

Knowledge transfer from agri-food scientific papers to a knowledge base

Rafał Trójczak, Robert Trypuz, Anna Mazurek, and Piotr Kulicki
The John Paul II Catholic University of Lublin
Department of the Foundations of Computer Science
Faculty of Philosophy
Al. Raławickie 14, Lublin, Poland
Email: trypuz@kul.pl

Abstract—We address the problem of the access to the results of scientific publications in the agri-food domain. We focus on the description of main contributions of the papers treating them as accepted or rejected beliefs of their authors expressed in the form of scientific laws. We define the structure of different kinds of scientific laws present in the domain in the form of an ontology. The main concern of the paper is a process in which we proceed from the abstracts of papers to the ontological representation of laws. Moreover, we present examples of SPARQL queries which show how the resulting knowledge base can be used. Among the uses we point out discovering new scientific hypotheses and incoherencies among scientific laws.

Index Terms—ontology, scientific law, knowledge extraction, scientific hypotheses, incoherency

I. INTRODUCTION

IN RECENT years the number of scientific publications has increased significantly. It is common knowledge in scientific community that there are so many works published every year that it is not possible for one person to read all of them, even if that person limits themselves to a narrowed specialization. The situation of professionals who are working at the crossroads of science and practice is even worse because they must not only struggle through countless scientific papers but they must also cope with the difficult scientific language in which these papers are written. This problem gets even more serious when it comes to scientific projects where research is being conducted in many loosely connected disciplines.

Some of the journals are already providing structured abstracts within the papers published there. One of the possible structures of such abstracts is to divide the abstract into three sections: background, results, and conclusion. And, additionally, there are keywords at the end¹ [1]. Another way to facilitate access to the content of a paper is using so called *highlights* which are “a short collection of bullet points that convey the core findings and provide readers with a quick textual overview of the article”.² There are also semantic solutions addressing this problem. For example, the authors of the Nanopublication and the Micropublications projects have created models to represent statements and argumentation from

scientific papers. We shall discuss the projects in the section dedicated to related works.

The research presented in the paper was conducted within ProOptiBeef³ project, oriented towards increasing the level of innovation in Polish beef sector. One of the tasks of the project is creating a knowledge base (henceforward KB) of recent results published in the top journals relevant to the domain of beef production and consumption. These results were selected by the community of the experts involved in the project. One of the main goals of constructing the KB is to help scientists decide which experiments should be conducted in order for the research to be innovative.

In order to transfer knowledge from the papers to our KB, we have created an ontology called Science. It is designed to represent proved and disproved statements extracted from scientific papers stored in the database of the ProOptiBeef project. We shall informally describe the Science ontology in section II of this paper. In section III we shall present the way in which we fill the KB with the information from scientific papers. The final element of our KB is a reasoning part. This element enables us to perform some reasoning algorithms on our KB in order to discover new scientific hypotheses and to detect incoherency among scientific laws which have already been represented. We shall present an implementation of three of these algorithms in the SPARQL query language in section IV. In section V we shall shortly describe related works. The last section, section VI, contains summary and describes our planned future work related to the project described in this paper.

II. ONTOLOGY OF SCIENTIFIC LAWS

As mentioned above, the project’s KB is structured according to Science ontology (see [2], [3] for the formal structure of the ontology and [4] for the philosophical sources of the ontology). The ontology is expressed in OWL and is composed of two parts: TBox and ABox. TBox (terminological component) consists of a taxonomy of scientific laws (see figure 1) and provides a formal characterization of each class of the

¹See also http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html. Journal of Biomedical Semantics can serve as an example.

²See: <http://www.elsevier.com/journal-authors/highlights>.

³Full name of the project: *ProOptiBeef – Optimizing beef production in Poland according to strategy “from fork to farm”*. Web page: <http://prooptibeef.pl>.

taxonomy⁴. Laws from different classes differ in their structure and therefore require different representation specified by the ontology. The top distinction in the taxonomy is between *quantitative and qualitative laws*⁵. Quantitative Law concerns the dependency between *qualities* of an entity that can be measured (e.g., time of (beef) aging influences (beef) pH value) while Qualitative Law concerns *perdurants* or *endurants* describing directly their properties or comparing them with respect to their qualities (e.g., meat of bulls housed in groups before slaughter was less tender than meat of bulls individually housed).

The meaning of terms “endurant”, “perdurant” and “quality” comes from the DOLCE ontology [5] which is used as a foundational ontology for Science. The terms are to be understood as follows:

- endurants are wholes that endure in time, e.g., beef, pasture, food;
- perdurants are entities that “happen in time”; they can have temporal parts or spatial parts, e.g., feeding, slaughtering, cooking;
- qualities are entities that can be perceived or measured, e.g., height, weight, age, protein content.

Quantitative Law is further divided into Correlation Law and Functional Law. Correlation laws state that there is a correlation between the values of the qualities, but the details of the correlation are unknown, i.e. it is not possible to determine the function (direction) of it (e.g., there is a significant correlation between sarcoplasmic protein solubility and both expressible moisture and color parameters). Functional laws say that there is a dependency between the values of two qualities. One of the qualities is called an independent parameter and the other one is called a dependent parameter. A functional law states that the value of a dependent parameter is a function of the value of an independent parameter (e.g., aging influenced instrumental hardness). If the function describing dependency in a law is monotonic (i.e., it is either increasing or decreasing), this law is an instance of Monotonic Law (e.g., fat reduction results in higher cooking losses).

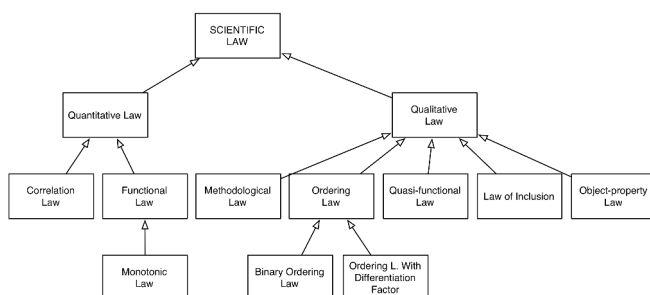


Fig. 1. The backbone of Science ontology.

⁴The TBox of Science ontology can be browsed here: <http://onto.kul.pl/webprotege/>.

⁵The way the laws are represented in Science has been widely described in [2], [3].

There are five subcategories of Qualitative Law: Methodological Law, Quasi-functional Law, Law of Inclusion, Object-property Law and Ordering Law. A methodological law states that a method is used to measure a certain property (e.g. Multiple Linear Regression is used to measure beef tenderness). A quasi-functional law—similarly to a functional law—expresses the dependency between two qualities but one of the qualities is qualitative (e.g., the sex of an animal has influence on the tenderness of its meat – “sex” here is a qualitative quality). A law of inclusion states that one class of objects is a subclass of the other (e.g. meat color is a meat quality). An object-property law states that an object possesses a quality (e.g., red meat is an important source of vitamin D). Ordering Law is divided into two subcategories: Binary Ordering Law and Ordering Law with Differentiation Factor. Each instance of Ordering Law has some ordering basis according to which classes of objects are compared. Binary ordering laws compare two classes of objects (e.g., dairy cow lean has a longer display color life than beef cow lean) with respect to some quality playing the role of order basis (e.g., display color life). Each instance of Ordering Law with Differentiation Factor describes a change of some quality of an object after the object has undergone some modification (e.g., magnesium decreased in beef during cooking).

It is also worth noting that the term “scientific law” which is used by us in the paper covers proved and disproved statements described in scientific papers. Proved and disproved statements are instances of classes: Accepted Scientific Law and Rejected Scientific Law, respectively (see figure 2). It suggests that we

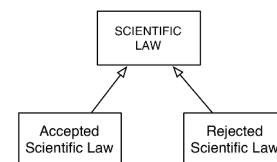


Fig. 2. Each scientific law is either accepted or rejected in a scientific paper.

treat scientific laws similarly to beliefs. Referring to [6] we assume that each scientific law has its “owners”, who are the authors of the paper which the scientific law comes from. The paper is also a source in which the law is exemplified. So modelling author role and the paper as the context in which a scientific law appears is very important for our representation ([7], [8]).

ABox of our KB is simply a set of individuals (i.e., concrete scientific laws extracted from concrete scientific papers) which are represented in accordance to the TBox just described.

Below we shall describe the process of extraction of scientific laws from the scientific papers and their translation into the language of Science ontology.

III. KNOWLEDGE ACQUISITION

A. Extraction of knowledge from scientific papers by domain experts

ABox contains a significant number of 8k scientific laws adequately representing the content of selected scientific articles. The selection of papers from the top journals—in the opinion of researchers involved in the ProOptiBeef—was prepared in the earlier phase of the project. Thus, the remaining task was to present the information from the selected papers in the form designed to represent scientific laws.

For an efficient realization of the task we decided to use abstracts of the papers as the main source of information. The UNESCO guide for the preparation of scientific papers for publication [9] states that “[t]he abstract should contain the results and conclusions of the paper in brief detail adjusted to the size allowed to the abstract and should, within these limits, refer to any new information which it contains. The abstract should not contain information or claims not contained in the body of the paper, nor should it include inessential details. [...] “*New information* should include observed facts, conclusions of an experiment or argument, the essentials of new methods or apparatus, etc.” Moreover, the guide postulates that “[t]he abstract should be *self-contained*”. In the light of these principles an abstract should contain enough information about the main results of the papers to reconstruct scientific laws.

Our experience shows that the practice in the journals from the agri-food domain in principle agrees with these postulates. However, authors use there technical vocabulary and discipline’s folklore. That makes them difficult to be accessed by non-specialists. Thus, the participation of agri-food science specialists in the process of extracting knowledge from abstracts is inevitable.

On the other hand most specialists are not familiar with the formal and computational tools of knowledge representation. After a short introduction to our schema of the representation of scientific laws the specialists invited to the project were aware of the key conceptual distinctions (object, qualities) and the key elements of the representation of theses – the influences between factors and correlations. However, to encourage them to co-operate we had to allow them to present the results of their analyses in a form quite close to natural language with only some structural restrictions.

As a result we introduced a multi step process whose main steps are:

- agri-food specialists extract information from abstracts and write them down in a semi-structured way,
- information is automatically transferred to a database,
- the database is further elaborated by ontologists and (this is itself a multi step process described below)
- finally the information is automatically converted into OWL ontology.

Let us now trace the process in detail.

a) *STEP 1*: An abstract (earlier divided into sentences that are numbered for easier reference) of a scientific paper

goes to the so called *Raiders of The Theses* (most of them are PhD students of broadly understood agri-food science). Their work is controlled and approved by senior members of the project team. The work with the abstract is based on a text form covering the following elements:

- a list of objects
- a list of properties (qualities)
- a thesis formulation from the abstract
- a new elaborated formulation of the thesis

The first two lists constitute a conceptual structure in which the laws presented in the article can be expressed. The description of the list of objects is usually rather complex and is a subject of further ontological elaboration. An example is presented later in the paper when we discuss the work of ontologists (see section III-B). The list of properties is rather simple and is directly transferred into a database where each quality is connected to the article and the thesis it comes from.

Example 3.1: Now let us concentrate on the formulation of theses and analyze an example. The original formulation from the abstract is as follows⁶: “W normalnej atmosferze, fotooksydacja była powierzchowna, ponieważ zaobserwowano odwrotną korelację między wagą plastra mięsa i stężenie COP na podłożu lipidowym, w przeciwieństwie do atmosfery bogatej w tlen (32%).” (“In normal atmosphere, photo[o]xidation was a superficial process, since an inverse correlation between meat slice weight and COPs content on a lipid basis was observed, unlike in a highoxygen (32%) atmosphere.”)

The specialist disambiguated the formulation and divided it into elementary facts forming laws of one of the type from our ontology. The statement from our example is presented as the two following facts:

- “Wraz ze spadkiem masy plastrów surowej wołowiny, poddanych pakowaniu w atmosferze powietrza, zwiększa się stężenie produktów oksydacji cholesterolu.” (“In normal atmosphere, with the decrease of meat slice weight the COPs content increases.”)
- “Wraz ze spadkiem masy plastrów surowej wołowiny, poddanych pakowaniu w atmosferze bogatej w tlen (32%), zmniejsza się stężenie produktów oksydacji cholesterolu.” (“In a highoxygen (32%) atmosphere, with the decrease of meat slice weight the COPs content decreases.”)

B. Pre-ontological elaboration of statements, qualities and objects

a) *STEP 2*: Statements formulated in *STEP 1* were automatically transferred to a database for further elaboration by the ontological staff of the project. Within this process all statements are annotated with unique identifiers. The two above statements from our example received numbers: t_652_3 and t_652_4 respectively. We shall further refer to them using these identifiers.

⁶Specialists worked on Polish translation – we present both Polish and English versions.

b) *STEP 3*: Each statement, whether proved or disproved, has an attributed type of thesis from Science.

For the selected types of scientific laws it is also necessary to determine the monotonic type. There are two main monotonic types: positive and negative. They express the direct or inverse proportion between qualities for the monotonic laws, increase or decrease of quality values for ordering laws. In special cases of ordering laws, additional monotonic types were introduced which are “equal”, “different” or “no monotonic type”. Exceptions, related to additional monotonic types, refer to these in which values of qualities are equal/stay unchanged or values are said to differ but the exact value differences are unknown. The latter exceptions refer to the ordering laws with differentiation factor, where the value of qualities is unknown.

At this point it is also determined for a particular thesis whether it is proved or disproved, by marking it as “accepted” or “rejected”.

Example 3.2: Let us describe this on the example of thesis *t_652_3* (it is a continuation of example 3.1). The statement is proved, so we mark it as “accepted”. It describes dependency between weight and concentration of cholesterol oxidation products. Both qualities are quantitative qualities and quality dependence is known; therefore the thesis should be assigned as an instance of Monotonic Law. Monotonic laws require determining monotonic type. At this case decreasing the weight causes increasing the concentration of cholesterol oxidation products, that indicates an inverse proportion between qualities: “negative” monotonic type.

c) *STEP 4*: The next step of the development refers to relating the qualities to particular scientific laws. The qualities were listed by the domain specialists in the earlier phase (see *STEP 1*), but the list was linked to a particular article not to a thesis (see *STEP 2*). So if there were a few theses extracted from one article (which was almost always the case) there was a need of selecting the qualities for each thesis. A particular thesis needs to have not only proper qualities attributed but also (if possible) their role determined. The latter depends on the type of thesis. To attribute qualities to a thesis, the proper qualities are selected from the list. The proposed names of qualities must precisely describe the qualities mentioned in the thesis. In other case they should be corrected together with their English translation.

For all scientific laws qualities can be attributed; the only exception might be the laws of inclusion, which may refer to two or any quality. If a statement expresses any kind of influence between two or more qualities, then we should characterize the roles of these qualities in the statement. It is so in the case of all functional, monotonic and quasi-functional laws. In such statements dependent and independent qualities are distinguished. For methodological, ordering and object-property laws single qualities are attached to the statement. The special cases are laws of inclusion and correlation laws. As it was said, the laws of inclusion might refer to two or any quality. In the first case two qualities are

assigned to a thesis in the following way: the first quality is a specialization of the second one. For correlation laws the direction of influence is unknown, for this reason each quality is assigned a number that corresponds to the order of its occurrence in the thesis.

Example 3.3: In our example (it is a continuation of example 3.2) two qualities: stężenie produktów oksydacji cholesterolu (concentration of cholesterol oxidation products) and masa (weight) were extracted from the abstract (to which refers thesis *t_652_3*). Both of them refer to the statement *t_652_3*. Since the statement belongs to Monotonic Law, qualities dependency must be characterized. Hence, “weight” should be attributed to the statement as independent quality and “concentration of cholesterol oxidation products” as a dependent quality.

d) *STEP 5*: In the next step the elaborated data are automatically translated into an OWL ontology. The result is the Science ontology with classified instances of scientific laws. Each concrete scientific law was assigned qualities in appropriate roles, monotonicity (if applicable), the identifier of the article it comes from, the original sentence from the abstract of the article which expresses the law, its Polish translation and expert’s elaboration.

e) *STEP 6*: The last part of the work is related to the description of objects (the experimental materials) the theses are about. The objects for the theses were described by the domain experts. Example 3.4 shows a description of the object for the aforementioned theses *t_652_3* and *t_652_4*.

Example 3.4 (Objects for theses t_652_3 and t_652_4):

- mięso wołowe (beef)
- plastry surowej wołowiny poddane działaniu światła fluorescencyjnego (raw beef slices exposed to fluorescent light)
- pakowane w: atmosferze powietrza / bogatej w tlen atmosferze (packed in: normal atmosphere / oxygen-rich atmosphere)
- tlenek cholesterolu 7k (cholesterol oxide 7k)
- tlenek cholesterolu 7-β-OH (cholesterol oxide 7-β-OH)
- tlenek cholesterolu 7-α-OH (cholesterol oxide 7-α-OH)
- tlenek cholesterolu β-epoksydowy (cholesterol oxide β-epoxy)

At this level, the concepts of objects are developed according to the top categories of DOLCE (endurants, perdurants, qualities) and established (for this domain) types of relations. At the first step of the formalization, all classes of endurants, perdurants, qualities, their instances and relations between them are extracted from the object description prepared by the domain experts. They are written out with the assigned serial numbers and a shortcut describing the type of category (endurant “e”, quality “c”, relation “r”). Each object has its English translation in brackets (see example 3.5 below).

Example 3.5 (Pre-ontological description of objects):

- 1.e. mięso (beef)
 - 2.e. bydło (cattle)
 - 3.c. surowe (raw)
 - 4.c. pokrojone w plastry (sliced)
 - 5.e. światło fluorescencyjne (fluorescent light)
 - 6.r. pakowane w (packed in)
 - 7.e. normalna atmosfera (normal atmosphere)
 - 8.e. bogata w tlen atmosfera (oxygen-rich atmosphere)
 - 9.e. tlenek cholesterolu 7k (cholesterol oxide 7k)
 - 10.e. tlenek cholesterolu 7-β-OH (cholesterol oxide 7-β-OH)
 - 11.e. tlenek cholesterolu 7-α-OH (cholesterol oxide 7-α-OH)
 - 12.e. tlenek cholesterolu β-epoksydowy (cholesterol oxide β-epoxy)
- ⟨1, o, 2⟩
 ⟨1, q, 3⟩
 ⟨1, q, 4⟩
 ⟨1, tr, 5⟩
 ⟨1, 6, 7⟩
 ⟨1, 6, 8⟩

Next the relations between objects (with use of given shortcuts) are written out in angle brackets. Firstly, the relations might occur between different objects. In that case we use triples: $\langle o_1, r, o_2 \rangle$, where “ o_1 ”, “ o_2 ” describe objects and “ r ” the relation between them (see example 3.5 above). In some cases we also use quadruples. It is the case for relations such as “is aged for”, “is stored in temperature”, and any other relation which refers to a quality and its value with a proper unit. The pattern for quadruple is the following: $\langle o, r, v, u \rangle$, where “ o ” describes object, “ r ” – relation, “ v ” and “ u ” respectively value and its unit.

For some frequently repeated relations special symbols were established. In our case those symbols are “ o ” and “ tr ”, which describe respectively the relation of “obtained from” (otrzymany z) and “is treated with” (potraktowany).

f) *STEP 7:* From the pre-ontological representation we automatically obtain the description of objects in OWL by translating tuples into axioms of the Science ontology.

IV. KNOWLEDGE MANAGEMENT

After the process described in section III is finished, the Science ontology is populated by scientific laws with their formal description. Then by performing SPARQL queries we may obtain many interesting results.

a) *query 1:* For instance by SPARQL queries we are able to find accepted (IRI so:c0000001) monotonic laws which refer to the same qualities in the same roles but having different monotonicity (see query 1).

query 1

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
```

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX so:<http://onto.beef.org.pl/science/>
```

```
SELECT ?class1 ?class2 ?thesis1 ?thesis2 ?m1 ?m2
WHERE {
?c1 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis2].
```

```
?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis2].
```

```
?thesis1 rdf:type so:c0000015; so:hasMonotonicType ?m1;
rdf:type so:c0000001.
?thesis2 rdf:type so:c0000015; so:hasMonotonicType ?m2;
rdf:type so:c0000001.
```

```
?c1 rdfs:label ?class1. ?c2 rdfs:label ?class2.
FILTER(lang(?class1) = "en"). FILTER(lang(?class2) = "en").
FILTER(?thesis1 != ?thesis2). FILTER(?m1 != ?m2).
}
```

The query gives 16 results in our KB (so 8 pairs of laws); four of them can be seen in table I. For instance the first row of the table informs that the class with english label "weight"@en is an independent parameter in theses with IRI so:t_652_3 and so:t_652_4 and the class with the English label "concentration of cholesterol oxidation products"@en is a dependent parameter in theses with IRI so:t_652_3 and so:t_652_4. But the thesis so:t_652_3 states that weight has a negative impact on concentration of cholesterol oxidation products, while so:t_652_4 states that the impact is positive. This result makes us curious about the reason for the change of monotonicity of the influence from “negative” to “positive”. Having access to the bearers of the mentioned qualities we can learn that weight and cholesterol oxidation products are qualities of meat slices and that in the experiment they were divided into two groups which were packed differently: one in normal atmosphere and the other in highoxygen (32%) atmosphere. So the interesting information is: the way in which meat slices are packed has impact on the direction of influence of weight on the concentration of cholesterol oxidation products.

b) *query 2:* By querying the ontology we are also able to find new laws, i.e. not explicitly expressed in the scientific articles. For instance, the theses which are linked with classes of qualities by so:isIndependentParameterIn and so:isDependentParameterIn roles can be examined in the following way (see query 2): find two functional laws which share the same class of qualities, lets call it a “transitive_class”, such that it is a dependent parameter in the first law and independent in the second; if they are found, the other classes appearing in the two laws—i.e. an independent parameter of the first, lets call it “class1”, and a dependent parameter of the second, lets call it “class2” — create a new law.

query 2

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
```

TABLE I
PARTIAL RESULTS OF QUERY 1

?class1	?class2	?thesis1	?thesis2	?monoton1	?monoton2
"weight"@en	"concentration of cholesterol oxidation products"@en	so:t_652_3	so:t_652_4	negative	positive
"level of cysteine"@en	"flavor"@en	so:t_2303_51	so:t_2303_38	positive	negative
"pressure"@en	"cooking loss"@en	so:t_1913_8	so:t_1913_9	positive	negative
"temperature"@en	"cooking loss"@en	so:t_557_5	so:t_557_6	positive	negative

```

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX so: <http://onto.beef.org.pl/science/>
SELECT ?class1 ?tc_label ?class2
?thesis1 ?m1 ?thesis2 ?m2
WHERE {
?c1 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis1].
?tc rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis2].
?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis2].
?thesis1 rdf:type so:c0000001.
?thesis2 rdf:type so:c0000001.
?c1 rdfs:label ?class1.
?tc rdfs:label ?tc_label.
?c2 rdfs:label ?class2.
FILTER (lang(?class1) = "en").
FILTER (lang(?tc_label) = "en").
FILTER (lang(?class2) = "en").
OPTIONAL {?thesis1 so:hasMonotonicType ?m1.}
OPTIONAL {?thesis2 so:hasMonotonicType ?m2.}
}

```

Table II presents four out of 704 results the query brings about in our KB. In the first row we find a new scientific law stating that tenderness (column ?class1) influences color value (column ?class2). The law is obtained from the two laws: so:t_837_8 and so:t_278_4. The first states that tenderness has impact on pH ultimate value and the second that color value depends on pH ultimate value.

c) *query 3*: We can also ask about qualities which have an influence on a selected quality, for instance a quality with the label "tenderness"@en (see query 3 below).

query 3

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX so: <http://onto.beef.org.pl/science/>
SELECT ?class ?monoton ?article ?thesis
WHERE {
?c rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis].

?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis];
rdfs:label "tenderness"@en.

?thesis rdf:type so:c0000001;
so:comesFromArticleId ?article.

?c rdfs:label ?class.
FILTER (lang(?class) = "en").
OPTIONAL {?thesis so:hasMonotonicType ?monoton.}
}

```

In table III we see partial results of the query. We get the information that aging time, total collagen content, country origin and the amount of connective tissue influence the tenderness of meat. It is also worth noting that in table III we find three different articles which support the existence of correlation between the amount of connective tissue and tenderness.

V. RELATED WORKS

At least a few attempts have been made to address the problem which we have presented in the introduction. Now, we are going to present some of these works which are, in our opinion, interesting and relevant to our work. Firstly, we shall present the SWAN project which “aims to develop a practical, common, semantically structured framework for biomedical discourse initially applied, but not limited, to significant problems in Alzheimer Disease (AD) research” [10, p. 739]. The ontology called SWAN was created within this project. The goals of the SWAN ontology are: 1) to be the schema of a knowledge base in AD research; 2) to link the information from this knowledge base with information from other sources. These goals are intended to be achieved by using the following groups of concepts of the SWAN ontology: people and organizations, discourse elements, bibliographical information, concepts from the domain of interest, and others. The SWAN ontology allows for a representation in which, for example, there is a hypothesis from a scientific publication connected with some journal article by the relation “citesAsEvidence”, with claims by “contains”, and with a person by “authoredBy”. In turn, one of these claims can be connected with another claim by the relation “inconsistentWith”. It is possible, of course, to create these “epistemic” connections among concepts automatically, but in the cited paper the authors seem that they want them to be represented manually. In turn our KB allows for detecting inconsistency between statements automatically.

The second example which we would like to discuss here is the Nanopublications project [11]. The authors of this project argue that the lack of context (i.e. experimental data, citations, argumentation) of a statement (extracted from a scientific paper) which is intended to be represented is a serious problem, because “the statement can only be validated scientifically if you take into consideration its context” [11, p. 51]. And because it is common practice in the Semantic Web to represent a statement without its context (which, traditionally, is implicitly included in a source document), it is important to enable people to add a context to the represented statement.

TABLE II
PARTIAL RESULTS OF QUERY 2

?class1	?tc_label	?class2	?thesis1	?m1	?thesis2	?m2
"tenderness"@en	"pH ultimate value"@en	"color value"@en	so:t_837_8	-	so:t_278_4	positive
"aging time"@en	"pH value"@en	"emulsion capacity"@en	so:t_837_8	-	so:t_1337_2	positive
"rate of spoilage"@en	"fat content"@en	"cooking loss"@en	so:t_910_2	-	so:t_936_3	-
"flaxseed flour content"@en	"protein content"@en	"mean growth rate"@en	so:t_398_6	negative	so:t_2184_10	positive

TABLE III
PARTIAL RESULTS OF QUERY 3

?class	?monoton	?article	?thesis
"aging time"@en	so:positive		so:t_1679_2
"total collagen content"@en	so:negative		so:t_569_21
"country of origin"@en	-		so:t_72_4
"amount of connective tissue"@en	-	"164"	so:t_164_3
"amount of connective tissue"@en	-	"1348"	so:t_1348_5
"amount of connective tissue"@en	so:negative	"1716"	so:t_1716_1

The authors of the project created the nanopublications – a way to connect a statement with its context. A nanopublication is a “set of annotations that refer to the same statement and contains a minimum set of (community) agreed-upon annotations” [11, p. 52]. An annotation is a triple in which the subject is a statement. A statement, in turn, is a triple which can be uniquely identifiable. In such a triple there are three concepts in three different roles (positions): subject, predicate, and object. According to the authors, this model can be realized through the use of Named Graphs which enable adding a URI to a given RDF graph. A statement is a Named Graph and “all annotations belonging to a nanopublication should be part of the same Named Graph” [11, p. 53]. A nanopublication should represent the information from a source publication. In the SWAN ontology the statements were treated as independent components, here the authors of the project laid emphasis on the fact that a publication is a whole, but may still be connected to other publications. The usage of triples for representing statements is a well known practice. In one of the previous sections we have used triples (and, in fact, quadruples) for describing objects. But here we have an additional use of Named Graphs which allows for connecting statements with URIs in triples.

Another example is the Micropublication project within which “a layered metadata model of scientific argumentation and evidence” [12, p. 2] was created in order to “organize, verify, assess, combine and absorb this information [from the world’s biomedical literature] in a comprehensive way” [12, p. 2]. The authors designed this model in such a way that it is able to represent both minimal and maximal forms of publications. In [12] the authors present nine use cases of this model. These use cases show that the model is very flexible and enables a wide range of possible usages: from a simple representation of a citable claim with supporting reference to a representation of a claim with a full chain of supporting evidence, citations etc. Also, the micropublication model allows for expressing statements in a natural language

due to the fact that it is easier for scientists to annotate their work in that way.

Although the above mentioned approaches are interesting, they fail to take into account our specific need, i.e. the ability to represent information in a very specific manner. It is due to the fact that we would like to conduct automatic reasoning on our knowledge base. The SWAN ontology is intended to represent “lonely” statements in a simple formal manner, more or less the same applies to nanopublications. Micropublications can be represented both in formal and natural languages. We want our statements to be not only composed from a few concepts, but they should also be internally structured. This is provided by the taxonomy of scientific laws. And in our representation it is also possible to connect a law with a statement in natural language (e.g. by data property) and with a sentence from the abstract.

VI. CONCLUSION AND PERSPECTIVES

We have presented an ontological representation of the main results of scientific articles from agri-food domain. The knowledge base consists of a TBox in which we define the structure of different kinds of scientific laws present in the domain and an ABox containing over 8k of particular laws. We have described in detail a way in which we proceed from abstracts of journal papers to assertions in ontology.

The work on the knowledge base is still in progress. It can be used as a core element of scientific information systems for the domain. We have not developed any such system in full by now but we have presented in the paper several SPARQL queries that show the kind of information that can be obtained from the knowledge base. The development of applications based on the knowledge base is one of the directions of future works.

The information extracted from the abstracts contains the detailed description of objects to which the laws pertain. The objects are connected to these laws in ontology. However, we have not developed algorithms that make extensive use

of them. Tackling this task is another direction of future work.

Yet another challenge is concerned with the unification of the present results with the earlier ontological works within ProOptiBeef project. They resulted in the Domain ontology [13], [14] containing concepts that were used as keywords in the scientific papers from the projects database. The set of concepts from Domain and the set of objects and features used in Science ABox overlap. We believe that it is worth to merge both ontologies.

ACKNOWLEDGMENT

Research was realized within the Project no. WND-POIG.01.03.01-00-204/09 Optimizing of Beef Production in Poland According to “from Fork to Farm” Strategy co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme 2007 – 2013.

REFERENCES

- [1] T. Nakayama, N. Hirai, S. Yamazaki, and M. Naito, “Adoption of structured abstracts by general medical journals and format for a structured abstract,” *Journal of the Medical Library Association: JMLA*, vol. 93, no. 2, pp. 237–242, April 2005, PMID: 15858627. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15858627>
- [2] P. Kulicki, R. Trypuz, R. Trójczak, J. Wierzbicki, and A. Woźniak, “Ontology-based representation of scientific laws on beef production and consumption,” in *MTSR*, E. Garoufallou and J. Greenberg, Eds., 2013. doi: 10.1007/978-3-319-03437-9_42 pp. 430–439.
- [3] —, “Semantic representation of proved and disproved statements extracted from scientific papers. Meat science case study,” *Information Processing In Agriculture*, pp. 66–72, 2014. doi: 10.1016/j.inpa.2014.06.002
- [4] R. Trójczak, R. Trypuz, and P. Kulicki, “Ontologia praw naukowych w kontekście reprezentacji i udostępniania wyników badań naukowych,” *Filozofia Nauki*, vol. [forthcoming], 2015.
- [5] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari, “WonderWeb Deliverable D18. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology,” 2003. [Online]. Available: <http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf>
- [6] P. Garbacz, M. Lechniak, P. Kulicki, and R. Trypuz, “Do you still want to vote for your favorite politician? Ask Ontobella!” in *In proceeding of: Formal Ontologies Meet Industry, Proceedings of the 4th Workshop FOMI 2009*, 2009. doi: 10.3233/978-1-60750-047-6-102 pp. 102–113.
- [7] R. Mizoguchi, K. Kozaki, and Y. Kitamura, “Ontological analyses of roles,” in *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012. ISBN 978-83-60810-51-4 pp. 489–496. [Online]. Available: <https://fedcsis.org/proceedings/2012/pliki/73.pdf>
- [8] K. Goczyla, A. Waloszek, and W. Waloszek, “Towards Context-Semantic Knowledge Bases,” in *FedCSIS*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012. ISBN 978-83-60810-51-4 pp. 475–482. [Online]. Available: <http://dblp.uni-trier.de/db/conf/fedcsis/fedcsis2012.html#GoczylaWW12>
- [9] A. Martinson, “Guide for the preparation of scientific papers for publication,” UNESCO, Paris, Tech. Rep. PGI-83/WS/10, 1983.
- [10] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, “The SWAN biomedical discourse ontology,” *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 739–751, 2008.
- [11] P. Groth, A. Gibson, and J. Velterop, “The Anatomy of a Nanopublication,” *Information Services and Use*, vol. 30, no. 1-2, pp. 51–56, 2010. doi: DOI 10.3233/ISU-2010-0613
- [12] T. Clark, P. N. Ciccarese, and C. A. Goble, “Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications,” *Journal of Biomedical Semantics*, vol. 5, no. 1, p. 28, 2014. doi: 10.1186/2041-1480-5-28. [Online]. Available: <http://www.jbiomedsem.com/content/5/1/28>
- [13] P. Kulicki, R. Trypuz, and J. Wierzbicki, “Towards beef production and consumption ontology and its application,” *Federated Conference on Computer Science and Information Systems*, pp. 483–488, 2012. [Online]. Available: <http://trypuz.pl/wp-content/papercite-data/pdf/weodia2012.pdf>
- [14] R. Trójczak, R. Trypuz, P. Grądzki, J. Wierzbicki, and A. Woźniak, “Evaluation of beef production and consumption ontology and presentation of its actual and potential applications,” in *FedCSIS*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2013. ISBN 9781467344715 pp. 275–278.
- [15] M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, 2012.