Protein Application Ontology (P-PRO)

H.E.J.M. (Hannelore) Heuer MSc, S.L. (Sander) van Leeuwen MSc, prof.dr.ir. J.L. (Jan) Top

PUBLIC

WAGENINGEN
UNIVERSITY & RESEARCH

Protein Application Ontology (P-PRO)

Authors: H.E.J.M. (Hannelore) Heuer MSc, S.L. (Sander) van Leeuwen MSc, prof.dr.ir. J.L. (Jan) Top

Institute: Wageningen Food & Biobased Research

This study was carried out by Wageningen Food & Biobased Research and subsidised by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWA.1418.20.009).

Wageningen Food & Biobased Research
Wageningen, December 2022

Public

Report 2377
DOI 10.18174/580266

WAGENINGEN
UNIVERSITY & RESEARCH

WFBR Project number: 6230209700
Version: Final
Reviewer: G. (Görkem) Simsek-Senel PhD
Approved by: dr.ir. H. (Henk) Wensink
Carried out by: Wageningen Food & Biobased Research
Subsidised by: Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWA.1418.20.009)
This report is: Public

The research that is documented in this report was conducted in an objective way by researchers who act impartial with respect to the client(s) and sponsor(s). This report can be downloaded for free at https://doi.org/10.18174/580266 or at www.wur.eu/wfbr (under publications).

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

Contents

Summary

Currently the world is facing a large population growth and the effects of climate change, which pose challenges regarding food security, including the need for adequate consumption of proteins. More sustainable protein sources, such as legumes, aquatic crops and insects will have to be used. There is an increasing need for expressing information on proteins and their sources and determining what processes are involved in the production and use of proteins. A number of ontologies already exist to support specific aspects of protein research. In addition, the protein application ontology 'P-PRO' aims to enable and enhance data sharing between the different parties involved in protein research and production.

In this report, we discuss the structure of the P-PRO ontology and how it is embedded in the web of ontologies regarding proteins. P-PRO includes information from the fields of protein technology and human nutrition and health research. There is a link between P-PRO and the Ontology of Units of Measures, OM; the sensor ontology SOSA, several anatomical ontologies and the NCBI Taxon database. P-PRO will be available as an open-source ontology. If P-PRO is adopted widely it can enrich standards that are already established.

1 Introduction

1.1 Background

Currently the world is facing the effects of population growth, climate change and geopolitical conflicts. These problems pose challenges regarding food security. As an alternative to animal-based protein sources, more sustainable protein sources, such as legumes, aquatic crops and insects will have to be used. The P-PRO project is part of the NWA project *Sustainable production of safe and healthy food* that aims to contribute to a radical change to the agricultural food system. One way to contribute to this goal is by focusing on the goal of a fast evolution towards sustainable protein sources such as plant, fungal, algal and insect protein. This emphasises the relationship between proteins and sustainability. Another way to achieve this is by concentrating on the *protein efficiency ratio*, which looks at the optimisation of the digestibility of protein, and is defined as the ratio between the consumed protein and the weight gain (Al-Sagheer, 2022). Therefore, it is important to look at proteins not just by focussing on their genetic composition and taxonomy, but also in their context. In a food context, proteins never exist on their own, but, rather, they are part of a protein containing food product. People do not eat pure proteins, but they are always a component of a protein source. Looking at proteins in this way, by viewing them in their broader context in protein sources, as well as focussing on detailed information about proteins themselves requires interaction between information on all of these levels. Furthermore, the relationship between proteins and sustainability requires the origin of proteins to be clear.

Thus, for the exploration, production and utilisation of new protein sources multiple data sources are needed, supported by information systems and ontologies that allow users to integrate and analyse these data. One way of attaining this is by making use of an ontology. An ontology is a machine and human readable graph of concepts and relationships between those concepts. Having such an ontology allows researchers, product developers, marketeers and others to have access to relevant data and models using their preferred software applications (Rijgersberg, 2012). It facilitates interoperability, one of the key elements of FAIR policy. The FAIR principles were introduced in 2016 to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital resources (Wilkinson, 2016).

There already exist some ontologies about proteins and their structure, such as the protein ontology PRO, gene ontology GO and the protein kinase ontology ProKinO. These ontologies represent knowledge at the level biochemical processes and molecular biology. PRO focuses on the taxonomy of proteins (Darren, 2011); GO on genes and biological processes that produce proteins (Ashburner, 2000; Carbon et al., 2020); and ProKinO provides information about protein kinases (Gosal et al., 2011). Furthermore, ontologies as Pfam and Interpro connect protein data to GO (Blum, 2020).

As is mentioned above, there is an increasing need for expressing information on linking proteins to their sources and determining what processes are involved in the production and use of proteins. Currently, when it comes to sourcing and applying proteins, communication still requires human intervention, which is expensive and error prone. This project supports the shift towards more digital information handling by participating in the development of international data standards, terminology and ontologies that can be used by different stakeholders to accelerate the protein transition. However, concepts and relations describing proteins and their properties at the functional and supply chain level are still lacking. Therefore, this project aims to construct a shared ontology called ‘P-PRO’ (with the first P referring to processing, production, and/or practical use).

1.2 Goals and research question

Our goal is to create a shared ontology called P-PRO, that supports the exploration, production and utilisation of new protein sources. The focus lies on providing the data structures that can be used both by universities and industry to have a standardized shared vocabulary, which they can use when developing their own data models. This enables and enhances data sharing between all different parties involved in protein research and production. With such an ontology in place, researchers, product developers and others can create and have access to interoperable data using their own terminology. This ontology will not only introduce new terms, but will also make use of other existing ontologies.

Our research question reads as follows:
What ontology can we build that enables the interoperability of data regarding protein sources and their practical use?

2 Method

The research question gives rise to two main activities, namely:

1. The exploration of existing ontologies.
2. Defining concepts and relations on proteins and protein sources that are not yet available.

The existing ontologies are used in two ways: as inspiration for the P-PRO ontology, as well as information sources to link to. The extent to which the links are explored can vary, because we anticipate that there are many protein-related data sources.

For the second activity, experts and stakeholders within Wageningen Food and Biobased Research who work in the field of proteins are interviewed to obtain information about which data about proteins is important.

We conducted interviews with the following research domains:

- Protein Technology experts provide knowledge about which processes are involved in protein production and processing, and which data is needed for these processes. These experts collect and structure data on the characteristics of protein-rich raw materials and protein isolates. Their aim is to facilitate a transition from animal-based proteins to plant-based proteins. They consider the functional, nutritional, and sensorial properties of raw materials to enable identification of plant-based protein that can be used in specific applications. Since they measure protein functionality on specific samples, the extraction and measurements on these samples are included in the ontology.
- Human Nutrition and Health experts study to what extent proteins are taken up by the human body. Digestibility measures play a key role in this research domain.
- Plant Research experts study how and which proteins grow in plants.

After obtaining information from these interviews, a first draft of the P-PRO ontology is constructed using the RDFs/OWL language. We use TopQuadrant's TopBraid program to edit and visualise the ontology.

Visit this link https://git.wur.nl/FoodInformatics/p-pro for the GitHub page. The ontology is published on Github along with a document describing the structure and motivation regarding the design of the ontology.

3 P-PRO structure

In this section, we describe the P-PRO ontology in detail. **Error! Reference source not found.** gives a graphical overview of the main concepts and relations.

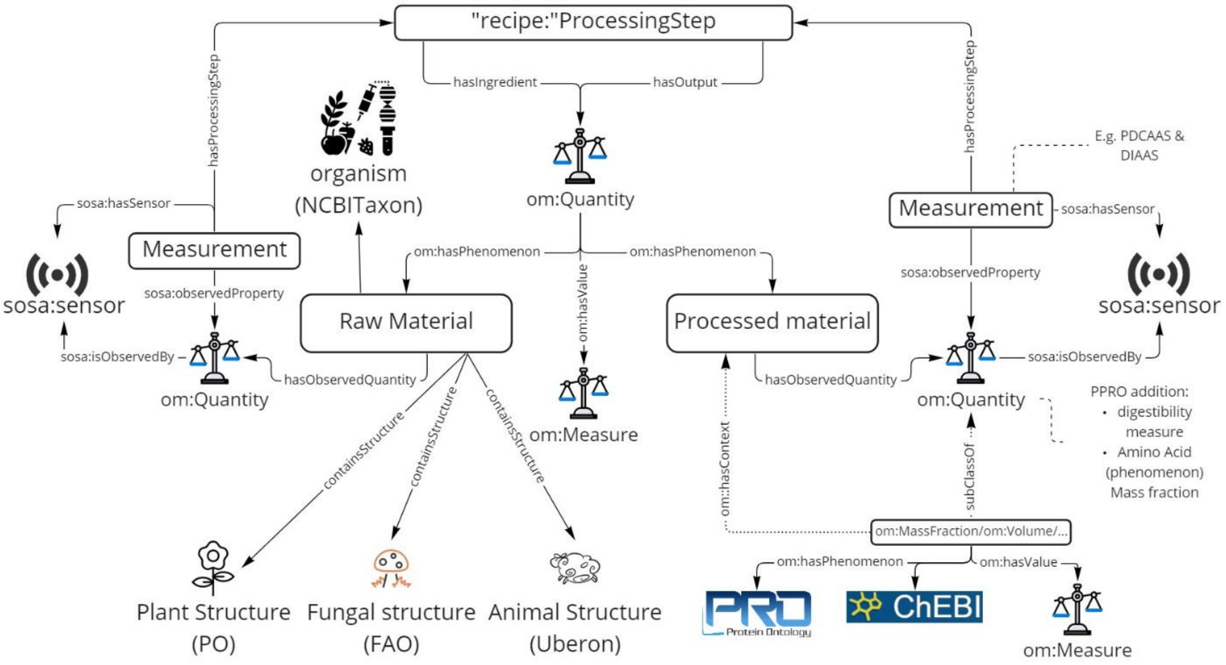


Figure 1 Structure of P-PRO, including links to other ontologies

The main concept within P-PRO is *processed material*. A processed material corresponds to a food substance containing proteins that has undergone at least one processing step in any way. In practice, *sample* and *food substance* are also used. The processed material can have a *brand,* a *sample id* and a *material type*. The processed material can only occur in a certain quantity and thus it is connected to the *quantity* concept from OM. This concept expresses, for example, in which fraction a certain substance, including protein, is present in a processed material. It can also quantify protein functionality, e.g. gelling ability.

In contrast, a *raw material* is a potential food product that is not yet processed and is not composed of more than one organism. It can have a *market class* and a *brand*. This raw material comes from a NCBI Taxon *species,* which can be further specified by its *variety* or *cultivar.* It contains (some part of) a plant, fungus, or animal. They can be linked to the Plant Ontology (PO), fungal structure ontology (FAO) and the animal structure ontology (Uberon). The raw material can reference a whole organism, like a pig, but also a part of an fungus, animal or plant, such as a pea of a certain variety or cultivar.

These concepts of raw and processed material can be used as an ingredient for a *processing step. Processed materials* can be the output of a *processing step*. A processing step is a step that is part of a process. Processing steps include the *method* used, one or more ingredients (*quantity*) and additives (*quantity*), a *duration, concentration, equipment, pH, temperature, material used,* and an output (*quantity*). It also contains a *next step*, so an order of processing steps can be defined.

Furthermore, the raw and processed material can have many *observed quantities*, which are defined as OM quantities, such as solubility, viscosity, etc. Via the *observed quantity*, the amount of protein in the raw and processed material can be specified. The names of these proteins can be taken from PRO or ChEBI (Hastings, 2015). In the observed quantities there are also concepts included about digestibility, such as the Amino Acid Score (AAS). There are specific ways to determine AAS, such as Protein Digestibility Corrected Amino Acid Score (*PDCAAS*) and Digestible Indispensable Amino Acid Score (*DIAAS*). Moreover, P-PRO also includes *sensory intensity*, which combines a specific *sensory category*, such as 'bland', 'burnt', 'cheesy', with a score for each of these categories.

Since it is important to keep track of how these quantities are measured, the concept *measurement* was added. These measurements can (but do not have to be) be done by a *sensor* from the SOSA sensor ontology (Janowicz, 2019), and a measurement can have multiple processing steps. For example, this can be used to describe digestibility measurements, such as PDCAAS and DIAAS, since there are many different methods by which to measure digestibility.

4 Related ontologies

In this section we describe ontologies that are linked to the P-PRO ontology. P-PRO is not the first ontology that includes information about proteins. There are several ontologies that already have included specific information about proteins, and some of these ontologies build upon knowledge from other ontologies by linking to their concepts. Figure shows an overview of all ontologies we have inspected that are related to or relevant for P-PRO. This idea of this overview is based on a similar overview made by FoodOn (see https://foodon.org/).

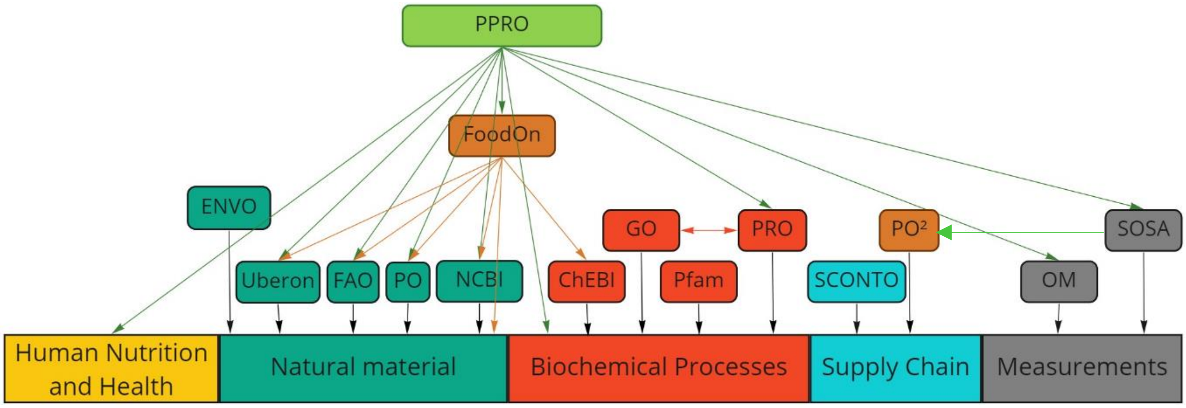
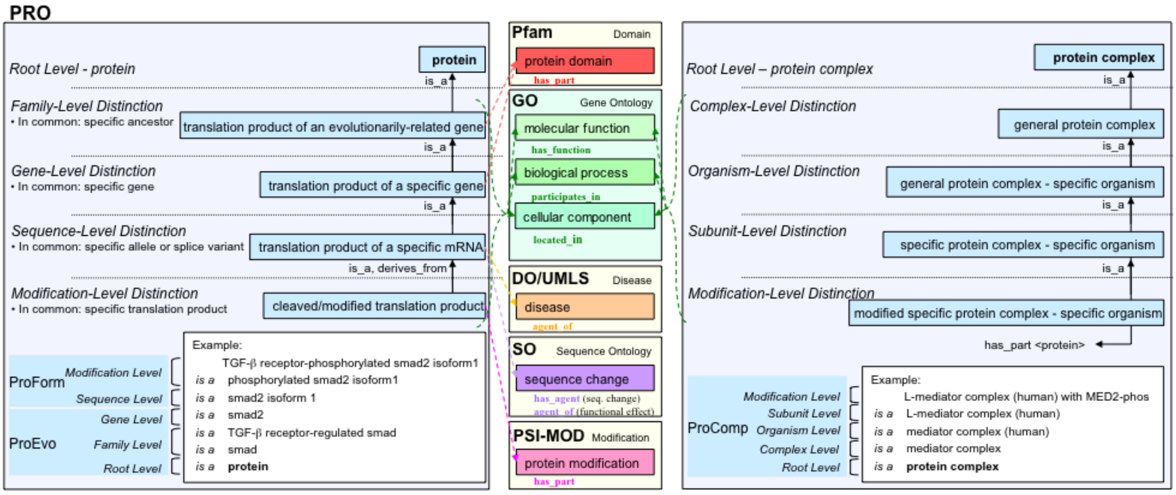


Figure 2 ***Overview of relevant ontologies related to proteins and their application, based on the overview that can be found on the FoodOn website***

Protein ontology

The Protein Ontology (PRO) focuses on protein-related entities and contains three sub-ontologies, namely: (1) proteins based on evolutionary relatedness (ProEvo); (2) protein forms produced from a given gene locus (ProForm); and (3) protein-containing complexes (ProComp) (Chen, 2020). Figure shows an overview from the Proconsortium website of PRO and its contents, as well as ontologies related to PRO.



PRO
Root Level - protein
protein
is_a
Family-Level Distinction
• In common: specific ancestor
translation product of an evolutionarily-related gene
is_a
Gene-Level Distinction
• In common: specific gene
translation product of a specific gene
is_a
Sequence-Level Distinction
• In common: specific allele or splice variant
translation product of a specific mRNA
is_a, derives_from
Modification-Level Distinction
• In common: specific translation product
cleaved/modified translation product
Example:
TGF-β receptor-phosphorylated smad2 isoform1
is a phosphorylated smad2 isoform1
is a smad2 isoform 1
is a smad2
is a TGF-β receptor-regulated smad
is a smad
is a protein
ProForm
Modification Level
Sequence Level
Gene Level
ProEvo
Family Level
Root Level
Pfam
Domain
protein domain
has_part
GO
Gene Ontology
molecular function
has_function
biological process
participates_in
cellular component
located_in
DO/UMLS
Disease
disease
agent_of
SO
Sequence Ontology
sequence change
has_agent (seq. change)
agent_of (functional effect)
PSI-MOD
Modification
protein modification
has_part
Root Level – protein complex
protein complex
is_a
Complex-Level Distinction
general protein complex
is_a
Organism-Level Distinction
general protein complex - specific organism
is_a
Subunit-Level Distinction
specific protein complex - specific organism
is_a
Modification-Level Distinction
modified specific protein complex - specific organism
has_part <protein>
Example:
L-mediator complex (human) with MED2-phos
is a L-mediator complex (human)
is a mediator complex (human)
is a mediator complex
is a protein complex
ProComp
Modification Level
Subunit Level
Organism Level
Complex Level
Root Level

Figure 3 ***Overview of PRO and related ontologies (Proconsortium, 2020). An explanation of the picture can be found at https://proconsortium.org/documents/framework_figure.pdf)***

Pfam database

The Protein Family database (Pfam) is not necessarily an ontology, but rather a large collection of protein families including descriptions of all of these concepts. The structure of the Pfam database is as follows: proteins are generally composed of one or more functional regions, called 'domains'. The way these domains are combined give rise to the diverse range of proteins found in nature and provides insights into their function; Pfam also generates higher-level groupings of related entries, known as 'clans'.

A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-Hidden Markov Model (Mistry, 2020). Therefore, it contains implicit concepts and relations that could be linked.

Gene ontology

The Gene Ontology (GO) is an ontology that goes into detail in the biological domain regarding three aspects, namely (1) molecular function by gene products; (2) cellular components; and (3) biological processes accomplished by several molecular activities, like DNA repair (see Figure). It is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species (The Gene Ontology Consortium, 2018).

FoodOn

The FoodOn ontology is an ontology that can be used to link data to all parts of organisms that can be used for food, food products and the processes used to make them. FoodOn aims to develop semantics for domains that have anything to do with food, e.g., food security, agricultural practices linked to food production and food processing (Dooley, 2018). Therefore, FoodOn connects to many other ontologies to cover these domains. Figure 4**Error! Reference source not found.** shows an overview of all other ontologies linked to FoodOn (FoodOn, n.d.).

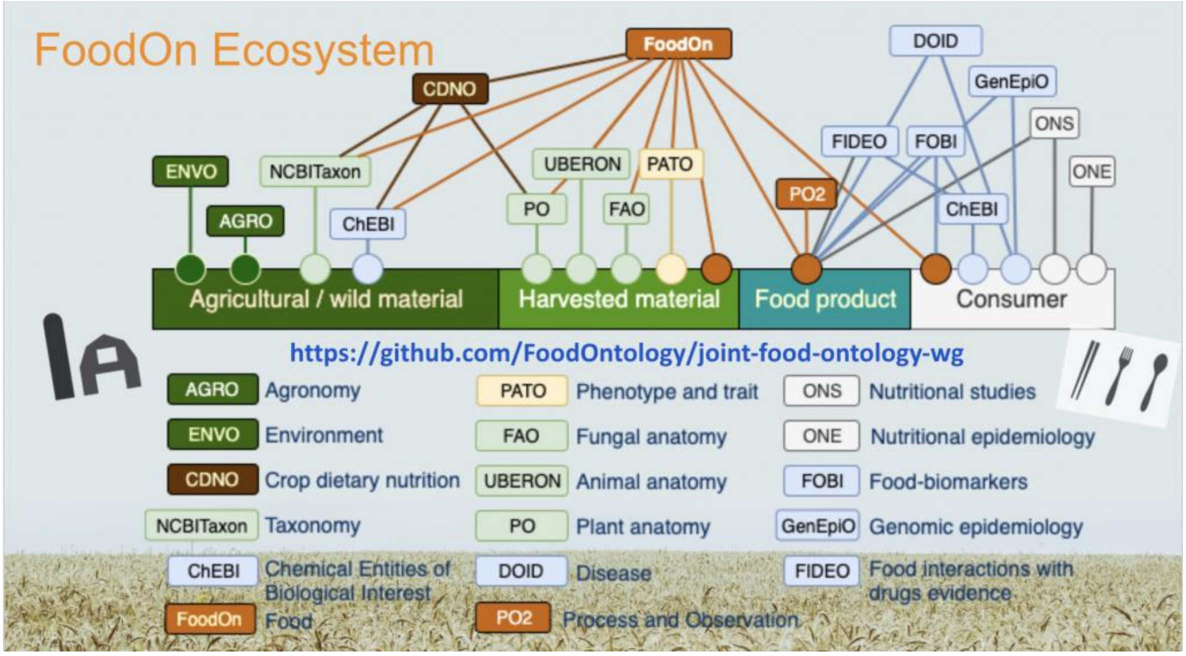


Figure 4 Overview of ontologies linked to FoodOn, see foodon.org (An OBO Foundry Ontology, n.d.)

NCBI Taxon

The NCBI Taxon ontology is developed by the National Center for Biotechnology Information (NCBI). It consists of a taxonomy database, which includes organism names and classifications for every sequence in the nucleotide and protein sequence databases of the International Nucleotide Sequence Database Collaboration (Schoch, 2020). The ontology contains classes of taxons from domain to species level. It also contains a few classes on variety level. This ontology is available on Bioportal (*National Center for Biotechnology Information (NCBI) Organismal Classification | NCBO BioPortal*, n.d.), but we have not been able to access the instance level at this location.

ChEBI

ChEBI stands for Chemical Entities of Biological Interest and is a freely available dictionary of small molecular entities. These entities encompass for example atoms, molecules and ions. The term small implies that the entities used in the dictionary are not directly encoded by the genome, so nucleic acids, proteins and peptides are not included. In the ontology, the relationships between compounds, groups of compounds are saved (Hastings, 2015).

Plant ontology

The plant ontology was developed as part of the common Reference Ontology Project (cROP). It provides a collection of reference and species-specific ontologies for plants and annotations to genes and phenotypes (Cooper, 2018). Part of the reference ontologies for plants are the Plant Ontology (PO); the Plant Trait Ontology (TO); the Plant Experimental Conditions Ontology (PECO); and the Plant Stress Ontology (PSO).

Fungal Structure Ontology

The Fungal gross Anatomy Ontology (FAO) contains descriptions of different structures that form all or part of a fungus (Mungall, 2020). Examples that can be found in the ontology are the hypha, sterigma and ascocarp (Fungal Gross Anatomy Ontology, 2020).

Uberon

Uberon is an ontology that focuses on animal anatomy. It includes body parts, organs and tissues in a variety of animal species. The focus lies on vertebrates. Efforts have already been made to connect it to other ontologies, including the Gene Ontology (Haendel, 2014).

Ontology of Units of Measure (OM)

The ontology of units of measure focuses on units, measures and quantities. It is designed in such a way that a quantity can be linked to a measure, which consists of a numerical value and a unit. Many units and quantities are already defined as instances in the ontology itself (Rijgersberg, 2011).

Sensor ontology (SOSA)

The semantic sensor network (SSN) ontology is an ontology for describing sensors and their observations, feature of interest and observed properties. SSN includes a self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties (Janowicz, 2019).

Supply chain ontologies

PO2 is a process and observation ontology (Ibanescu, 2016). Another related ontology is the Supply Chain Ontology (SCONTO), with the subontology ‘Supply Chain Process Ontology’ (SCOPRO) (Vegetti, 2021). It is a general supply chain ontology, not focussed on proteins or food per se. A direct link between PO2 and P-PRO has not been made, but parts of PO2 have been used as an inspiration of the P-PRO structure. A direct link between SCONTO and P-PRO has not been investigated yet, but since the application of proteins also includes supply chains, it is worth mentioning. This can be further investigated during tests with P-PRO in the industry.

The Environment Ontology

The Environment Ontology (ENVO) is an ontology that describes environmental entities, and can be used in multiple domains including biomedicine, natural and anthropogenic ecology, omics, and socioeconomic development (Buttigieg, 2016). A direct link between ENVO and P-PRO has not been investigated yet either, but ENVO is nevertheless worth mentioning, since sustainability measures are and will be relevant for protein research.

5 Mapping P-PRO concepts to other ontologies

In the previous sections, we introduced the P-PRO ontology (see section 3 and Figure 1) and also described the ontologies that are linked to it (see section 4 and Figure 2). Within these existing ontologies there are already some concepts defined that are similar to concepts in P-PRO. Most of them do not cater to the exact needs we anticipate for protein application. In this section we discuss these concepts and dive into the similarities and differences.

5.1.1 Processing step

P-Pro’s *Processing step* is defined in some form in several ontologies: as *Food transformation process* in FoodOn; *Process* in ScoPro; *Procedure* in Sosa; and *step* in PO2. However, they do not necessarily align with our definition of a processing step. FoodOn’s *Food transformation process* only refers to processes involving a physical transformation from a food product or food source to a food product. In P-PRO's case *processingStep* is broader than that, as it does not only involve these kinds of processes. On the one hand, this more general concept is useful, because more information can be captured within the same concept. On the other hand, when using P-PRO the user should be aware that distinguishing between processing steps for either measurements or food processing can be more difficult. The *Process* defined in ScoPro focuses more on industrial processes, and therefore also does not quite capture the meaning of P-PRO's *Processing step.* Sosa describes a *Procedure,* whose functionality corresponds to the *Processing step* concerning *measurement* in P-PRO. However, it is not entirely clear whether Procedure refers to P-PRO's *method* or *processing step*. *Procedure* is described as a workflow, protocol, plan, algorithm, or computational method (Semantic Sensor Network Ontology (w3.org)). A procedure can be used for an observation, sample or actuation. Our processing step only includes a link to the measurement (loosely linked to sosa:observation) and the sample (linked to sosa:sample). Therefore, Procedure seems to be a broader term of the *processingStep*. The ontology that has a concept most similar to P-PRO's definition of *processingStep* is the PO2 ontology. This ontology describes a *process* (corresponding to P-PRO's *method*) which can have a *step* (corresponding to *processingStep)* with an input and output. However, one of the differences is that they make use of the property *hasForSubStep* to indicate a sub-step of a step. This property has the same domain and range as P-PRO's *hasNextStep* to indicate a next step. Since they have the same domain and range, this might lead to conflicting inferences being made when using PO2 in P-PRO. The concept of a sub-step is an interesting one as it is not included in P-PRO yet, but then we would design it as either a subclass of step or indicating a *skos:broader/skos:narrower* relation between sub-steps and steps.

5.1.2 Raw & processed material

P-PRO's *raw material* is linked to NCBITaxon. However, NCBITaxon only contains classes up until species level. This can cause a problem when you want to define a cultivar or variety. A more extended file of NCBITaxon containing varieties should exist, e.g. different varieties of peas (see https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=3888), but we have not managed to get it. Therefore, P-PRO has its own class describing variety and cultivar if needed. The *variety* and *cultivar* instances of this can also be linked to the NCBITaxon *species*.

In FoodOn multiple concepts can refer to either a *processed* or *raw material.* FoodOn describes *food product, food material, harvested food material, food product by organism, food fermented and multicomponent food product* which can all refer to a *processed material*. FoodOn also has the concept *basic food reference,* which is similar to *raw material.* However, the scope of P-PRO's raw material also includes non-edible substances; the concept is broader than food only, and therefore it is not included in the ontology. FoodOn’s basic food reference can be added as a subclass of raw material or a skos:broader link between them can be introduced.

FoodOn handles multicomponent food products differently than the way P-PRO handles them. A multicomponent food product comes from a process that includes multiple ingredients. If there is a need to label multicomponent food products, they can be detected this way in P-PRO.

In PO2, the input and output of a *step* refer to a *component. Components* can have a *isComposedOf* relation with another *component,* but no distinction is made between an initial or raw material and a processed material (Buche, 2020). PO2's *component* has a broader scope than P-PRO's raw or processed material, because it can also contain processes that are not linked to food or proteins. However, if such a link is necessary for e.g. the packaging of food, it can be a useful addition to P-PRO. PO2's *component* is a subClassOf sosa's *feature of interest*, which is an even more general term.

5.1.3 Measurement

The sensor ontology (ssn/sosa) describes the class *observation,* which is an observation made by a sensor and is similar to the *measurement* class. Since not all *measurements* will be using a sensor, the possibility to add information about sensors is optional. Therefore, we decided not to include the *observation* class, but constructed a *measurement* class instead. *Observation* can be added as a subclass of *measurement.*

Furthermore, P-PRO makes use of the Ontology of units of Measure (OM) as well. P-PRO conforms to the OM structure. This way, the quantity of the output and input of a *processingStep* can be given, as well as the quantity for the measurements. However, for users not interested in quantities and only in ingredients and processes it might be more complicated to use, because they have to address *om:quantity* even though it is not of interest of them. This can be solved by generating a dummy/mock instance for *om:quantity*.

6 Discussion

The P-PRO ontology has been developed due to the need of better communication between the research fields within WUR. The ontology as designed takes much of its inspiration from the PPP Protein Compass project (WFBR project 6229086701, Protein Compass (SFI)). The Protein Compass database contains a lot of detailed information, which was captured with the more general processing step and measurement concepts linked to the processed material. However, to make sure P-PRO covers all the information needed in the protein technology domain, more feedback is needed. Some information from the Protein Compass project was too specific to include, and, therefore, information might be lost due to the applied generalisation.

Protein compass contains a number of properties too specific for P-PRO to include. A good example is fractionation, where Protein Compass makes a distinction between two types of fractionation: whether it is dry or wet fractionation and the fractionation method used. In P-PRO it is not possible to directly indicate (other than by comment in the *processingStep*) to define these specific properties. A solution can be to add wet fractionation to the method class and let the *processingStep* refer to two methods.
Other specific properties exist that can also use this solution (solvent type, injection type). Some specific properties contain additional information on the process or processed material (e.g definition of gel). This information can be added as a note or a comment to a *processingStep* or *processedMaterial* if needed.

In some cases a process is defined as a text describing different processes following each other. To make this compatible with P-PRO the processes describe in the text with their properties should be entered separately as *processingSteps*. This may require some manual entries.
In the nutritional database of Protein Compass more information is given on what type an ingredient is. E.g. sucrose is a sugar. This information is not included in P-PRO, because P-PRO's focus is on proteins. Therefore, there is information present on amino acids, but not on sugars, minerals etc.
In the sensorial database of Protein Compass, different processes are described for sample processing, but only one duration is given. In P-PRO it is possible to give the duration per process, but not for multiple processes in one. A solution can be to define a processing step for sample processing and have the sub-processes refer to this through skos:broader

The ontology PO2 is an ontology for processes, which contains classes and properties concerning processes similar to P-PRO. It would be interesting to see how we could connect this ontology with P-PRO, but due to time constraints it could not be done in this project.

We also would like to include information on supply chains. We could make a connection with SCONTO to achieve this. However, to include this connection in a correct way, we would need data on supply chains to test whether SCONTO would be correctly connected to P-PRO. Unfortunately, we did not have time to gather such data for this project.

Another interesting aspect to include would be information on sustainability of proteins. E.g. we would like to express the sustainability of plant and animal proteins. More research is needed whether this can be done with other ontologies like the ENVO ontology.

7 Conclusions

In conclusion, we can say that the P-PRO ontology connects interesting concepts in the protein domain with other existing ontologies. We have looked at the links between P-PRO and other ontologies to make sure it can be an addition to existing ontologies.

As was mentioned before, in a food context, proteins never exist on their own, but, rather, they are part of a protein containing food product. People do not eat pure proteins, but they are always a component of a protein source. The P-PRO ontology captures this coherence by these links. Thereby, it strengthens the link between protein research and research that is (getting more) related to proteins. P-PRO facilitates transparent and independent data connection and expresses data in a machine-readable way. In this way, it enables the interoperability of data regarding protein sources and their practical use.

7.1 Recommendations

To make sure the P-PRO ontology facilitates the interoperability of protein data, there is more that can be done. First of all, by applying the knowledge captured in P-PRO in various applications. Currently it is mainly tested on the Protein Compass project, but developing more applications using P-PRO will increase its quality.

Second, by showing a use case of linking data from various sources. One way to do this would be by checking if the information provided by the P-PRO ontology is adequate for a business setting as opposed to the research environment it is tested in now. To do this properly, we would need example data from different parties collecting data on proteins outside of academia. Information on supply chains and sustainability are examples of information that can still be added to P-PRO. Companies involved in the supply chain for protein-based products or researchers investigating the sustainability of proteins can be interviewed to get this information.

Literature

Al-Sagheer, A. A., Abdel-Rahman, G., Elsisi, G. F., & Ayyat, M. S. (2022). Comparative effects of supplementary different copper forms on performance, protein efficiency, digestibility of nutrients, immune function and architecture of liver and kidney in growing rabbits. Animal Biotechnology, 1–11. https://doi.org/10.1080/10495398.2022.2084746

An OBO Foundry ontology. (n.d.). FoodOn. Retrieved 4 July 2022, from https://foodon.org/

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., . . . Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/10.1093/nar/gkaa977

Buche, P., Cufi, J., Dervaux, S., Dibie, J., Ibanescu, L., Oudot, A., & Weber, M. (2020, September). Food transformation process description using PO2 and FoodOn. In *Integrated Food Ontology Workshop (IFOW)@ ICBO*.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, *7*(1). https://doi.org/10.1186/s13326-016-0097-6

Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., . . . Elser, J. (2020). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Chen, C., Huang, H., Ross, K. E., Cowart, J. E., Arighi, C. N., Wu, C. H., & Natale, D. A. (2020). Protein ontology on the semantic web for knowledge discovery. *Scientific Data*, *7*(1). https://doi.org/10.1038/s41597-020-00679-9

Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., & Jaiswal, P. (2017). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, *46*(D1), D1168–D1180. https://doi.org/10.1093/nar/gkx1152

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., Schriml, L. M., Brinkman, F. S. L., & Hsiao, W. W. L. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Science of Food*, *2*(1). https://doi.org/10.1038/s41538-018-0032-6

Fungal Gross Anatomy Ontology. (2020, July 5). Bioportal. Retrieved 5 July 2022, from https://bioportal.bioontology.org/ontologies/FAO?p=classes&conceptid=root

Gosal, G., Kochut, K. J., & Kannan, N. (2011). ProKinO: An Ontology for Integrative Analysis of Protein Kinases in Cancer. *PLoS ONE*, *6*(12), e28782. https://doi.org/10.1371/journal.pone.0028782

Haendel, M. A., Balhoff, J. P., Bastian, F. B., Blackburn, D. C., Blake, J. A., Bradford, Y., Comte, A., Dahdul, W. M., Dececchi, T. A., Druzinsky, R. E., Hayamizu, T. F., Ibrahim, N., Lewis, S. E., Mabee, P. M., Niknejad, A., Robinson-Rechavi, M., Sereno, P. C., & Mungall, C. J. (2014). Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics*, *5*(1), 21. https://doi.org/10.1186/2041-1480-5-21

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2015). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Ibanescu, L., Dibie, J., Dervaux, S., Guichard, E., & Raad, J. (2016, November). PO^2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels. In *Research Conference on Metadata and Semantics Research* (pp. 155-165). Springer, Cham.

Janowicz, K., Haller, A., Cox, S. J., le Phuoc, D., & Lefrançois, M. (2019). SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, *56*, 1–10. https://doi.org/10.1016/j.websem.2018.06.003

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mungall, C., & Harris, M. (2020, May 7). *fungal-anatomy-ontology: A structured controlled vocabulary for the anatomy of fungi*. GitHub. Retrieved 5 July 2022, from https://github.com/obophenotype/fungal-anatomy-ontology/

National Center for Biotechnology Information (NCBI) Organismal Classification | NCBO BioPortal. (n.d.). https://bioportal.bioontology.org/ontologies/NCBITAXON

Ontologies | Planteome. (n.d.). Planteome.Org. Retrieved 5 July 2022, from https://planteome.org/node/1

Proconsortium (2020). *Protein Ontology*. Retrieved 9 May 2022, from https://proconsortium.org/

The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. (2018). *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Rijgersberg, H., Wigham, M., & Top, J. (2011). How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, *25*(2), 276–287. https://doi.org/10.1016/j.aei.2010.07.008

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, *2020*. https://doi.org/10.1093/database/baaa062

Vegetti, M. M., Böhm, A., Leone, H. P., & Henning, G. P. (2021, March). SCONTO: A modular ontology for supply chain representation. In *Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing*.

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi-org.ezproxy.library.wur.nl/10.1038/sdata.2016.18

WFBR project number: 6229086701, "Protein Compass (SFI)" , TKI-AF-18017

To explore
the potential
of nature to
improve the
quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
E info.wfbr@wur.nl
wur.nl/wfbr

Report 2377

The mission of Wageningen University & Research is “To explore the potential of nature to improve the quality of life”. Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 6,800 employees (6,000 fte) and 12,900 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.