View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Repositorio Universidad de Zaragoza

Agricultural recommendation system for crop protection

Javier Lacasta, F. Javier Lopez-Pellicer, Borja Espejo-García, Javier Nogueras-Iso, F. Javier Zarazaga-Soria

Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Spain

Abstract

Pests in crops produce important economic loses all around the world. To deal with them without damaging people or the environment, governments have established strict legislation and norms describing the products and procedures of use. However, since these norms frequently change to reflect scientific and technological advances, it is needed to perform a frequent review of affected norms in order to update pest related information systems. This is not an easy task because they are usually human-oriented, so intensive manual labour is required. To facilitate the use of this information, this work proposes the construction of a recommendation system that facilitates the identification of pests and the selection of suitable treatments. The core of this system is an ontology that models the interactions between crops, pests and treatments.

Keywords: Ontology creation, Ontology population, Data integration, Intelligent systems, Pest control

1. Introduction

Agriculture is a vital sector in the economy of any country, but depending on the crop between 26% and 80% of the agricultural production is lost because of pests (Oerke, 2006). Crop protection is vital but also challenging due to the multiple pests that affect them, such as insects, plant pathogens and weeds, and the toxic effects of most of the existing solutions (Alavanja, 2009). Because of these effects, most countries have established strict regulations for their use and promote non-chemical solutions (European Parliament, 2009).

In general, the norms about pest control are published in heterogeneous and human oriented formats, so intensive manual labour is required to identify the most suitable solution for a given pest. An example of this heterogeneity can be found in the data collections provided by the Spanish Ministry of Agriculture[1] where the description of how to control each type of pest is distributed among multiple heterogeneous textual sources. For example, each document has a layout slightly different from the rest and the names of the pests in the document title are variants of those used in the pest description. This lack of interoperability affects critically tasks requiring some degree of data integration such as identifying the different crops affected by a single organism, finding similitude in the treatment of different species, and comparing the approved pesticides in different countries. Additionally, as new products and techniques are frequently approved, a continuous review is required (Ricci et al., 2010). This happens not only in Spain, but also in many other countries such as United Kingdom[2], United States[3] and Canada[4].

To facilitate the usability of this information, we need systems able to provide it in an integrated and harmonized way. For this task, in this paper, we propose the "Pests in Crops and their Treatments" Ontology (PCT-O). To populate it, we suggest a conversion process for the transformation of non-ontological heterogeneous resources into ontological ones. As use case, this process is applied to transform content from selected Spanish data sources into instances according to PCT-O model. Finally, we describe the structure of the information retrieval (IR) system and the recommendation process that simplifies the identification of a pest and the selection of a suitable treatment.

[1] http://www.mapama.gob.es/

[2] https://secure.pesticides.gov.uk/pestreg/

[3] https://www.epa.gov/pesticide-registration

[4] https://www.canada.ca/en/health-canada/services/consumer-product-safety.html

2. State of the art

The use of ontologies is a classical solution to deal with heterogeneity and interoperability problems. In the biology area, Walls et al. (2012a) remark how semantic models facilitate the creation of intelligent applications that manage living species information. The inference capability of ontologies are especially relevant in the biology area, because it can be used in the taxonomic structures used for classification to simplify conceptual interoperability, data integration and search. However, the creation of ontologies is difficult. The main challenges are the modelling of the information for the desired task, the availability of data for population, and the data transformation complexity. Data modelling is difficult due to different interpretations of the selected knowledge area. With respect to data availability, the availability of data sources conditions the extension and depth of a semantic model. Something similar happens with data transformation. Too complex or too heterogeneous data collections may not be added to the model due to transformation costs.

Several works in the literature categorize living species, the interactions between them or the effects produced by chemical substances. This section describes the main works in these fields, remarks the parts of these models that can be used to describe pest control information, and indicates the shortcomings solved by the proposed PCT-O.

With respect to living being descriptions, the Integrated Information Taxonomic System (ITIS) (Integrated Taxonomic Information System, 2010) contains taxonomic information of aquatic and terrestrial flora and fauna, the Catalogue of Life model (Jones et al., 2000) describes 2 million of species, and the NCBI taxonomy (Gene Ontology Consortium, 2004; Federhen, 2012) stores the organism names and taxonomic lineages in the INSDC database. All these models provide a comprehensive collection of species but they do not provide very detailed information about their features and behaviour. The search capabilities of the portals providing them are limited to the use of names or database codes.

Other works provide extended taxonomies with additional information such as species descriptions, biology, lifecycle, habitat, and interaction with other species. An example of this type of works is Wikispecies (Wikimedia foundation, 2017), which contains near half a million of species, although the information provided for each species is limited. Focusing on plants, the U.S. plants database (Natural resource conservation service, 2016) includes a quite detailed textual description of U.S. plant, their distribution, life cycle, and common pests. Another system is the European Nature Information System (EUNIS) (Davies et al., 2004). It includes a large collection of species obtained from other databases and indicates the geographical distribution and the level of extinction threat of those species. A relevant work is the Encyclopedia of Life (Li et al., 2004), which provides more detailed information about a million of species and even a basic description of the interaction between species. However, it does not detail the kind of interaction they have (predator, prey, symbiosis, and so on). Sini (2009) describes the AGROVOC vocabulary, an agriculture thesaurus. A part of it provides a taxonomy of living beings that includes the main used crops and pests in the form of hierarchically related concepts. DBpedia (Auer et al., 2007) also contains a formal structure for the information about living species in Wikipedia and Wikispecies. However, the number of provided species is more limited. Finally, GeoSpecies (DeVries, 2013) relates each concept to the Encyclopedia of Life, Wikipedia, Wikispecies, NCBI, ITIS, and other similar systems. Instead of providing proper information about the stored species, it focuses on providing equivalences between the aligned models. The search capabilities in these systems are more complete, allowing textual search in the data content. In the semantic models, such as AGROVOC, DBpedia and GeoSpecies, arbitrary searches are also possible.

Some works specifically focus on the interactions between species. Rodríguez-Iglesias et al. (2017) propose an ontology that details the pathogens that affect plants. It integrates data related to both plant physiology and plant pathology with the objective of facilitating the interpretation of phenotypic responses and disease processes. Similar to this, Walls et al. (2012b) analyse the infectious diseases of plants and the pathogens that cause them. They reuse vocabularies from other plant, pathogen and disease ontologies such as the Infectious Disease Ontology (IDO) (Cowell and Smith, 2010). Finally, the Plant Ontology Consortium (2002) defines a set of ontologies to describe plants, their genes, diseases and growing process that include the relation between plants and harmful virus and bacteria. All these models, as in the previous cases, provide semantic searches that make possible detailed queries and precise results.

With respect to crop treatments, PubChem model (Fu et al., 2015) describes chemical structures, biological activities and biomedical annotations. This includes pesticides and the environmental effects they produce. However, this information is text-based and it is not linked to any living species model. ChEBI ontology is another model describing chemical substances (Degtyarenko et al., 2008). It contains natural molecular entities and synthetic products that affect living organisms. However, it also lacks a semantic relation with the species affected by each chemical product. Here, depending on the part of the models, textual or semantic searches are possible.

Other works integrate parts of all these and other agricultural aspects together. Damos (2013) proposes the definition of ontologies that allow describing all the characteristics of cultivations. He also indicates the need to link the created models to other related data collections that complement them. Damos et al. (2017) show an ontology to describe pest and the treatments approved by the Greek Ministry of Rural Development and Food. The core of the ontology contains the pests that are related to the affected crops and existent treatments. On a broader context, Athanasiadis et al. (2009) describe several ontologies for data integration in the agricultural field. Especially relevant is their agricultural activities ontology for crop management. Goumopoulos et al. (2009) describe an ontology for precision agriculture. It focuses on describing plants and all the technological and electronic devices that surround them in precision agriculture. Finally, Rehman and Shaikh (2011) describe another precision agriculture ontology whose core includes concepts for describing crops and their pests.

The objective of the ontology proposed in this paper (PCT-O) is to connect crops, pests and treatments into a unified model. The formal description of living species taxonomies can be managed with the previously described ontologies such as NCBI taxon or GeoSpecies, the description of plant pathologies is covered by Rodríguez-Iglesias et al. (2017) illnesses ontology, and PubChem covers the application of chemical substances. However, they do not model all the crop protection aspects. Specifically, they do not cover the relation between crops, pests that affect them, and the solutions approved by each country to deal with them. Only Damos et al. (2017) make a proposal to relate information about pests and treatments to the affected crops. However, they propose a high-level model that does not provide detailed properties about each of the proposed classes. The proposed PCT-O allows describing the conditions required by a pest to produce outbreaks and the restrictions on the treatments.

3. Structure of the PCT-O

This section describes the ontology created for the description of pests, crops and their treatments. The core of the proposed model can be considered as an extension of the disease triangle described in Rodríguez-Iglesias et al. (2017), which consists of a virulent pathogen, a susceptible host, and a propitious environment. It has been extended to include non-pathogen pests and the definition of treatments for the pests. We have also modelled the provenance of the information to allow updates and correction of errors in the sources and in the generation process.

The ontology has been created with the Methontology methodology (Gómez-Pérez et al., 2004). Specifically, the modelling has been guided to answer the following competence questions: Which is the pest that is affecting a given crop? Which treatment do I have to apply to deal with the pest? When do I have to apply the treatment? What are the sanitary/environmental restrictions of the treatment?

In the construction process of the PCT-O, we have put a special emphasis on reusing existing models to improve the ontology interoperability. Specifically, we have analysed widely used models of living species (which include both crops

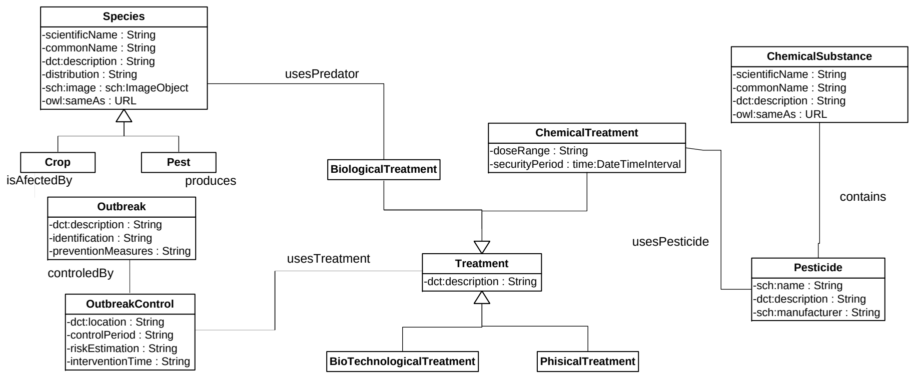

Figure 1: Plant affections and their treatment ontology

and pest) and chemical substances (which include pesticides) described in the state of the art section. The core *Species* and *ChemicalSubstance* classes in the model have DBpedia equivalents, and their instances are linked to NCBI taxon, PubChem, ChEBI ontology instances and the Spanish Wikipedia pages (using *owl:sameAs*). The connection between these elements has been guided according to the information provided in the Spanish guides for pest diagnosis and management.

The Spanish guides that detail the pest characteristics and treatments have provided us the terminology and relations used to construct the proposed ontology. However, their lack of structure has forced us to use a coarse level of granularity for properties, leaving many of them as simple text fields. A finer granularity level is possible, but extracting the concepts and relations from the guides would require the definition of complex natural language processing (NLP) rules specific to each property. This issue is detailed in the discussion section.

Figure 1 shows the conceptual view of PCT-O. The main concept is the *Species* concept, which describes the name and characteristics of the included species. It has been specialized into *Crops* grown by farmers and *Pests* that harm the *Crops*. Crops that act as weeds can be classified as both types. The attributes are the common and scientific name the species, a description, its distribution, images, and equivalency relations with other species models.

The *Outbreak* class models the interaction between crops and pests. It contains a textual description of the produced symptoms, the identification and analysis procedures used to establish that a pest is affecting a crop and the existent prevention measures to reduce the risk of infection. It is based on the IDO ontology, but our ontology also covers insects, plant pathogens and weeds. It has been simplified because of the complexity of filling the description of symptoms from the data sources.

The *OutbreakControl* class models the procedure to control a specific kind of *Outbreak* and its location restrictions. Humidity and temperature are the main triggers of outbreaks. Therefore, control procedures and recommendations may vary depending on the climatology of each region. This class includes the period of time in which the pest is harmful to the crop, the description of a way to estimate the infection risk, the description of the best moment to take action to reduce the damages, and the list of treatments approved in the location for dealing with the pest.

The *Treatment* class describes four kinds of treatments: *Biological*, *Biotechnological*, *Physical* and *Chemical*. Biological treatments make use of predators, physical treatments describe manual measures such as removing infected fruits, bio-technological measures mostly use traps and pheromones, and chemical treatments use pesticides. Each treatment has a description of the treatment itself. The chemical treatments are linked to the pesticides approved by the government (*Pesticide* class), the regulated amount and the legal period between the application and the harvest.

The ontology describes the substances dangerous to the environment contained in pesticides through the *ChemicalSubstance* class. It includes the common and scientific names of the substances and a description of the effects caused and interactions with other species. We link the substances to PubChem, ChEBI ontology and the Spanish Wikipedia through the *owl:sameAs* property. PubChem link is especially relevant as it contains information about the environmental hazards produced by the chemical substances, and the recommended restrictions of use (e.g. many chemical substances must not be used near water sources or some protected/commercial species). We think this information is vital to be able to select appropriately the least aggressive solution among the existent ones for a given place at a given time.

The ontology instances contain information extracted from multiple sources. In this context, knowing the provenance of each piece of information is vital if errors are detected or the sources change. Rodríguez-Iglesias et al. (2016) proposes the use of a named graph structure in which the URI of the named graphs are the base URI of the involved resources. We implement a similar solution by using the PROV ontology (Lebo et al., 2013), which is recommended by W3C for provenance description in the web. From PROV, we have used the *Bundle* class and *hasDerivedFrom* property as our goal is to store the instance sources. A *Bundle* is a named set of provenance descriptions that describe the common provenance properties of a set of elements. *Bundles* contain the *hasDerivedFrom* property that links the *Bundle* to the source file of the controlled elements. The direct implementation of a *Bundle* is using a named graph. Named graphs define collections of resources in a semantic repository under a single name and can be annotated with the necessary properties. The combination of the *Bundles* provides the complete view of the provenance of the crops, pests and treatments. Figure 2 shows an application example where the information extracted from the "Agrotis Ipsillon" diagnosis guide is stored in a named graph and then integrated with the rest of the instances for query. Since the information obtained from each source is stored in different named graphs, it is possible to identify their provenance by querying about the named graph that contains it.

4. Ontology construction and population

The backbone of the ontology instances are the NCBI taxon and the Spanish Wikipedia for living species (crops and pests) and PubChem, ChEBI ontology, and the Spanish Wikipedia for pesticide substances. The NCBI taxon, PubChem and ChEBI ontologies are well-known models in their respective fields

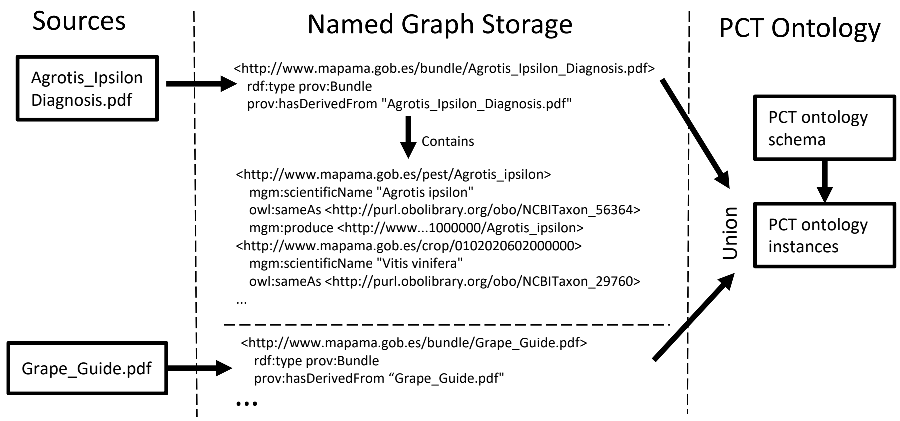

Figure 2: Example of provenance modelling

and provide the scientific names for each element (crop, pest and chemical substances). Specifically, NCBI taxon provides a hierarchy of species useful for identification of families of crops. The Spanish Wikipedia provides alternative scientific and common names that are helpful in the disambiguation process. Each model has additional information about species and chemical substances such as taxonomic relations, definitions, chemical formula and so on. We do not currently use this information, but the linkage makes it accessible for future improvements.

To populate the PCT-O we have focused on the official information about crops and authorised pesticides maintained by the government of Spain. This section describes the data sources, the ontology construction and the process developed to extract the available information and represent it according to the ontology model.

4.1. Tools used for ontology construction

We have selected OWL (McGuinness et al., 2004) as the description model for our ontology and its instances. OWL is the most common RDF-based description model in the semantic field and it enriches the description capabilities of RDF/RDFS (Brickley et al., 2014) by supporting complex relations between classes and detailed characterization of properties. The construction of the ontology has required the use of multiple tools and libraries to define the model and populate it from the selected sources. The ontology has been created using the Protégé editor[5], a tool designed to facilitate the creation of OWL schemas. With respect to the ontology population, it has required the extraction of information from multiple PDF files. This has been done using Apache PDFBox[6], a Java library for PDF processing. For the processing of the extracted content, a workflow that fills an Apache Jena [7] triple-store (a RDF database that support named graphs) has been created using Spring Batch[8]. Finally, the recommendation tool is a very simple text interface that uses SPARQL (Prud et al., 2006) (a language for querying RDF graphs) to extract the desired information from the Jena triple-store.

4.2. Data sources used for population

The description of the effects that each pest has in each crop and the processes established to detect and treat them have been obtained from the following heterogeneous document collections provided by the Spanish Ministry of Agriculture: The laboratory diagnosis sheets of noxious species for crops created by the phytosanitary diagnosis and survey laboratory, which is a collection of 464 scanned PDF documents describing plants, insects, bacteria and virus (scientific and common names of the pests that affect crops, their distribution in Spain, symptoms, detection measures and identification procedures); the guides for the integrated control of pests created by the national plan for sustainable use of pesticides, which is a collection of 21 digital PDF documents that describe the crops affections in Spain and the recommendations for their treatment (common name of the crops, the common and scientific name of the noxious species, control and prevention measures, and available non chemical treatments); and the registry of pesticides approved by the national institute for agrarian research

[5]https://protege.stanford.edu/
[6]https://pdfbox.apache.org/
[7]https://jena.apache.org/
[8]https://projects.spring.io/spring-batch/

and technology, which is a repository containing 2375 PDF records detailing the pesticides allowed in Spain, their composition and use restrictions.

The content of these sources connects the living species information with the chemical substances used on them. The main issue of these collections is their heterogeneity. None of these data sources is completely structured and uniform. Some parts have a tabular structure, but most of them are described as paragraphs of plain text. The text sections are similar between documents but not exactly equivalent. Additionally, the quality of several scanned documents is low, making data extraction difficult.

4.3. Population process

We have followed the population process described in Figure 3. The first step has been to extract the textual content and available images from the source PDF files. Then, each type of source has been parsed to identify the elements required in the ontology. Textual content is used for filling the different properties of the instances, while the images are stored as a graphical representation of each concept. All the extracted images are stored, independently of the relevance of their content. To simplify data integration, each extracted resource is aligned to the previously described ontologies using the common and scientific name of crops, pests and chemical substances as matching text. Having identified the species/chemical substances in the resources, their integration is direct. The first half of the process is dependent of the selected sources, but the second half can be directly used for integrating future additional data collections.

In the data extraction step, if the origin of the PDF file is analogical (scanning of a printed document), the OCR process in the PDFBox library is applied to extract the text. However, scan quality of the source files limits the quality of the extracted content. Most of the extracted text contains minor errors due to bad recognition of some characters, but a few have higher error rates. In addition to this, the non-plain text parts of the documents are not correctly extracted due to PDFBox limitations (e.g., captions of photos or tabular information).

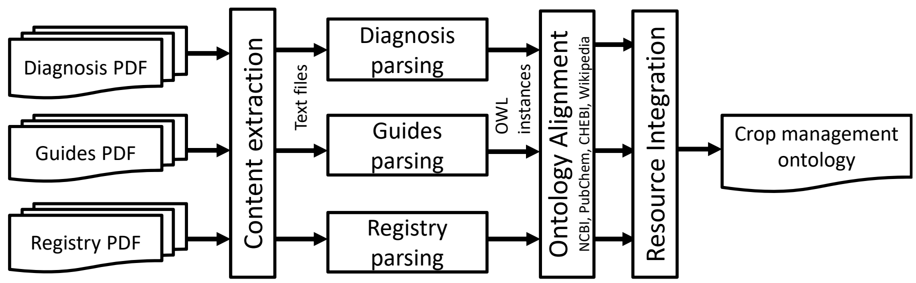

Figure 3: Ontology population process

The parsing step makes use of the fact that all the analysed sources are divided into sections whose content mainly corresponds with properties of the defined model. It identifies these sections according to a list of predefined headers for each type of document that contain all the variant forms found for the sections names and structure of the source documents. Additionally, we have defined specific rules containing syntactic patterns describing textual constructions in the documents when describing the common or scientific name of a species. The extracted information and its provenance information is stored according to the PCT-O model.

The alignment step matches the extracted resources describing species (crops and pests) with the NCBI taxon and the Spanish Wikipedia, and the chemical substances with respect to the Spanish Wikipedia, PubChem, and ChEBI ontologies. The alignment of the species is used to directly merge the information of the involved data collections. The alignment of the chemical substances is used to facilitate the identification of equivalences between the different products used to deal with the pests.

The alignment has been performed looking for equivalences in the scientific names of species and chemical substances contained in the documents. The complexity of this alignment process has come from the need of identifying and correcting the errors in the sources, and because of the existence of synonyms and variants of names of the living beings and chemical substances. To deal with these problems, we have performed the following alignment sub-steps. First, we

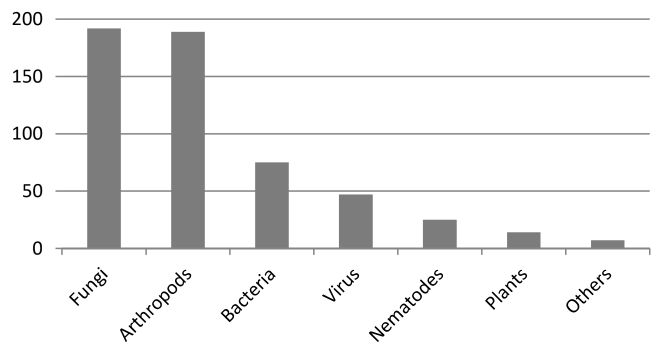

Figure 4: Classification of pests

have extended the available synonyms and variant names for each extracted crop/pest with additional names obtained from the Spanish Wikipedia. This has been done looking for the common names in the Spanish Wikipedia and extracting the scientific ones contained in the corresponding info-boxes. Then, all the scientific names are matched (exact match) with the corresponding ontology/model (NCBI, PubChem, ChEBI). If a match is found, the alignment is established. If there is no correspondence, we have used the Levenshtein distance (Levenshtein, 1966) to identify matches with minor errors and variants of the scientific names. For this comparison, the scientific names are normalized removing abbreviations, numbers, and texts in brackets. Name heterogeneity has led us to use a threshold of 20% of the name size to decide if the most similar name can be aligned or not. Therefore, shorter names allow smaller differences than longer ones. This threshold has been selected experimentally to reduce the number of incorrectly aligned concepts (we prefer to leave them unaligned).

The resulting ontology consists of 549 pests that affect 462 crops through 3471 outbreaks. Figure 4 shows the pests in the model aggregated by family. It can be observed that most of them are fungi and arthropods. In addition to those, there are virus, bacteria, nematodes and other plants. A few pests are from species that do not fit in the previous categories. To deal with these pests, there are 42397 different chemical treatments involving 2109 pesticides with 566 different chemical substances, and 219 alternative treatments.

A manual review of the ontology has shown that 96.12% of the species (pests and crops) have been correctly aligned to their scientific name in NCBI Ontology. The main source of errors are problems in the description of the names of the sources (e.g., "summer cereals"), the use in the sources of the fruit name instead of the plant name or the lack of equivalences for some of the used common names. We have also reviewed the quality of the extracted description of the species, the symptoms and the information related to prevention and intervention time. Here the quality is worse due to the difficulty of extracting the content. There are almost no records without syntactic errors. Most of them are small, but to be usable, it is required to correct them through a manual proofreading. Something similar happens with treatments: the extracted information has been correctly assigned to the corresponding concepts in the ontology, but there are many syntactic errors caused by the extraction. Finally, we have also reviewed the alignment of the chemical substances with the ChEBI database (PubChem is linked to it). The result shows that just 59.9% of the chemical substances have been correctly aligned, 27.7% of them are left unaligned and the rest (12.4%) are incorrectly aligned. This alignment problem is caused by the lack of correspondence between the Spanish common/scientific names for the chemical substances in the sources and the Spanish Wikipedia. The Spanish Wikipedia has proven to be a good source to align common and scientific names of living species but its coverage for chemical substances is much worse. It does not describe many specific substances, thus the Spanish names cannot be aligned to the English ones in the selected ontologies.

From these data, it can be observed that current crop protection is completely focused around the use of chemical products. There are many more chemical solutions than alternative ones, and their amplitude of action is also broader because they affect several pests. With respect to alternative approaches, they are only able to deal with a small set of the pests (mainly insects) but they do not have secondary effects for humans or nature.

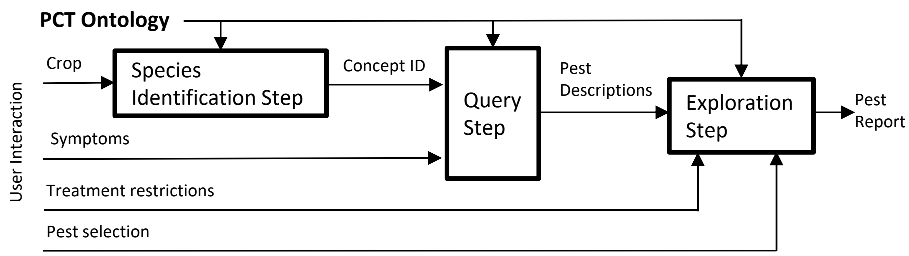

Figure 5: Query process

4.4. Recommendation system scenario

This section describes the developed IR-based recommendation system, constructed on top of PCT-O to obtain directly complex information useful for crop protection, and describes its potential and limitations. Figure 5 shows the different components of this process. These components use SPARQL queries to process the ontology and provide the results. The species identification step finds the crop concepts that correspond with the ones used in the query. Here, all the registered variants of common and scientific names are matched with the query term and the concept that matches it is returned. The query step identifies the pests that affect a crop with the symptoms indicated by the user. Since the species are defined in a taxonomical way and several of the relations are at category level (e.g., citric or fungus), any search by a member of these categories can be expanded to obtain all the pests affecting to its category. Finally, the exploration step starts when the user selects a pest from the set obtained through the query step. Then, the local pest information and treatments are selected based on the user restrictions. If information from additional countries were added, it would be also possible to restrict solutions for products cultivated for exportation or even to identify better solutions than the one currently approved in the residing country.

Because of the coarse granularity level of the ontology, the query and exploration restrictions have to be done on text fields. This is a system limitation as text match solutions have problems related to synonymy, polysemy and multiple variant forms that reduce match quality. In this system, we have not used provenance information, because their main purpose is for tasks related to model updates, and versioning.

This recommendation system shows how PCT-O facilitates identification tasks, but PCT-O also allows direct queries to list all the available treatments for a pest in a crop. In this case, there is no ambiguity problems because it is a direct query about specific elements that are perfectly identified.

```
1. Query = Crop:"Lemon tree", Symptoms:"Brown leaves", Treatments:"Biological"
2. Species identification step:
Select ?crop where {{?crop mgm:scientificName ?name. FILTER regex(?name, "Lemon tree", "i" )}
    union {?crop mgm:commonName ?name. FILTER regex(?name, "Lemon tree", "i" )}
    Result: http://www.mapama.gob.es/crop/0102020104000000 <- Citrus Limon URI
3. Query step:
Select ?outb where {{<http://www.mapama.gob.es/crop/0102020104000000> mgm:isAfectedBy ?outb}
    union {<http://www.../0102020104000000> skos:broader+ ?crop. ?crop mgm:isAfectedBy ?outb}.
    ?outb dc:description ?descr. FILTER regex(?descr, "Brown leaves", "i" )}
    Result: http://www.mapama.gob.es/ourbreak/0102020100000000/Tetranychus_urticae
            http://www.mapama.gob.es/ourbreak/0102020100000000/Citrus_exocortis_viroid_(CEVd)
4. Exploration step:
Select ?treatment where {<http://www.../Tetranychus_urticae> mgm:isControledBy ?control.
    ?control mgm:usesTrearment ?treatment.
    ?treatment rdf:type <http://www.mapama.gob.es/vocabulary#BiologicalTreatment>}
```

Figure 6: Example of query specification and SPARQL queries performed

As a summarised example of this IR flow, we describe how the query depicted in Figure 6 is executed (it is simplified and just the concept identifier is returned). The current query interface allows introducing the query terms to search in the crop name, symptoms produced by the pest, and restrictions in the treatment. The selected query (1) searches for a pest affecting the “Lemon tree” that produces “Brown leaves” and how to treat it with a biological treatment. The species identification step (2) directly matches the “Lemon tree” species name with the “Citrus limon” concept in the ontology. “Citrus limon” has no direct specification of pests as they are common to all ”Citrus” family. Thus, the query step (3) expands the query to the ”Citrus” species and finds two different pests, “Citrus exocortis viroid” and “Tetranychus urticae”, that produce “Brown leaves”. For this expansion, we use a crop taxonomy extracted

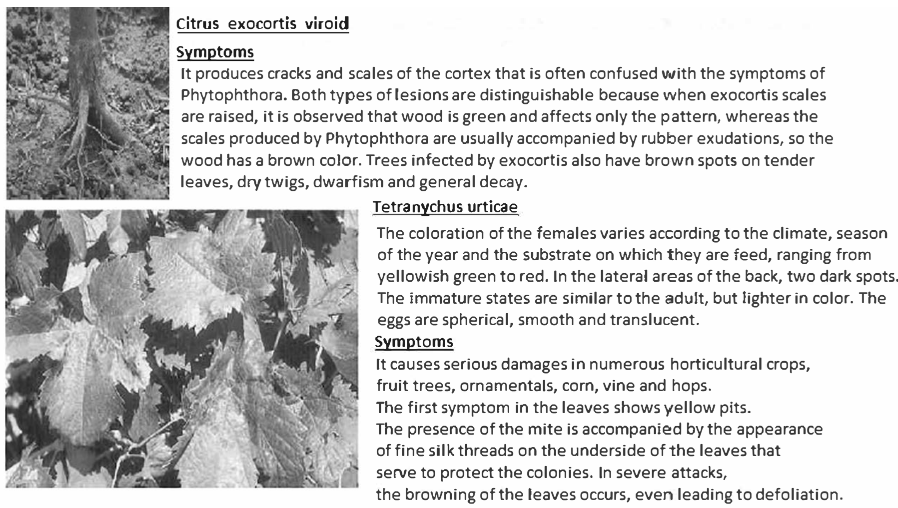

Figure 7: Example of information returned by the Query Step

from the sources, but since NCBI is liked to the concepts, it also could be used for this task. Figure 7 shows a composition of the information that can be returned in the Query Step (the original Spanish text has been translated to English to facilitate its understanding). Finally, given the "Tetranychus urticae", the exploration step (4) returns the available biological treatments for it, which consists in releasing predators such as Amblyseius (Neoseiulus) californicus, Phytoseiulus persimilis and Diptera Feltiella acarisuga.

Two problems have been found in this query system. First, source information is sometimes imprecise or incomplete. This is the case of the "Citrus exocortis viroid" that has no description. This lack of information can limit the ontology usability. The second issue is related to the generality of the information. For species that attack multiple crops, sources only provide the most general and representative examples. In this case, the "Citrus exocortis viroid" image is focused on roots, because the main symptom focuses there (leaves coloration is secondary). In the "Tetranychus urticae" case, the image shows a leaf affected by the pest, but from a plant different from the "Citrus limon". Correcting both issues would require to increase the amount and precision of the data sources available.

5. Discussion

As indicated in the state of the art section, there are several models for the description of species and chemical substances, but only Damos (2013) and Damos et al. (2017) provide some relation between crops, pests, and treatments. PCT-O goes a step further by including the description of the conditions of these relations. Therefore, in PCT-O, it is possible to specify the period of time when a pest is harmful, when it is needed to react, and the nature of the treatments. PCT-O also includes provenance information to keep track of the data sources. The next closest solution is the PubChem database (and ontology) that describes thousands of chemical substances and their application in the industry. For the appropriate substances, it indicates the common name of the crops to which the substance can be applied according to USA legislation. However, it is not linked to any species ontology and may be ambiguous. Additionally, it indicates neither a detailed list of the noxious species the chemical substance can deal with, nor the symptoms, periods of control or chemical alternatives.

In the analysed scenario, we have shown how PCT-O helps in terms of interoperability and data integration between crops, pests and treatments information. Thanks to it, it is possible to construct a semantic recommendation system that helps to determine the pests that affect each crop and how to treat them. The crops, pests, and pesticides are linked to commonly used ontologies and taxonomies. This removes name ambiguity and allows comparing solutions adopted in different regions or countries.

The population of the ontology with Spanish official data has illustrated the complexity of obtaining a complete model from the available official sources. Data quality has been an issue that has complicated the data transformation and it has added errors. We have found several cases where a correct equivalence has not been found and chemical substances have been incorrectly aligned. The cause of this is mainly due to the incompleteness of Spanish Wikipedia in biology/chemistry area and the similarity between some scientific names of species/chemical substances. Another identified issue is related to the completeness and overlap of the data sources. Each data source was created by its producer with a different purpose and they do not completely overlap. For instance, the guides only cover a subset of species described in the diagnosis files. As a result, the populated ontology does not have a uniform coverage: some species are very detailed, other ones contain very limited information. These restrictions reduce the usability of the extracted information, but it is a good starting point for future improvement.

Because of the automatic nature of the population process and the heterogeneity of the sources, the resulting collection requires manual validation. For this task, the stored provenance information becomes vital as incorrect or poorly described instances can be traced to the original sources, allowing the detection of the source documents with errors, so they can be fixed.

Although we have focused on Spain data for the population step, information from other countries could be added. Countries such as U.S., United Kingdom or Canada also provide the information required to populate this ontology in heterogeneous formats, but specific extraction and transformation steps for each new source format would be required. The step that align each species/chemical with the selected ontologies and the final integration phase could be reused.

A limitation of PCT-O is the selected semantic granularity of the model. The information contained in fields such as pest description, control period, identification procedures, or intervention time is described as plain text, so queries on these fields are imprecise. For example, when querying for"Brown leaves" as pest symptom, pests that only produce brown leaves in some specific situations will be returned with the same importance than pests with brown leaves as representative symptom. Solving this problem would require to extend the ontology to allow a precise description of such content. However, available information is so heterogeneous that cannot be automatically interpreted only with the information contained in the source files. For example, in the period of control of a crop, it is important to consider the growth stage, temperature and humidity. The growth state is sometimes properly described (e.g., flowering), but other times it is referenced using periods of months or seasons (e.g., May). This must be interpreted depending on the place and the climate conditions of a given year. The same happens with the humidity or temperature. Some descriptions are quite clear (e.g., temperature under 25 degrees), but others need human interpretation (e.g., high temperature). In this context, a semantic baseline for each crop must be defined to allow the mapping of all the imprecise descriptions to measurable values. We have done a preliminary processing to identify the common temperature and humidity patterns in the source documents and more than 80 different rules have been needed. Additionally, we had to perform approximations that are crop and pest dependent. For instance, many documents say that a crop is vulnerable to a pest with high temperature, but how much temperature is "high"? To model it semantically, this must be translated to a numerical range (as it is in many other descriptions). However, with the source information alone it is not possible to determine a precise value, and an approximation must be given. Due to these approximations, we think that the fine grain semantic extraction can only be useful as an initial step in IR process. The final decision must be taken by the user who has interpret the original description.

6. Conclusions

This work proposes the PCT-O ontology, a model to describe the outbreaks that pests produce to crops and the approved ways to treat them. Currently, there are several ontologies to describe taxonomies of living beings but none allows describing their inter-relations as the PCT-O ontology. As use case for this ontology, we propose a recommendation system that helps to identify the pests affecting a crop and their treatments.

The ontology has been populated with official information in Spain about crops, pests and approved treatments. This process has been complex due to the heterogeneity, format and quality of the data sources. The extraction and source errors, complemented with synonymy and name variants, have forced us to use a disambiguation process of scientific names based on the alignment of species and chemical substance records with ontologies such as NCBI, PubChem, ChEBI and Wikipedia. The resulting model has been tested in a suggestion use case to determine how to identify a pest and select a treatment. Additionally, it can be used for tasks such as the identification of outbreaks, identification of location-based related conflicts with the treatments, and comparison of solutions between country legislations.

A first area of future work is to integrate treatments adopted by other countries for the same illnesses/pests in the population of the ontology. This will require extending the extraction and parsing step to deal with the additional data sources, but it will allow complementing the pest descriptions and comparing the approved treatments to detect differences between regions. These differences may show gaps in country legislations, and allow identifying better solutions for a region than the currently approved ones.

Another interesting extension would be to include other aspects of the use of chemical substances in the land. For example, PubChem repository contains information about the hazards of the use of the chemical substances, such as "Very toxic to aquatic life with long lasting effects"). This information merged with water flow, crops or protected species distribution maps can be useful to determine the areas where a product can be used, or suitable alternatives for areas that forbid it. A complementary source of this information is the EU - Pesticide Database (European Commission, 2005) that stores the list of substances approved in each European member state for their use as pesticides. Finally, the ontology could be extended to integrate more detailed information about crops and their varieties. For example, the Spanish Ministry of Agriculture provides a collection of descriptive sheets containing information about the different crop varieties used in Spain. This collection provides information about the growth conditions, performance and resistance of the different varieties of species. This could be used to recommend the best variety for a field given its climate and the distribution of the registered pests.

Acknowledgments

This work has been partially supported by the Spanish Government through the project TIN2017-88002-R. The work of Borja Espejo-Garcia has been partially supported by Aragón Government through the grant number C38/2015.

References

M. C. R. Alavanja. Pesticides use and exposure extensive worldwide. *Reviews on Environmental Health*, 24(4):303–309, 2009.

I. N. Athanasiadis, A. E. Rizzoli, S. Janssen, E. Andersen, and F. Villa. Ontology for seamless integration of agricultural data and models. In *Conf. on Metadata and Semantic Research*, pages 282–293, 2009.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735, 2007.

D. Brickley, R. V. Guha, and B. McBride. RDF Schema 1.1. *W3C recommendation*, 2014.

L. G. Cowell and B. Smith. Infectious disease ontology. In *Infectious disease informatics*, pages 373–395. Springer, 2010.

P. Damos. Semantics and emergent web-3 technologies: Modern challenges for integrated fruit production systems towards internationalization. *IOBC-WPRS Bulletin*, 91:133–142, 2013.

P. Damos, S. Karampatakis, and C. Bratsas. Representing and integrating agro plant-protection data into semantic web trhough a crop-pest ontology: The case of the greek ministry of rural development and food (GMRDF) ontology. *IOBC-WPRS Bulletin*, 123:122–127, 2017.

C.E. Davies, D. Moss, and M. O. Hill. EUNIS habitat classification. Technical report, European Environment Agency-European Topic Centre on Nature Protection and Biodiversity, 2004.

K. Degtyarenko, P. de Matos, M. Ennis, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(1):344–350, 2008.

P. J. DeVries. GeoSpecies knowledge base, 2013.

European Commission. EU pesticides database. Online database, 2005.

European Parliament. Regulation (EC) 1107/2009 of the European Parliament and of the Council. Technical report, EU, 2009.

S. Federhen. The NCBI taxonomy database. *Nucleic acids research*, 40(1): 136–143, 2012.

G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton. PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *Journal of Cheminformatics*, 7(34), 2015.

Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(1):258–261, 2004.

A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering*, chapter Methodologies and methods for building ontologies. Methontology, pages 125–142. Advanced Information and Knowledge Processing. 2004.

C. Goumopoulos, A. D. Kameas, and A. Cassells. An ontology-driven system architecture for precision agriculture applications. *International Journal of Metadata, Semantics and Ontologies*, 4(1-2):72–84, 2009.

Integrated Taxonomic Information System. Integrated taxonomic information system on-line database, 2010.

A. Jones, X. Xu, N. Pittas, et al. Spice: A flexible architecture for integrating autonomous databases to comprise a distributed catalogue of life. In *Int. Conf. on Database and Expert Systems Applications*, pages 981–992, 2000.

T. Lebo, S. Saho, and D. McGuinness. PROV-O: The PROV ontology. Recommendation, W3C, April 2013.

V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals*, volume 10, pages 707–710. 1966.

W. Li, R. W. Byrnes, J. Hayesa, et al. The encyclopedia of life project: grid software and deployment. *New Generation Computing*, 22(2):127–136, 2004.

D. L. McGuinness, F. Van Harmelen, et al. OWL web ontology language overview. *W3C recommendation*, 2004.

Natural resource conservation service. The plants database, 2016.

E. C. Oerke. Crop losses to pests. *Journal of Agricultural Science*, 144(31-43), 2006.

Plant Ontology Consortium. The plant ontology consortium and plant ontologies. 3(2):137–142, 2002.

E. Prud, A. Seaborne, et al. SPARQL query language for RDF. 2006.

A. Rehman and Z. Shaikh. Ontagri: scalable service oriented agriculture ontology for precision farming. In *Int. Conf. on agricultural and biosystems engineering vols*, pages 1–2, 2011.

P. Ricci, M. Barzman, F. Bigler, et al. Integrated pest management in europe. Technical report, ENDURE Network, 2010.

A. Rodríguez-Iglesias, A. Rodríguez-González, A. Irvine, et al. Publishing fair data: an exemplar methodology utilizing phi-base. *Frontiers in plant science*, 7:641, 2016.

A. Rodríguez-Iglesias, M. Egana Aranguren, A. Rodríguez-González, and M. D. Wilkinson. Plant-pathogen interactions ontology (PPIO). In *Int. Conf. on Bioinformatics and Biomedical Engineering*, 2017.

M. Sini. Semantic technologies at FAO. *Agricultural information management standards, International Society for Knowledge Organization (ISKO)*, 3, 2009.

R. L. Walls, B. Athreya, L. Cooper, et al. Ontologies as integrative tools for plant science. *American journal of botany*, 99(8):1263–1275, 2012a.

R. L. Walls, B. Smith, J. Elser, et al. A plant disease extension of the infectious disease ontology. In *ICBO*, pages 1–5, 2012b.

Wikimedia foundation. Wikispecies: free species dictionary, 2017.