 (2017)