



**HAL**  
open science

## Mapping bread wheat trait ontologies for semantic interoperability

Claire Nédellec, Sophie Aubin, Clara Sauvion, Liliana Ibanescu, Sonia Bravo, Jacques Le Gouis, Thierry C Marcel, Cyril Pommier, Robert Bossy, Michael Alaux

► **To cite this version:**

Claire Nédellec, Sophie Aubin, Clara Sauvion, Liliana Ibanescu, Sonia Bravo, et al.. Mapping bread wheat trait ontologies for semantic interoperability. 2024. hal-04717147

**HAL Id: hal-04717147**

**<https://hal.science/hal-04717147v1>**

Preprint submitted on 1 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.









Distributed under a Creative Commons Attribution 4.0 International License



## RESEARCH ARTICLE

# Mapping bread wheat trait ontologies for semantic interoperability [version 1; peer review: awaiting peer review]

Claire Nédellec <sup>1</sup>, Sophie Aubin <sup>2</sup>, Clara Sauvion<sup>1</sup>, Liliana Ibanescu <sup>3</sup>,  
Sonia Bravo<sup>2</sup>, Jacques Le Gouis<sup>4</sup>, Thierry C. Marcel <sup>5</sup>, Cyril Pommier <sup>6,7</sup>,  
Robert Bossy<sup>1</sup>, Michael Alaux <sup>6,7</sup>

<sup>1</sup>MaIAGE, Université Paris-Saclay, INRAE, Jouy-en-Josas, France

<sup>2</sup>DipSO, INRAE, Paris, France

<sup>3</sup>Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, Palaiseau, France

<sup>4</sup>INRAE, Université Clermont Auvergne, UMR GDEC, Clermont-Ferrand, France

<sup>5</sup>Université Paris-Saclay, INRAE, UR BIOGER, Palaiseau, France

<sup>6</sup>Université Paris-Saclay, INRAE, URGI, Versailles, France

<sup>7</sup>BioinfOmics, Plant Bioinformatics Facility, Université Paris-Saclay, INRAE, URGI, Versailles, France

**V1** First published: 27 Sep 2024, 13:1102  
<https://doi.org/10.12688/f1000research.154860.1>

Latest published: 27 Sep 2024, 13:1102  
<https://doi.org/10.12688/f1000research.154860.1>

## Open Peer Review

**Approval Status** *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

## Abstract

### Background

The Wheat Crop ontology was created to annotate phenotypic experimental data (i.e. field and greenhouse measurements standardized and integrated in databases). The Wheat Trait and Phenotype ontology was created to annotate information on wheat traits from the literature (i.e. text found in the abstract, results and discussion of scholarly articles). To enable seamless data retrieval on wheat traits from these complementary sources, the classes in the two ontologies have been aligned.

### Methods

All pairs of ontology classes were examined and categorized in nine groups based on the nature of their relationships (e.g. equivalence, subsumption). General principles emerged from this process which were formalized into rules. The Simple Standard for Sharing Ontological Mappings (SSSOM) representation was chosen to represent the mappings in RDF (Resource Description Framework), including their metadata such as creators, reviewers, and justification (including rules).

## Results

The mapping dataset is publicly available. It covers 77% of the ontology classes. Most labels of the aligned classes differed significantly and required domain expertise for decisions, especially for traits related to biotic stress. Consequently, most mappings are close mappings rather than exact equivalents.

## Conclusions

We present the end-to-end manual process used to select and represent mappings in SSSOM within the specific domain of wheat traits. We derive general lessons from the complex alignment process that extend beyond the specific case of these two ontologies and more generally apply to alignments of specialized ontologies for information retrieval purposes. This work demonstrates the relevance of SSSOM for representing these mappings.

## Keywords

bread wheat traits and phenotypes, data interoperability, information retrieval, ontology alignment, ontology mapping representation, wheat breeding

**Corresponding authors:** Claire Nédellec ([claire.nedellec@inrae.fr](mailto:claire.nedellec@inrae.fr)), Michael Alaux ([michael.alaux@inrae.fr](mailto:michael.alaux@inrae.fr))

**Author roles:** **Nédellec C:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation; **Aubin S:** Conceptualization, Funding Acquisition, Methodology, Software, Supervision, Writing – Original Draft Preparation; **Sauvion C:** Data Curation, Writing – Review & Editing; **Ibanescu L:** Conceptualization, Methodology, Writing – Original Draft Preparation; **Bravo S:** Data Curation, Software, Writing – Original Draft Preparation; **Le Gouis J:** Data Curation, Writing – Review & Editing; **Marcel TC:** Data Curation, Writing – Review & Editing; **Pommier C:** Conceptualization, Data Curation, Funding Acquisition, Validation, Writing – Original Draft Preparation; **Bossy R:** Formal Analysis, Resources, Software, Writing – Review & Editing; **Alaux M:** Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** The project was funded by Agence Nationale de la Recherche under grant number [ANR-18-CE23-0017 D2KAB] *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2024 Nédellec C *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Nédellec C, Aubin S, Sauvion C *et al.* **Mapping bread wheat trait ontologies for semantic interoperability [version 1; peer review: awaiting peer review]** F1000Research 2024, 13:1102 <https://doi.org/10.12688/f1000research.154860.1>

**First published:** 27 Sep 2024, 13:1102 <https://doi.org/10.12688/f1000research.154860.1>

## Introduction

Plant phenotyping aims to understand how genetic variations and environmental factors influence plant traits. It plays a critical role in crop improvement and precision agriculture to enable sustainable food production and adaptation to climate change. Protocols are being developed to measure phenotypes of plants such as wheat in order to speed up breeding. Wheat is one of the most widely cultivated cereal crops playing a vital role in human nutrition, agriculture, cultural and economic development.<sup>1,2</sup>

Phenotyping generates large amounts of heterogeneous data such as field and greenhouse observations and measurements.<sup>3</sup> These phenotypic experimental data include raw and computed traits standardized and FAIRified in database.<sup>4</sup> They inform and validate scientific conclusions published in scholarly articles. Therefore, phenotypic experimental data and the description of their corresponding traits in scholarly articles differ in scope and type. On the first hand, experimental data quantify measurable properties of the plant within a limited spatial and temporal scope (cf. Table 1). On the other hand, traits descriptions in the text of scholarly articles qualify the characteristics of wheat varieties (cf. Figure 1). Table 1 gives an example of the observation data available in the GnpIS information system: <https://urgi.versailles.inrae.fr/ephep/ephep/viewer.do#dataResults/trialIds=801>.

It has been collected for the trial BTH\_Lusignan\_2014\_SetB1 in Lusignan, France in 2014. The Septoria score (Septoria tritici blotch incidence) value of Apache variety for this specific trial is 5 on a scale 1 to 9, from resistant to susceptible.

The example from the scholarly article<sup>5</sup> in Figure 1 summarizes the conclusion of the analysis of several trials on resistance phenotypic trait of Apache variety to Septoria tritici blotch disease. Compared to experimental data, this excerpt presents a more general and synthesized finding.

The aim of our study is to integrate information coming from these two sources, (1) experimental data, (2) scholarly articles, and to propose generic and reusable methods and tools beyond our use case. Our operational goal is their integration in the WheatIS and FAIDARE data discovery portals (<https://urgi.versailles.inrae.fr/wheatis/>, <https://urgi.versailles.inrae.fr/faidare/>).<sup>6</sup> They are web-based one-stop portals developed to access all available plant data resources that result from the international initiatives ELIXIR (<https://elixir-europe.org/>) and the WheatIS expert working group of the Wheat Initiative (<https://www.wheatinitiative.org/>).<sup>7</sup> The data interoperability is ensured by semantic indexing by ontologies.

The two sources of experimental and scholarly articles data we considered in this study use specific conceptualizations (concepts and relations), and are managed in two different information systems, but both address bread wheat phenotyping.

The two types of data are indexed by two ontologies, respectively, the Wheat Crop Ontology (CO\_321)<sup>8</sup> and the Wheat Trait and Phenotype Ontology (WTO; <http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE>).<sup>9</sup> For example, the experimental data scores of Table 1 are indexed by CO\_321:0000919 which label is *Septoria tritici blotch incidence* measured on the whole plant by the direct measurement method. The scholarly article example of Figure 1 is indexed by

**Table 1. Extract of the data collected in the trial BTH\_Lusignan\_2014\_SetB1 in Lusignan, France in 2014.**

Lot Number	Accession Number	Accession Name	sept: Septoria score
Grapeli	36801	GRAPELI	3
AO13031	AO13031	AO13031	3,5
Apache	13481	<b>APACHE</b>	<b>5</b>
AO14001	AO14001	AO14001	5,5

The Septoria score value of Apache variety is 5.

*The present data show that both cvs. Apache and Balance contributed specific resistance to the DH population. The resistance in both parents could be easily differentiated using the 30 M. graminicola isolates panel [...]*

**Figure 1. Extract of the scholarly article<sup>5</sup> about Apache resistance to Septoria tritici blotch disease.** The specific resistance value of the trait has been drawn from the synthesis of multiple trials measures.

WTO:0000554 *resistance to Septoria leaf blotch*. The choice of ontology for indexing both data sources was made based on historical and conceptual considerations, specifically due to differences in the scope and nature of their conceptual classes. In order to fully capitalize on the complementary attributes of the two primary data sources, the essential goal of facilitating federated search requires mapping the classes of the two ontologies. Indeed, the WheatIS and FAIDARE data portals offer full-text search and resource annotation that can both take advantage of a mapping between CO\_321 and WTO to easily find experimental and scholarly article trait-related data.

The conceptual differences between the two ontologies make the alignment a complex task. Although their hierarchical structures are similar, organizing the trait classes from general to specific, the trait classes of CO\_321 and WTO represent different levels of abstraction and information aggregation. The first is tailored to the needs of data producers and experimenters, while the latter is organizing scientific knowledge as synthesized in scholarly articles.

Indeed, WTO mostly defines and ontologically organizes atomic concepts that represent different traits, while Crop Ontology is a list of phenotypic variables based on the entity-attribute-value (EAV) model. Each variable is a triplet semantically aggregating a trait, a measurement method and a unit or scale.

(1) is an illustrative example.

- (1) The single *stay-green* trait of WTO characterizes the wheat variety's general capability to maintain its green leaves and photosynthetic capacity after anthesis. In CO\_321 multiple traits qualify the measurement of observable vegetative greenness and canopy photosynthetic size, including the *Normalized difference vegetation index* or the *Canopy green normalized difference vegetation index*. Each of these traits is associated to a method and scale.

Therefore, to enable an integrated search of heterogeneous data in WheatIS and FAIDARE data portals, we studied the conceptual differences, defined alignment procedures and identified all alignments between the WTO trait classes and the CO\_321 trait classes.

Before describing the alignments between the two ontologies (Result section), we will introduce the method used to obtain them and the Simple Standard for Sharing Ontological Mappings (SSSOM) representation (Method section). The next section (Background section) is dedicated to the two ontology specificities and alignment methods and representation.

## Background

### Ontologies

The Crop Ontology (CO) was created by the CGIAR in 2008 to describe the traits of 17 crops. (<https://cropontology.org/>). CO\_321 is dedicated to bread wheat ([https://github.com/Planteome/CO\\_321-wheat-traits](https://github.com/Planteome/CO_321-wheat-traits)). It contains 467 trait classes and 87 synonyms in 16 subtrees of depth 2. The crop ontologies follow a conceptual model that defines a phenotypic variable as a combination of a trait, a method and a scale. The CO format is adopted by the Minimum Information About a Plant Phenotype Experiment (MIAPPE, <https://www.miappe.org/>)<sup>10</sup> and the Breeding Application Programming Interface (BrAPI, <https://brapi.org/>).<sup>11</sup> CO\_321 has been used for the phenotypic variables of the experimental data generated by the European Whealbi (<https://wheat-urgi.versailles.inrae.fr/Projects/Achieved-projects/Whealbi>) (12 trials) and the French BreedWheat (<https://wheat-urgi.versailles.inrae.fr/Projects/BreedWheat>) (60 trials) projects.

The Wheat Trait Ontology (WTO) was developed by the MaIAGE research laboratory within the SamBlé project to facilitate the annotation of textual data using ontologies. The WTO is accessible on AgroPortal (<https://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE/>) in both OBO and OWL formats. It comprises two main sections: the Environmental Condition subtree, which includes 236 classes, and the Plant Property subtree. The latter is further divided into the Trait subtree (WTO:0000006) and the Phenotype subtree (WTO:0000005), resulting in a total of 514 classes. These classes encompass all aspects of plant properties across six primary categories: development, growth, morphology, quality, reproduction, and response to environmental conditions. The WTO model is deeply structured to facilitate navigation and reuse for data and knowledge discovery. The maximum depth of WTO is 9 and includes 486 synonyms. The plant property trait and phenotype classes are respectively used to annotate the trait and phenotype values in scientific documents.<sup>12</sup> The SamBlé application provides an AlvisIR-based search engine where plant properties are linked to genes and varieties in PubMed references ([https://bibliome.migale.inrae.fr/wheat/alvisir/webapi/search?q=wheat+LR\\*+%28resistance+to+pathogen%29](https://bibliome.migale.inrae.fr/wheat/alvisir/webapi/search?q=wheat+LR*+%28resistance+to+pathogen%29)).

Both ontologies are presently employed within the FAIDARE and WheatIS portals for indexing wheat experimental and PubMed literature data. This functionality enables users to filter data using the Ontology Annotation facet and search for results associated with ontology terms (e.g., mapped, synonyms). However, the federated information retrieval currently utilizes only the alignment of 59 class labels that are identical in both ontologies. The presence of many more shared concepts could be exploited with the creation of new formal alignments.

### Ontology alignment

Ontologies provide a formal description of a set of concepts and the relations between them in a domain area. Therefore, it represents the shared meaning of this domain, that is machine-processable and allows reasoning, i.e. generation of new knowledge, and automatic detection of inconsistencies in the semantic model. Ideally, each domain should be described by a single ontology, but numerous overlapping ontologies exist, having their own scope, and purposes and annotating a specific type of information. This is the case for WTO and CO\_321 ontologies.

Ontology matching is the process of identifying semantic correspondences between ontology concepts and relations.<sup>13</sup> A number of automatic and semi-automatic ontology matching methods were developed.<sup>14–16</sup> Since 2004, OAEI (*ontology alignment evaluation initiative*) (<https://oaei.ontologymatching.org/>) organizes annual campaigns aiming at their evaluations on different test sets. Results for OAEI 2023 were published in<sup>17</sup> and highlights the diversity of the methods.

As demonstrated in Ref. 18, the majority of semi-automated and automated alignment techniques primarily focus on identifying one-to-one equivalence and subsumption relationships. However, real-world ontologies often require more sophisticated set operations, such as union, intersection, disjunction, and cardinality restriction, collectively referred to as complex alignments. The evaluation of complex alignments was first introduced in OAEI 2018,<sup>19</sup> which underscored the absence of benchmarks that include complex relations, as well as the lack of appropriate metrics for assessing the outcomes of ontology matching techniques. Both WTO and CO\_321 are available on the AgroPortal ontology repository.<sup>14</sup> AgroPortal integrates a basic ontology alignment method, called LOOM, which finds 271 mappings between WTO and CO\_321 classes.

The 2021 version of AgreementMakerLight (AML) [<https://www.semantic-web-journal.net/content/agreementmaker-light-0>], one of the ontology matching available systems, found 93 mappings. A deep analysis of the complexity of potential alignments with respect to the state-of-the-art convinced us to build them manually.

### Alignment representation

In order to make the result of the alignment task (re) usable by both humans and machines, the choice of its representation is essential. Taking into account the goals of the application that exploits the WTO and CO\_321 alignments as well as the community recommendations to produce FAIR mappings,<sup>20</sup> the target representation model should meet several needs. The mappings need to be documented with comprehensive and adequate metadata, including 1) the purpose and application domain of the alignment, here information retrieval, 2) the alignment method used, and 3) the scientific justification for each mapping. The model must allow the use of standard, possibly oriented, mapping predicates from the semantic web community, like *owl:equivalentClass* and *skos:match* and the representation of complex (1 to N) mappings. As for other data, it is also important that resources involved in the alignment are precisely identified, successive versions of the alignment are marked, and the authors and reviewers of the mappings are credited for their work.

We found a couple of representation models that were good candidates to represent alignments independently of the semantic artifacts aligned. The summary of their evaluation with respect to the needs is given in [Table 2](#).

First of all, as the ontologies were already hosted in the AgroPortal, we considered the model used in this repository of ontologies. Indeed, AgroPortal, like other OntoPortal instances, makes it possible to store, describe and retrieve alignments between the hosted resources. It includes some elements of context such as the author (meta:creator) or an open text field to provide additional details (meta:comment). But it quickly became clear that the metadata available in this model did not sufficiently document the alignment. For example, we could not indicate a direction in the alignment, which can be problematic in the case of “narrower” or “broader” mappings.

The *Alignment Format Model*<sup>13</sup> allows complex alignments to be represented but it does not fully meet the needs formulated above. Its extension, the *Expressive and Declarative Ontology Alignment Language* (EDOAL) model is much more complete [<https://moex.gitlabpages.inria.fr/alignapi/edoal.html>] but also more focused on complex alignments. Still, it is dedicated to research purposes and evaluation challenges.

**Table 2. Evaluation of the alignment models with respect to the project criteria.**

Needs	Criteria	Model of Ontoportal	Alignment format model	EDOAL	SSSOM
To know what is aligned	Ontology: identifier	Y	Y	Y	Y
To know what is aligned	Ontology: version	N	N	N	Y
To know what is aligned	Concept/Class: identifier	Y	Y	Y	Y
To be user friendly	Concept/Class: preferred label	Y	N	N	Y
To be interpretable	A direction between the two concepts	N	Y	Y	Y
To be interpretable	Alignment predicate from various standards (owl, SKOS)	Y	N	N	Y
To be interpretable	Cardinality (1:1, 1:n)	N	Y	Y	Y
To be interpretable	Restriction of application of mapping	N	in an add-on	Y	being
discussed					
To be trustworthy	Justification of the mapping	N	N	N	Y
To be trustworthy	Method used (tool, algorithm, etc)	Y	Y	Y	Y

Expressive and Declarative Ontology Alignment Language (EDOAL); Simple Standard for Sharing Ontological Mappings (SSSOM).

During the benchmarking work, the scholarly article “A Simple Standard for Sharing Ontological Mappings (SSSOM)” was published, presenting the model developed by the biomedical community involved in the OBO Foundry.<sup>21</sup> This standard meets the majority of our expectations and the SSSOM community is open to suggestions for improvement. SSSOM was developed to meet several objectives: to offer a standard easy-to-use representation, with rich metadata to best represent alignments and facilitate their understanding, integration and reusability. For example, the SSSOM framework provides both RDF (*Resource Description Framework*) and OWL (*Web Ontology Language*) serialization for people working in the Semantic Web framework and a TSV format for a wider audience, including domain experts.

We therefore used the SSSOM framework to represent the alignment set. Compared to others, this emerging format offers several advantages for both the alignment producers and users as discussed above. Yet, SSSOM does not allow the representation of compound (or complex) mappings, which has been identified as a current limitation by the authors and is subject to discussions in fora and community events to propose evolutions.

Interestingly, SSSOM meets our need to share alignments that respect the FAIR principles and the open science approach. In addition, it is strongly endorsed within the *European Open Science Cloud* (EOSC, <https://eosc.eu/>) and other international projects to facilitate the interoperability of participating information systems that share scientific data.

## Methods

### Alignment principle

The overall goal for aligning the classes of WTO and CO\_321 ontologies is to enable information retrieval from the indexed datasets regardless of the ontology used to index the data. The target users are researchers, experimenters and breeders who are not familiar with the indexing and inference principles. We defined the following competency questions that cover various topics and are representative of their needs.

- CQ1: Which scholarly articles and experiments pertain to insect resistance in wheat varieties?
- CQ2: Which scholarly articles and experiments evaluate the flour protein content of wheat varieties?
- CQ3: Which scholarly articles discuss the resistance of wheat varieties to specific diseases or pathogens, and which experiments report on wheat responses?

- CQ4: Which variety has the highest grain manganese content?
- CQ5: Which variety exhibits the greatest soil coverage?
- CQ6: What scholarly articles and experiments are available on the milling quality of wheat?

To answer these kinds of questions, information retrieval involves two types of inference, (1) the usual subsumption mechanism that retrieves the data indexed by all classes that are *logically subsumed* by the class of the query and (2) the logical equivalence that retrieves the data indexed by all classes that are *logically equivalent* to the class of the query. The first involves inference within the same ontology while the second uses the alignment between the classes of two ontologies. The two combined inference mechanisms go down into the ontologies and back and forth from one ontology to the other. The answer set of a user query class C contains the data annotated with C and all its subclasses C, and the data annotated with the aligned classes of C and C and their subclasses. Figure 2 shows the graphical representation of an abstract example.

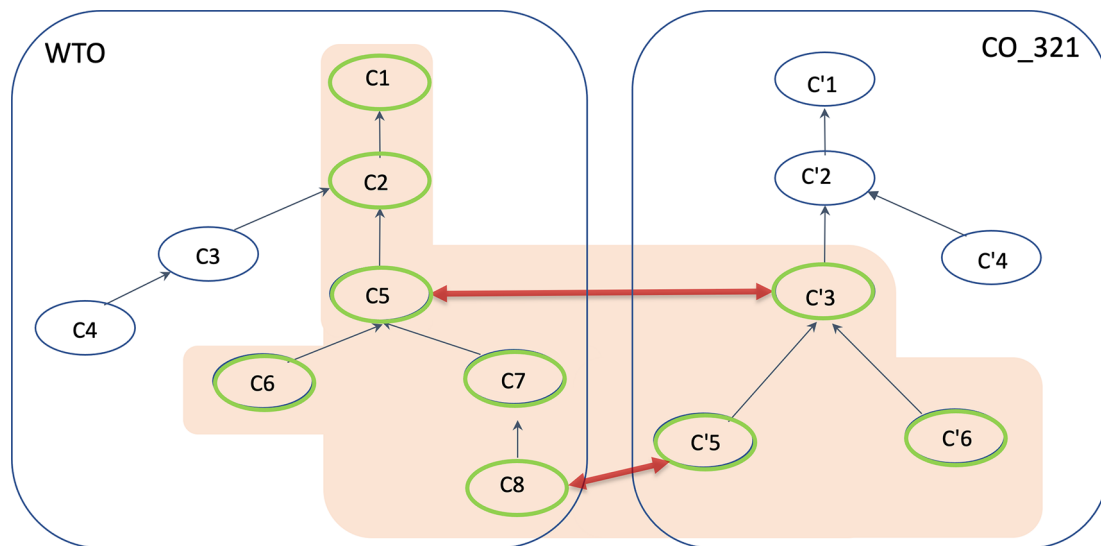
To illustrate this with an example, we handle the competency question CQ6 “What scholarly articles and experiments are available on the milling quality of wheat?”. Figure 3 illustrates that the *Milling quality* class in WTO has two subclasses: *Amount of damaged starch* and *Moisture content of grain*. *Milling quality* in CO\_321 is a subclass of *Quality trait*, while *Grain moisture content* is a subclass of *Agronomical trait*. We identified two alignments:

<align1, WTO:Milling quality, CO\_321:Milling quality>

<align2, WTO:Moisture content of the grain, CO\_321:Grain Moisture content>

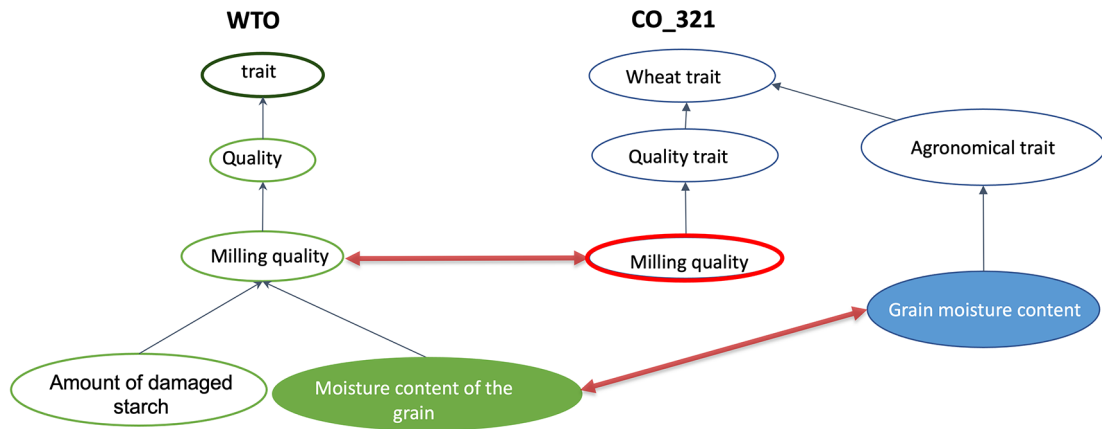
The answer for the query Q:“milling quality” retrieves the scholarly article information indexed by WTO:Milling quality and its subclasses and the experimental data indexed by CO\_321:Milling quality according to align1 alignment. Using align2 alignment, information retrieval also gathers the experiments indexed by CO\_321:Grain moisture content.

It is worth noting that the procedure is formally complete. For instance, although the ancestors of the two classes “CO\_321:Grain moisture content” and “WTO:Moisture content of the grain” are different, the answer set will be the same regardless of the starting class of the query belonging to WTO or CO\_321. Our mapping decision principle is task-oriented. Two classes will be declared aligned if the user’s expectation is to retrieve the data indexed by both classes. This leads to a less restrictive meaning of the equivalence relation.



**Figure 2. Graphical representation of an information retrieval example.** The query is on class C1. By subsumption in WTO, the data indexed by the classes C2, C5, C6, C7 and C8 are retrieved together with the initial class C1 indexed data. Given the equivalence alignment between C5 and C'3, the data indexed by C'3 are also retrieved. The subsumed classes C'5 and C'6 in CO\_321 are then considered. Finally, the equivalence alignment between C'5 and C8 also retrieves the data they index. The final derived class set is C1, C2, C5, C6, C7, C8 and C'3, C'5 and C'6.





**Figure 3.** Schema of the information retrieval process using CO\_321 and WTO alignments.

Until then we have considered the aligned classes as equivalent. It happens that there is no equivalence possible for a class C but only a more specific candidate C'. In this case, our strategy is to create an asymmetric alignment between C and C' where C' is subsumed by C. For instance “WTO: Leaf senescence time” is more general than “CO\_321:Flag leaf senescence time” and WTO does not contain better candidate to align with. Considering such asymmetric alignments extend the retrieval capability of the system while preserving its soundness.

**Alignment mismatches**

Regardless of differences in structure and the inference choices that result, we encountered differences in the traits due to the two ontology purposes and expert disagreement. We present in this section a detailed analysis of these differences and the alignment principles that we have adopted for aligning ontology classes.

**Class naming variations**

Some of the variations among the class names are shallow and are more related to a mismatch at the language level, i.e. using different terms to denote the same concept.<sup>22</sup> An example is the pair <WTO:Moisture content of the grain, CO\_321:Grain Moisture content> where the name differs by a syntactical variation. Such variations are easy to handle in an automatic way but less frequent than deeper semantic differences.

The first type of discrepancy is caused by the differences in the objective of the data annotation by the ontologies, which are the annotation of the *characteristics* of the plant for WTO, and the annotation of the *plant precise observations* for CO\_321. We distinguish four cases related to (E1) whole plant vs plant part description, (E2) measurement methods, (E3) different words for expressing stress effects, and (E4) differences in modality. Examples for each of them are given in Table 3.

**Table 3.** Examples per type of class name variation.

(E1)	<u>Whole plant vs plant part description</u> <i>Fusarium head blight spikelet incidence</i> (CO_321:0000711) qualifies the effect of head blight disease as measurable on spikelets. The phenotypic value is calculated as the ratio of infected spikelets over the total number of spikelets. Conversely, <i>resistance to Fusarium head blight</i> (WTO:0000483) qualifies the resistance level of the cultivar to the disease. Its value ranges from highly resistant to highly susceptible.
(E2)	<u>Measurement method</u> <i>Flour SDS sedimentation</i> (CO_321:0000146) and <i>Flour SDS sedimentation index</i> (CO_321:0000160) both qualify flour protein content but have different calculation methods. The index is the result of dividing the sedimentation volume by the flour protein content. Conversely, WTO defines a single trait regardless of the calculation method, i.e., <i>Sodium dodecyl sulfate-sedimentation volume</i> (WTO:0000610)
(E3)	<u>Different words for expressing stress effects</u> <i>Bird damage</i> (CO_321:0000087) and <i>response to bird damage</i> (WTO:0000674) are two examples of naming variations for denoting the plant response.
(E4)	<u>Difference in modality</u> <i>Aluminium toxicity</i> (WTO:0000450) and <i>Aluminium tolerance</i> (CO_321:0000079)

CO\_321 does not fully follow the EAV model in that the names of the traits contain information about the entity observed (e.g. whole plant, plant part) and the method (e.g. alveograph measure), which are also formally defined. This leads to name variations corresponding to a single trait in WTO that omits the entity observed and method names from trait labels when relevant. This leads to the cases (E1) and (E2).

(E1) is an illustrative example where CO\_321 trait distinguishes the part of the plant where the trait is measured, while WTO traits qualify the properties of the whole plant.

Different traits in CO\_321 may qualify the same property but differ by the method that is used to observe or calculate their values. (E2) is an illustrative example.

CO\_321 trait labels combine the names of biotic or abiotic stresses with eleven different words that express the effect of the stress and depend on the method and the plant part considered: *score*, *notes*, *incidence*, *severity*, *response*, *coefficient of infection*, *index*, *AUDPC*, *damage*, *tolerance*, and *susceptibility*. Conversely, the ability of the plant to resist microbial biotic stresses, the main source of stress, is expressed in WTO trait labels by the single word *resistance*. The words *tolerance* (to micronutrient deficiency and to extreme temperature), *susceptibility* (to lodging), *toxicity* (i.e., *rhyzotoxicity*), and *response* (to macronutrient deficiency, to animal damage, and to general stresses) are used in a complementarily way. Example (E3) illustrates these naming variations.

To deal with cases (E1), (E2) and (E3), in order to achieve our overall interoperability goal, we considered a *multiple alignment* of the single class in one ontology with the distinct classes in the other ontology, so that queries on any of the classes retrieve the same dataset indexed by any of the aligned classes.

It may also happen that the labels induce negative or positive values in opposite directions in the two ontologies, as in (E4), although trait names should be neutral. The decision to consider them aligned despite this requires that the user using the retrieved data be informed.

In these four cases, our mapping strategy in line with our overall interoperability goal is to align the classes if they do not otherwise differ.

### **Agent and disease misalignment**

A large number of the class labels and definitions relate to the names of diseases and pathogenic agents and how plants respond to the biotic stress they cause. The second major source of discrepancy is the disagreement among the experts about the identification of the pathogen agents that cause diseases (cf. examples in [Table 4](#)).

First, the diseases caused by the same pathogen on *different organs* can be different diseases. The classes that relate to resistance to these diseases must therefore not be aligned. (E5) is an illustrative example.

The names of diseases are sometimes different in Europe, where WTO is developed, and the USA, where CO\_321 is mostly developed, or may even differ more locally. In addition, the names of diseases may change over time as scientific knowledge advances.

Two different diseases in CO\_321 may be considered the same in WTO, or conversely. It yields two resistance traits on one side and a single resistance trait on the other side. (E6) is an example of this case. This violates the principle of concept uniqueness.

In a simpler case, each ontology defines a single class for a given disease resistance trait, but they *differ on which species names are synonyms* for the pathogen agents involved. Example (E7) illustrates this case. We decided to consider the two classes aligned despite their taxonomic choice differences because in this case, the two ontologies agree on the causal pathogen agents of the disease, and the disagreement concerns naming. Therefore, the decision is to align the classes. It is worth noting that the identification and naming of pathogens evolve over time; for some, no definitive conclusions can be drawn. The pathogens may have many names that are not all reported in the ontologies.

A more complex case, as illustrated by (E8), is the case where the causal agents of the diseases to which resistance is reported, are different depending on the ontologies. Using an external taxonomic reference may help to build the alignments. Similarly to the disease distinction case above, multiple alignments have to be considered to reconcile the *different agent-disease relation* points of view and achieve our interoperability objective.

**Table 4. Examples per type of misalignment due to disease and causal agent differences.**

(E5)	<u>Different diseases caused by the same pathogen on different organs</u> <i>Black chaff</i> and <i>Leaf streak</i> diseases are both caused by the bacteria <i>Xanthomonas campestris</i> pv. <i>Translucens</i> . <i>Black chaff</i> affects the glume of the plant, and <i>leaf streak</i> affects the leaf. Therefore, the two classes <i>resistance to black chaff</i> (WTO:0000494) and <i>Bacterial leaf streak severity</i> (CO_321:0001019) are not aligned.
(E6)	<u>Multiple diseases viewed as a single disease</u> WTO considers <i>Leaf blotch</i> , <i>Septoria blotch</i> and <i>Septoria tritici blotch</i> the same diseases leading to a single trait. In CO_321 they correspond to two different diseases, i.e. <i>Septoria tritici blotch</i> and <i>Septoria blotch</i> , leading to two distinct traits.
(E7)	<u>Difference in pathogen species definition</u> WTO:0000510 <i>resistance to wheat blast</i> synonyms includes <i>resistance to Magnaporthe grisea</i> , to <i>Magnaporthe oryzae</i> , and to <i>Pyricularia grisea</i> . The first name refers to a species other than the last two, which are synonyms according to the NCBI reference taxonomy. CO_321:0001031 <i>Wheat blast severity</i> is defined as caused by the agent <i>Magnaporthe grisea</i> ( <i>Pyricularia oryzae</i> ) where the two names are presented as synonyms, referring to the same species as opposed to NCBI taxonomy.
(E8)	<u>Different agent-disease relation</u> <i>Resistance to eyespot</i> (WTO:0000482) is related to two pathogen species, <i>Oculimacula yallundae</i> (anamorph <i>Helgardia herpotrichoides</i> (Fon)) and <i>Oculimacula acuformis</i> (anamorph <i>Helgardia acuformis</i> (Nirenberg)). CO_321 defines two different classes, <i>Eyespot plant response</i> (agent <i>Tapesia yallundae</i> = <i>Oculimacula yallundae</i> ) that correspond to <i>Resistance to eyespot</i> , but also <i>Susceptibility to Cercospora</i> , which refers to <i>Cercospora herpotrichoides</i> species. Following the NBI taxonomy and the Encyclopedia of Life, <i>Cercospora herpotrichoides</i> is synonym of <i>Oculimacula yallundae</i> . The two CO_321 classes should then not be distinct. <i>Resistance to eyespot</i> (WTO:0000482) and <i>Susceptibility to Cercospora</i> (CO_321:1000252) are then also aligned through the pathogen name synonymy link, creating a multiple alignment.
(E9)	<u>Difference in taxonomy and missing classes</u> The <i>Nematode resistance</i> (WTO:0000335) class is more general than the <i>Cyst nematode damage</i> (CO_321:0000111) class. Conversely the <i>Resistance to cereal cyst nematode</i> (WTO:0000495) class is more specific than the <i>Cyst nematode damage</i> (CO_321:0000111) class. None of these classes are equivalent. Two subsumption asymmetric alignments are then needed here. They prevent retrieving data on non-cyst nematode when data on cyst nematode data is queried and retrieving data on non-cereal cyst nematode when data on cereal cyst nematode is queried.

Finally, the level of precision in taxonomy may also vary. One ontology may group pathogen subspecies together, while the other will distinguish them into different classes. The (E9) case involves a difference in the taxonomic level where WTO class species A and B are subspecies of CO\_321 class species C. The desired behavior of the information system is that queries on C would retrieve data indexed by A and B, not *vice versa*. Therefore, we define an asymmetric alignment to meet this requirement.

It should be emphasized that the knowledge about the identification of the causal agents of wheat diseases is constantly being revised, which explains these ontology differences.

### Traceability and justification of the alignment decisions

To ensure consistency and to formalize the alignment principles at a fine-grained level, we have defined three strategies that we describe in this section.

A set of rules formalizes the naming variations of the label classes that are acceptable to align them according to the principles above (section Alignment mismatches).

A typology of alignment types defines the nature of the lexical and semantic relationships between the aligned classes.

Additional documentation describes the external source of information when it is needed, e.g., the project expertise, scientific documents, and reference taxonomy.

### Alignment rules

We identified and defined 17 rules for naming variations of the label classes (Table 5). Each rule has a name, a definition in the form of a regular expression. It includes some illustrative examples, an explanation when relevant, and a comment to clarify its meaning or limitation.

**Table 5. List of alignment rules.**

(R1.1) Bio_Score (R1.2) Bio_Incidence (R1.3) Bio_Severity (R1.4) Bio_Plant_response (R1.5) Bio_Coefficient of infection (R1.6) Bio_Seedling response (R1.7) Bio_Incidence_plantpart (R1.8) Bio_Disease index (R1.9) Bio_ISK index (R1.10) Bio_AUDPC	(R2) Resistance_Damage (R3) Abio_Toxicity_Tolerance (R4) Abio_Resist_Toler (R5) Abio_Resist_Suscept (R6) Plant_opt (R7) Toler_Suscept (R8) Agro_Resist_Incidence (R9) Abio_Suscept_Incid
--	---

**Table 6. The Bio\_Plant\_response rule, (R 1.4).**

Ontology	WTO	CO_321
Rule	resistance to + 'name of disease'   'name of the pathogen	'name of disease' + plant response
Example	Resistance to Stripe Rust (WTO:0000562)	Stripe rust plant response (CO_321:0000179)
Explanation	WTO defines general trait classes of response or resistance to pests and diseases. CO_321 defines traits about the observable degree of affection. The observation indexed by CO_321 may be done on the whole plant, or a subpart of it while WTO defines resistance to pests and diseases for the whole plant. We consider that the user of the information retrieval function, given a pathogen or a disease, would like to retrieve all data, regardless of the way the disease is observed. As a consequence, the retrieval terms response and resistance are considered similar.	
Comment	The plant response is measured on the adult plant	

It defines as aligned the classes related to the plant response and plant resistance to a specific pathogen or disease.

For instance, the rule R1.4 Bio\_Plant\_response in Table 5 is about the term used to name the response to a biotic stress and is detailed in Table 6. Notice that the name of the rule was chosen to include the main words involved in the lexical variation. The full list of rules is given in <https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/ZLJYQO>.

### Alignment typology

We identified and defined alignment types to characterize the relation logically (Tables 7 and 8). The equivalence type (T1) is subdivided into five subtypes with regard to the reasons for the alignment of the classes.

**Table 7. Simple alignment types.**

Label	Type of the alignment
T1	Equivalence of the classes.
T1.1	The two traits have the same meaning. Their labels are the same or differ slightly.
T1.2	The two traits have the same meaning. The label of one of the classes is identical or slightly different from a synonym in the other class.
T1.3	The two traits have different labels, but their definitions are in agreement.
T1.4	The labels of the two classes match according to one of the rules of Table 4.
T1.5	The two traits match according to supplementary information from external sources (e.g., scientific publications).
T7	The CO_321 class is more general than the WTO class.
T8	The WTO class is more general than the CO_321 class.

These types represent equivalence and subsumption relationships between class pairs.

**Table 8. Complex alignment types, definitions and examples.**

(T2)	The phenotypic value of the trait can be formally derived or computed from values of other traits	definition
	the <i>tiller number</i> (CO_321:0000190) per area depends on the <i>tillering capacity</i> (WTO:0000640) (=number of tillers per shoot) and on the <i>plants per area</i> (CO_321: 1010059)	example
(T3)	The phenotypic value of the trait depends on the values of other traits, but the relationship cannot be formalized.	definition
	The value of <i>grain manganese content</i> (CO_321:0500033) depends on the <i>manganese use efficiency</i> (WTO:0000238) through a complex physiological relationship	example
(T4)	The phenotypic value of the trait of WTO can be deduced from the value of the trait of CO_321 given thresholds (e.g., discretization).	definition
	The characteristics of the <i>high soil coverage</i> (WTO:0000663) phenotypic value can be derived from a threshold and from the <i>Crop ground cover</i> (CO_321:0000014) trait value which is the percentage of coverage.	example
(T5)	The phenotypic value of the trait of WTO can be deduced from the value of the trait of CO_321 given external condition values, but the relationship cannot be formalized.	definition
	<i>Response to water deficiency</i> (WTO:0000259) and <i>Leaf rolling incidence</i> (CO_321:0001529). Leaf rolling is an indicator of the plant ability to respond to water stress.	example
(T6)	The phenotypic value of the WTO trait can be derived from the value of the CO_321 trait given a reference.	definition
	<i>Short awned</i> (WTO:0000054) and <i>Awn length</i> (CO_321:0000026). Given the measure of the length, the shortness of the awn is determined by comparison with a reference.	example

Five other alignment types have been defined for the documentation of complex alignments (Table 8). Although they have been defined, the alignments of these types are not included in the published alignment file because they cannot be easily represented in SSSOM. Only the simple alignment types in Table 5 are included in the formal mapping set.

### Methodology for alignment

#### Competencies and roles of the participants of the alignment task

The highly technical nature of the phenotyping field calls for the participation of two wheat experts and five knowledge engineers who elicit the knowledge of the domain experts and represent it according to the formal framework. Characterizing the relationship between the two ontology sets of classes involves broad expertise not only in phenotyping measurement, but also in plant biology, physiology, pathology, agronomy, and food processing. The role of the knowledge engineers was divided between formulating alignment proposals, reviewing them with the help of experts when necessary, revising them, and coordinating the overall task.

#### Collaborative process and tools

Naturally, we have examined the candidate classes for mapping by subfields in order to concentrate the interactions into limited areas taking advantage of the highly structured nature of WTO. The classes of each subfield with their properties, labels, definitions, and synonyms were listed in a depth-first order in separate tables. Tables were used because no suitable user-friendly interface could be found to edit and visualize the alignments in a collaborative and flexible way. Once all WTO classes had been handled, we examined the remaining CO\_321 classes that were still not aligned. The flat structure of CO\_321 means that the knowledge engineer was constantly moving from one topic to another.

The successive versions of the alignment tables were kept in the INRAE GitLab (<https://forgemia.inra.fr/urgi-is/ontologies/-/tree/feat/mapping-wheat/mapping>). The two main knowledge engineers used GitLab issues to discuss with the rest of the team alignment proposals that fell outside their expertise and the scope of the name variation rules.

The issues that required the in-depth expertise of the two specialists have been dealt with in a separate file, which was more convenient for them than GitLab issues. It gathers 72 numbered questions in 32 pages broken down into sub-fields which were answered by their respective experts. Each question was answered with a paragraph of varying length depending on the complexity of the issue. The answer is followed by the decision and closed with a 'done' mark.

### Validation and revision procedure

Once the discussion was closed, the main arguments and decisions were summarized in the comment section of the alignment tables. The mapping type, the reasons for the alignment in free text form, and the external references, when relevant, were also recorded in the alignment tables. All alignments were reviewed by one or two knowledge engineers with backgrounds in wheat breeding and knowledge representation. The review of the alignment has generated new sequences of discussions and revisions in an iterative way.

### Results

#### Alignment figures

The traits of the whole CO\_321 and the traits of WTO subtree have been examined and aligned when appropriate. **Table 9** gives the total number of classes and the number of classes for which an alignment has been found; only 140 WTO and 109 CO\_321 classes could not be aligned, i.e., 23% of the total.

**Figure 4** shows the distribution of the mapping types from 0 to 8 (section Alignment typology). As expected, the most frequent types are 0 (no alignment found) and 1 (equivalence). The 7 and 8 alignment types represent the asymmetric relations where the pair members are not equivalent but one is more general than the other. As expected, type 7 (the WTO class is more specific) is less frequent than 8 (the WTO class is more general) since the depth of CO\_321 is low with only 3 levels.

Only alignments of type 1, 7 and 8 can be represented in SSSOM.

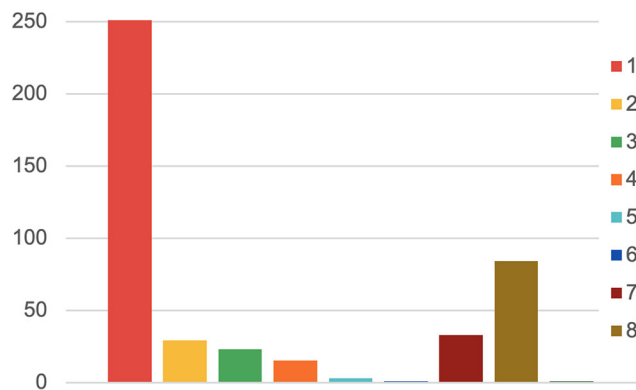
- The subcategories of type 1 (equivalence) have been represented by skos:exactMatch (types 1.1, 1.2 and 1.3), skos:closeMatch (type 1.4) and skos:relatedMatch (type 1.5).
- Types 7 and 8 are represented respectively by skos:narrowMatch and skos:broadMatch.

**Table 10** gives the number of alignments represented in SSSOM per type of skos match.

As detailed in section Class naming variations, the lexical and semantic variations among the two class sets led to extending the bijective frame to multiple mappings where a class in one ontology may have more than one image class in the other ontology. **Figures 5a** and **5b** show the distribution of multiple mappings of the two ontologies. Multiple

**Table 9. Number of trait classes aligned classes per ontology.**

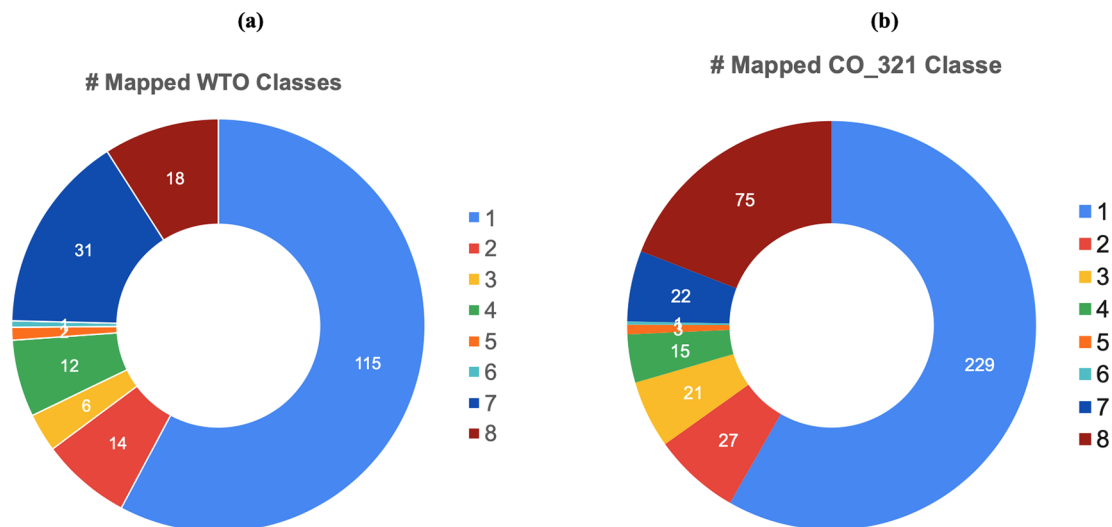
# WTO trait classes	596
# WTO aligned classes	456
# CO_321 classes	467
# CO_321 aligned classes	358



**Figure 4. Mapping distribution by type.** 1: equivalence; 2 to 6: complex; 7: broader WTO class; 8: broader CO\_321 class.

**Table 10. Number of published alignments per type of skos match.**

Type of alignment	Number
skos:exactMatch	76
skos:narrowMatch	81
skos:broadMatch	32
skos:closeMatch	161
skos:relatedMatch	0
TOTAL alignments	350



**Figure 5. Distribution of multiple mappings per type for (a) WTO classes and (b) CO\_321 classes.** The types are numbered from 1 to 8 according to Tables 5 and 6.

mappings have different causes that are related to differences in class granularity and disagreements on pathogen agents causing diseases (section Agent and disease misalignment).

The high number of multiple mappings (up to 8 WTO classes mappings to CO\_321 classes) are due to three causes: (E1) whole plant vs plant part description, (E2) measurement methods, and (E3) different words for expressing stress effect measure (e.g. note, score, incidence). Multiple classes detailing the different plant parts affected, the different measurement methods used slightly varying by their labels were considered as equivalent in order to ensure the retrieval of all data related to a given trait.

Table 11 gives an example of a multiple mapping of type T1 (equivalence) (Table 5) of the class *resistance to Fusarium head blight* (WTO:0000483). It is aligned to 7 classes from CO\_321 that differ by the words expressing biotic stress effect measures.

These alignments are all justified by the application of one of the rules. Therefore, the alignment type of Table 9 example is T1.4, i.e. *The labels of the two classes match according to one of the rules*, according to the alignment typology.

### Files and documentation

The result of this alignment task is a mapping set published as a tabular file (TSV format) that conforms to the SSSOM specifications<sup>1</sup> and includes the information listed in Tables 12 and 13. In the TSV format, the mapping set metadata is described in the header of the file (Table 12). In the body of the file, each line contains a mapping (subject - predicate - object) with its own metadata (Table 13).

**Table 11. Multiple mappings of the class *resistance to Fusarium head blight* (WTO:0000483).**

Type	Rule name	CO_321 class id	CO_321 class label
T1.4	Bio_Incidence (1.2)	CO_321:0000924	Fusarium graminearum incidence
T1.4	Bio_Severity (1.3)	CO_321:0000926	Fusarium graminearum severity
T1.4	Bio_Plant_response (1.4)	CO_321:0000925	Fusarium graminearum plant response
T1.4	Bio_Plant_response	CO_321:0000929	Fusarium head blight spike response
T1.4	Bio_AUDPC (1.10)	CO_321:0000651	Fusarium head blight AUDPC
T1.4	Bio_ISK index (1.9)	CO_321:0500021	Fusarium head blight ISK index
T1.4	Bio_Disease index (1.8)	CO_321:0500019	Fusarium head blight disease index

**Table 12. SSSOM metadata elements used to describe the mapping set.**

Property	Description	Example
creator_id	Identifies the persons or groups responsible for the creation of the mapping.	"https://ror.org/02kvxyf05"
creator_label	A string identifying the creator of this mapping.	"INRAE"
curie_map	A valid curie map that allows the unambiguous interpretation of CURIEs	#skos: "http://www.w3.org/2004/02/skos/core"
license	A url to the license of the mapping. In absence of a license, we assumed no license.	"https://www.etalab.gouv.fr/licence-ouverte-open-licence/"
mapping_set_id	A globally unique identifier for the mapping set (not each individual mapping). Should be IRI, ideally resolvable.	"https://doi.org/10.57745/ZLJYQO"
object_source	URI of vocabulary or identifier source for the object.	"https://cropontology.org/term/CO_321"
object_source_version	Version IRI or version string of the source of the object term.	"July 2018"
subject_source	URI of ontology source for the subject.	"http://opendata.inrae.fr/wto"
subject_source_version	Version IRI or version string of the source of the subject term.	"3.0"

**Table 13. SSSOM metadata elements used to describe a mapping.**

Property	Description	Example
author_id	Identifies the persons or groups responsible for asserting the mappings. Recommended to be a list of ORCID IDs or otherwise identifying URIs.	"https://orcid.org/0000-0002-1110-8004"
author_label	A string identifying the author of this mapping. In the spirit of provenance, consider using author_id instead.	"Claire Nédellec"
comment	Free text field containing either curator notes or text generated by a tool providing additional informative information.	"The definitions of both traits correspond to the date of ear emergence."
curation_rule	A (potentially) complex condition executed by an agent that led to the establishment of a mapping.	"https://doi.org/10.57745/MRGHPA" - Rule 4
mapping_cardinality	A string indicating whether this mapping is from a 1:1 (the subject_id maps to a single object_id), 1:n (the subject maps to more than one object_id), n:1, 1:0, 0:1 or n:n group. Note that this is a convenience field that should be derivable from the mapping set.	"1:1"



**Table 13.** *Continued*

Property	Description	Example
<b>mapping_date</b>	The date the mapping was asserted. This is different from the date the mapping was published or compiled in a SSSOM file.	2022
<b>mapping_justification</b>	A mapping justification is an action (or the written representation of that action) of showing a mapping to be right or reasonable.	semapv: ManualMappingCuration
<b>object_id</b>	The ID of the object of the mapping.	CO_321:0000982
<b>object_label</b>	The label of object of the mapping.	Anther extrusion
<b>object_type</b>	The type of entity that is being mapped.	owl class
<b>predicate_id</b>	The ID of the predicate or relation that relates the subject and object of this match.	skos:exactMatch
<b>reviewer_id</b>	Identifies the persons or groups that reviewed and confirmed the mapping. Recommended to be a list of ORCIDs or otherwise identifying URIs.	" <a href="https://orcid.org/0000-0001-9356-4072">https://orcid.org/0000-0001-9356-4072</a> "
<b>reviewer_label</b>	A string identifying the reviewer of this mapping. In the spirit of provenance, consider using reviewer_id instead.	"Michael Alaux"
<b>subject_id</b>	The ID of the subject of the mapping.	WTO:0000065
<b>subject_label</b>	The label of subject of the mapping.	anther extrusion
<b>subject_type</b>	The type of entity that is being mapped.	skos concept

All the metadata elements are explained in the SSSOM documentation available online [1]. We associated the mapping set with a PDF file containing the mapping rules, i.e. the domain-related explanations for creating the mappings. References to these rules are included in the mapping file for 123 mappings, i.e. the mapping of type T1.4 (*the labels of the two classes match according to one of the rules*). When producing mappings manually, it is crucial to document the work done by the experts and be transparent on the reasons for creating the mappings. The curation rule and mapping justification metadata contribute to the trustworthiness and reusability of the mapping set.

## Conclusion

The objective of our research is to consolidate experimental and scientific data pertaining to wheat breeding sourced from phenotyping experiments and scholarly articles. These datasets are indexed using two distinct ontologies CO\_321 and WTO which both delineate plant traits through hierarchical class structures. However, discrepancies in ontology modeling have arisen due to conceptual differences. To leverage the wealth and complementarity of both data reservoirs, we recognize the mapping of classes as a crucial step toward ensuring data interoperability.

We qualified each alignment with the appropriate semantic relation depending on the degree of equivalence and we provided mapping justification for an effective intended use in the information retrieval task. These alignments, curated and validated through an iterative procedure involving wheat experts, are a valuable resource for researchers in wheat breeding. Furthermore, the decision to make these alignments available in SSSOM format<sup>21</sup> greatly enhances their interoperability and usability within the broader context of knowledge graphs and ontological mappings. The conclusions drawn from this specific alignment task appear to be valuable within the broader context of integrating scientific and experimental data.

## Discussion and future work

### Perspectives for the WheatIS and FAIDARE data discovery exploitation

The FAIRification of phenotyping data through WheatIS and FAIDARE data discovery portals is a major challenge as there is no generic repository for trait data.<sup>4,23</sup> WheatIS and FAIDARE are widely used by the international wheat research community and the European plant research community to find and exploit these data.

The initial application of this research will involve enhancing data discovery within WheatIS and FAIDARE through formal alignments of the WTO and CO\_321 ontologies. Prior efforts within the European OpenMinTeD project

<sup>1</sup><https://mapping-commons.github.io/sssom/spec/#sssom-metadata-elements>

The screenshot shows the WheatIS search results for 'Stripe Rust'. The search bar contains 'Stripe Rust' and a 'Search' button. The results are displayed in a list format, with the first result highlighted. The ontology annotation filter is open, showing a list of terms related to 'Stripe Rust' and 'Wheat'. The filter is set to 'Filter on Ontology annotation...'. The list of terms includes:

- stripe rust (WTO:0000518) [162]
- leaf rust (WTO:0000466) [123]
- resistance to Stripe Rust (WTO:0000562) [93]
- stem rust (WTO:0000517) [88]
- Puccinia striiformis (WTO:0000421) [87]
- resistance to Leaf Rust (WTO:0000549) [80]
- Puccinia triticina (WTO:0000422) [75]
- resistance to Stem Rust (WTO:0000561) [68]

Other results are available. Refine your search.

The search results list includes:

- Results 1 to 20 of 387
- 10.1094/PHYTO.2003.93.7.881 - OpenMinTeD@GnplS
- Bibliography **Triticum**
- Microsatellite markers for genes Lr34/Yr18 and other quantitative trait loci for leaf rust and stripe rust resistance in bread wheat. 2003
- Microsatellite markers for genes Lr34/Yr18 and other quantitative trait loci for leaf rust and stripe rust resistance ... (expand)
- 10.1371/journal.pone.0222755 - OpenMinTeD@GnplS
- Bibliography **Triticum** **Triticum aestivum**
- Evaluation of a global spring wheat panel for stripe rust: Resistance loci validation and novel resources identification. 2019 Evaluation of a global spring wheat panel for stripe rust: Resistance loci validation and novel resources identification **Stripe r** ... (expand)
- 10.1162/PAKJAS/16.2503 - OpenMinTeD@GnplS
- Bibliography **Triticum**
- MOLECULAR GENETIC VARIATION FOR STRIPE RUST RESISTANCE IN SPRING WHEAT. 2016 MOLECULAR GENETIC VARIATION FOR STRIPE RUST RESISTANCE IN SPRING WHEAT **Stripe rust**, caused by Puccinia striiformis f. sp. tritici, is a major biotic constraint to global wheat production ... (expand)
- 10.1094/PDIS-07-13-0707-RE - OpenMinTeD@GnplS
- Bibliography **Triticum**
- Genetic Analysis of Resistance to Leaf Rust and Stripe Rust in Wheat Cultivar Francolin#1. 2014 Genetic Analysis of Resistance to Leaf Rust and Stripe Rust in Wheat Cultivar Francolin#1. Leaf rust and stripe rust are important diseases of wheat and can be controlled by ... (expand)
- 10.1007/s11032-012-9798-4 - OpenMinTeD@GnplS
- Bibliography **Triticum** **Triticum turgidum**
- Identification and mapping of leaf, stem and stripe rust resistance quantitative trait loci and their interactions in durum wheat. 2013
- Identification and mapping of leaf, stem and stripe rust resistance quantitative trait loci and their interactions in durum wheat ... (expand)

**Figure 6. WheatIS screenshot of the ontology annotation filter used in a search for 'Stripe Rust'.**

established an initial set of alignments between WTO and WIPO (*Wheat INRA Phenotyping Ontology*), subsequently integrated with CO\_321. This process involves a preprocessing saturation step, expanding the “annotation id” field (for CO\_321 and WTO) and “observation variable id” (for CO\_321), and incorporating all equivalent and subsumed class labels and synonyms. This augmentation simplifies the information retrieval process to a filtering task (see Figure 6).

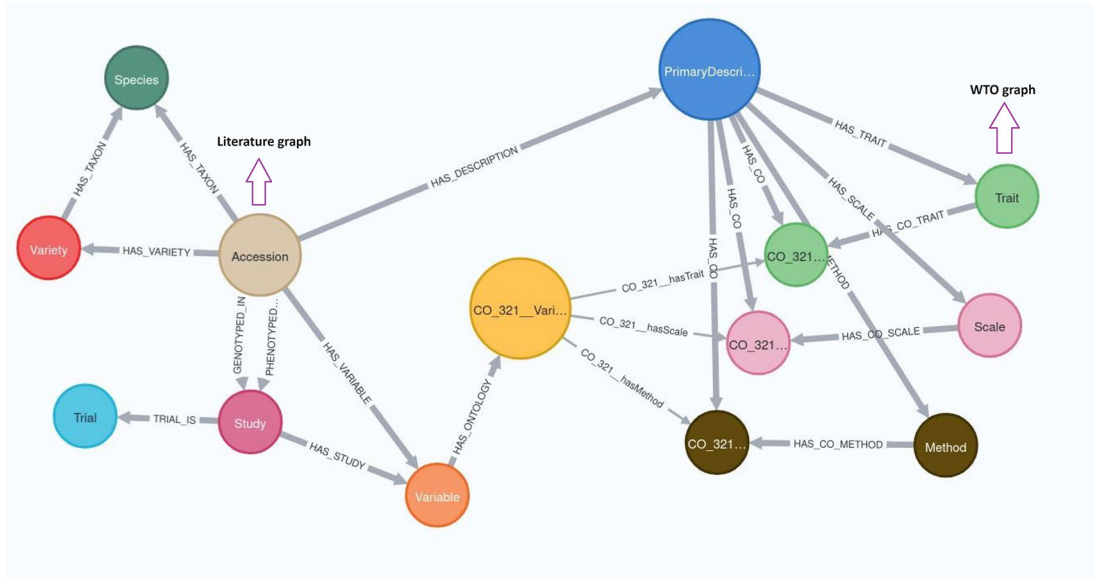
We will enhance and refine the existing scripts to leverage ontology mapping, facilitating the retrieval of results indexed with classes mapped from one ontology to another (CO\_321 to WTO or WTO to CO\_321). This enhancement will enable the correlation of experimental data with literature data and *vice versa*. The implementation of this development is anticipated to be swift and will significantly augment the functionality of the tool.

In the medium term, we envision establishing a graphical representation of the data sources based on annotation classes and mapping. This visualization will empower users to intuitively explore the mappings and navigate through them with ease.

Furthermore, these mapped ontologies and corresponding data will be seamlessly integrated into the wheat knowledge graph, under development at INRAE-URGI, leveraging Neo4J (*Network Exploration and Optimization 4 Java*) technology (<https://neo4j.com/>). This knowledge graph encompasses accessions and related phenotypic experimental data annotated with CO\_321 (Figure 7). The integration of literature data annotated with WTO in this graph will create a comprehensive knowledge system for exploring the biological mechanisms underlying phenotypic expressions.

### Update and life cycle

The information retrieval system within the data discovery portals is intended for wheat breeders, experimenters and researchers. A test phase is planned to identify any errors or deviations from the intended use. Beta versions of the WheatIS and FAIDARE data discovery portals will be set up to test queries, the consistency of the results and the clarity of the explanations of retrieved data. These explanations will enable users to scrutinize the data retrieval mechanisms and suggest revisions. In addition, future updates will be triggered by revisions to the WTO and CO\_321 ontologies, including the addition of new classes, as well as class merging or deletion.



**Figure 7.** Neo4J graph schema including CO\_321 variables.

**Representation choice**

We chose SSSOM as the standard for representation due to its relevance in depicting alignments, their sources, and justifications. In selecting SSSOM we sought to identify a methodology that was relevant in this particular context. Following this representation work, we identify a number of avenues for further reflection regarding the limitations of our use of SSSOM. SSSOM offers functionalities that we have not yet used, which could prove useful in the future. The level of confidence of the experts is one of them. In our work, the experts were confident in their alignments. Uncertainties, such as those concerning disease-causing pathogens, are shared by the scientific community as a whole.

The main issue we encounter in the use of SSSOM concerns multiple mappings that associate a class from one ontology with several disjoint classes from the other ontology, as exemplified by the example of Table 11. The SSSOM metadata mapping\_cardinality provides insight into the alignment quality by indicating the cardinality of relations, thus potentially detecting inconsistencies in SKOS mappings. The mapping\_cardinality metadata takes values of 1:n in our dataset. Specifically, in the 24 cases of multiple mappings, some are of the exactMatch type. According to SKOS specifications, skos:exactMatch should have a cardinality of “1:1”. Any deviation from this rule may indicate either erroneous mappings or duplicate classes or concepts in one or both of the aligned semantic resources. We assume that the mappings of our dataset are correct. Moreover, the designers of both ontologies consider that the classes in their ontologies are not duplicated: neither the relations in WTO nor those in CO\_321 include equivalence relations between classes, and they are not linked by skos:exactMatch relations.

It is the mapping process that creates pseudo-duplication of classes from the perspective of the source ontology onto the target ontology. Therefore, our current usage contravenes SKOS specifications without us having identified a satisfactory solution. Our SSSOM representation fulfills our need for formalization and sharing of our results, but its utilization will need to consider this peculiarity of multiple correspondences.

We believe that the issue raised by aligning sets of classes with different degrees of precision is a general representation problem that deserves further investigation, particularly within the context of indexing scientific and experimental data by a shared conceptualization.

**Ethical considerations**

Not applicable.

**Ethics and consent**

Not applicable.

### **Author contributions**

Conceptualization: Claire Nédellec, Sophie Aubin, Michael Alaux, Cyril Pommier, Liliana Ibanescu,

Data curation: Clara Sauvion, Claire Nédellec, Michael Alaux, Cyril Pommier, Jacques Le Gouis, Thierry C. Marcel

Formal Analysis: Robert Bossy

Funding acquisition: Claire Nédellec, Sophie Aubin, Cyril Pommier, Michael Alaux

Investigation: Claire Nédellec, Michael Alaux

Methodology: Claire Nédellec, Sophie Aubin, Michael Alaux, Liliana Ibanescu

Project administration: Claire Nédellec

Resources: Robert Bossy

Software: Robert Bossy, Sonia Bravo, Sophie Aubin

Supervision: Claire Nédellec, Michael Alaux, Sophie Aubin

Validation: Cyril Pommier

Writing – original draft: Claire Nédellec, Sophie Aubin, Cyril Pommier, Michael Alaux, Liliana Ibanescu, Sonia Bravo

Writing – review & editing: Jacques Le Gouis, Thierry C. Marcel, Robert Bossy, Clara Sauvion

### **Data availability**

Recherche Data Gouv: Alignment of WTO and CO\_321 ontology classes [Dataset]

The dataset file is provided in the SSSOM TSV format together with the full rule list file at the French research open data portal:

The project contains the following underlying data:

1. S1.Alignment rules.pdf
2. WTO-CO321\_mappings\_V1.0\_02\_2024.tsv

<https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/ZLJYQO>.<sup>24</sup>

The dataset “Alignment of WTO and CO\_321 ontology classes” is available under the terms of the [etalab 2.0](#) license. This license has been designed to be compatible with any free license that at least requires an acknowledgement of authorship, and specifically with Creative Commons’ “Creative Commons Attribution” (CC-BY). The repository is the French national public repository for research data.

### **Acknowledgements**

The authors want to thank Mariya Evtimova, Thomas Letellier, Maud Marty, for their contribution on the class mapping; Catherine Faron, Raphael Flores, Nicolas Francillonne, Yosra Maestri, Franck Michel, Nadia Yacoubi for their contribution in the knowledge graphs developments (RDF and Neo4J); Clément Jonquet for its contribution to AgroPortal development and D2KAB project coordination.

The authors acknowledge the following projects, networks and communities.

OpenMinTeD project (H2020-EINFRA-2014-2) under grant agreement no. 654021 and the Crop Ontology project for their contribution in the development of WTO and CO\_321 ontologies.

FSOV 2010H SAM Blé for its contribution in the development of WTO and the text mining workflow.

Investment for the Future PIA BreedWheat project funded by the French Research National Agency (ANR-10-BTBR-03), FranceAgriMer (2013-0544), and the French fund to support breeding research (FSOV-2012D); and the Whealbi project under grant agreement FP7-613556 for providing phenotypic experimental data.

Wheat Information System expert working group of the Wheat Initiative and the ELIXIR plant science community for the development of the WheatIS and FAIDARE data discovery portals.

Saclay Plant Sciences-SPS (ANR-17-EUR-0007) for supporting INRAE-URGI.

## References

- Bonjean A: **The saga of wheat—the successful story of wheat and human interaction.** Bonjean A, *et al.*, editors. *The world wheat book: a history of wheat breeding*. Vol 3. Paris: Lavoisier; 2016; pp. 1–90.
- Oury F, Godin C, Mailliard A, *et al.*: **A study of genetic progress due to selection reveals a negative effect of climate change on bread wheat yield in France.** *Eur. J. Agron.* 2012; **40**: 28–38.  
[Publisher Full Text](#)
- Tardieu F, Cabrera-Bosquet L, Pridmore T, *et al.*: **Plant phenomics, from sensors to knowledge.** *Curr. Biol.* 2017; **27**(15): R770–R783.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Pommier C, Michotey C, Cornut G, *et al.*: **Applying FAIR Principles to Plant Phenotypic Data Management in GnpIS.** *Plant Phenomics.* 2019 Apr 30; **2019**: 1671403.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ghaffary SM, Robert O, Laurent V, *et al.*: **Genetic analysis of resistance to septoria tritici blotch in the French winter wheat cultivars Balance and Apache.** *Theor. Appl. Genet.* 2011 Sep; **123**(5): 741–754.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Alaux M, Dyer S, Sen TZ: **Wheat Data Integration and FAIRification: IWGSC, GrainGenes, Ensembl and Other Data Repositories.** Appels R, Eversole K, Feuillet C, *et al.*, editors. *The Wheat Genome. Compendium of Plant Genomes*. Cham: Springer; 2024.  
[Publisher Full Text](#)
- Sen TZ, Caccamo M, Edwards D, *et al.*: **Building a successful international research community through data sharing: the case of the Wheat Information System (WheatIS).** *F1000Res.* 2020; **9**: 536.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shrestha R, Matteis L, Skofic M, *et al.*: **Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice.** *Front. Physiol.* 2012; **3**: 326.  
[Free Full Text](#)
- Nédellec C, Ibanescu L, Bossy R, *et al.*: **WTO, an ontology for wheat traits and phenotypes in scientific publications.** *Genomics Inform.* 2020; **18**: e14.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Papoutsoglou EA, Faria D, Arend D, *et al.*: **Enabling reusability of plant phenomic datasets with MIAPE 1.1.** *New Phytol.* 2020; **227**(1): 260–273.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Selby P, Abbeloos R, Backlund JE, *et al.*: **BrAPI—an application programming interface for plant breeding applications.** *Bioinformatics.* 2019; **35**(20): 4147–4155.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nédellec C, Bossy R, Valsamou D, *et al.*: **Information extraction from bibliography for marker-assisted selection in wheat.** *Metadata and Semantics Research: 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings*. Vol. 8. Springer International Publishing; 2014; 8: pp. 301–313.  
[Publisher Full Text](#)
- Euzenat J: **An API for ontology alignment.** *Proc. 3rd international semantic web conference (ISWC), Nov 2004, Hiroshima, Japan.* pp.698–712. [ff10.1007/978-3-540-30475-3\\_48ff](#). [ffhal-00825931](#).
- Jonquet C, Toulet A, Arnaud E, *et al.*: **AgroPortal: A vocabulary and ontology repository for agronomy.** *Comput. Electron. Agric.* 2018; **144**: 126–143.  
[Publisher Full Text](#)
- Harrow I, Balakrishnan R, Jimenez-Ruiz E, *et al.*: **Ontology mapping for semantically enabled applications.** *Drug Discov. Today.* 2019; **24**(10): 2068–2075.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Liu X, Tong Q, Liu X, *et al.*: **Ontology Matching: State of the Art, Future Challenges, and Thinking Based on Utilized Information.** *IEEE Access.* 2021; **9**: 91235–91243.  
[Publisher Full Text](#)
- Shvaiko P, Euzenat J, Jiménez-Ruiz E, *et al.*: **Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023).**
- Zhou L, Thiéblin E, Cheatham M, *et al.*: **Towards evaluating complex ontology alignments.** *Knowl. Eng. Rev.* 2020; **35**: e21.  
[Publisher Full Text](#)
- Thiéblin É, *et al.*: **The First Version of the OAEI Complex Alignment Benchmark.** *International Workshop on the Semantic Web.* 2018.
- Le Franc Y, Parland-von Essen J, Bonino L, *et al.*: **D2.2 FAIR Semantics: First recommendations (1.0 DRAFT).** *FAIRsFAIR.* 2020.  
[Publisher Full Text](#)
- Matentzoglou N, Balhoff JP, Bello SM, *et al.*: **A Simple Standard for Sharing Ontological Mappings (SSSOM).** *Database (Oxford).* 2022 May 25; **2022**: baac035.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Noy NF: **Ontology mapping.** *Handbook on ontologies.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2009; pp. 573–590.
- Alaux M, Rogers J, Letellier T, *et al.*: **Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data.** *Genome Biol.* 2018 Aug 17; **19**(1): 111.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Claire N, Mariya E, Sophie A, *et al.*: **Alignment of WTO and CO\_321 ontology classes. [Dataset].** *Recherche Data Govv.* 2023; **V4**.  
[Publisher Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**