L'ontologie E-Phy, une base de connaissances pour le catalogue des produits phytopharmaceutiques autorisés en agriculture en France

Syphax Bouazzouni,[1,2] Clement Jonquet[1,3]

[1] LIRMM, Univ. de Montpellier, CNRS, France
[2] Ecole Nationale Supérieure d'Informatique, Alger, Algérie
[3] MISTEA, Univ. de Montpellier, INRAE, Institut Agro, France

gs_bouazzouni@esi.dz, jonquet@lirmm.fr

Résumé

Le développement de ressources sémantiques (ontologies, vocabulaires, graphes de connaissances, etc.) est une activité clé pour faciliter l'intégration et l'interopérabilité des données en agriculture. Bien souvent, les catalogues ou les référentiels officiels ne sont pas du tout "FAIR," et n'existent pas dans un format RDF, comme dans notre cas, en agriculture, le catalogue E-Phy, produit par l'ANSES, qui contient l'ensemble des produits phytopharmaceutiques et de leurs usages, des matières fertilisantes et des supports de culture autorisés en France. *Dans ce travail, nous détaillons notre démarche pour formaliser le catalogue E-Phy sous forme d'une base de connaissances OWL constituée d'un modèle ontologique, de ses instances et d'alignements vers d'autres ontologies. Nous montrons les points difficiles rencontrés dans ce processus, et les limites de la modélisation actuelle restée rétro-compatible avec la base de données d'origine. Nous illustrons également la valeur ajoutée de l'ontologie E-Phy via des requêtes SPARQL qui valorisent la sémantique et les alignements et permettent des interrogations impossibles sur les données d'origine.*

Mots-clés

Ontologie, développement de ressource sémantique, RDFisation, agriculture, ANSES, E-Phy, produits-phytosanitaires ou phytopharmaceutiques

Abstract

The development of semantic resources (ontologies, vocabularies, knowledge graphs, etc.) is a key activity to facilitate data integration and interoperability in agriculture. Often, catalogs or official reference lists do not respect the FAIR principles, and do not exist in an RDF format, as in our case, in agriculture, the E-Phy catalog, produced by ANSES, which contains all the plant protection products (phytosanitary) and their uses, fertilizers and growing media authorized in France. *In this work, we detail our approach to formalize the E-Phy catalog in the form of an OWL knowledge base consisting of an ontological model, instances and alignments to other ontologies. We show the various issues encountered in this process, and the limitations of the current model, which is still backward compatible with the original database. We also highlight, with a few SPARQL queries, the added value of the E-Phy ontology's semantics and alignments with queries impossible on the original data.*

Keywords

Ontology, semantic resource development, RDFisation, agriculture, ANSES, E-Phy, phytosanitary or phytopharmaceutical products

1 Introduction

De nombreux référentiels ou catalogues officiels existent dans le domaine de l'agriculture, en France ou dans le reste du monde. Ces référentiels sont produits par des organismes différents le plus souvent accrédités pour les maintenir, et avec des formats, des processus de production, et de maintenance variés. Par exemples, le *Catalogue officiel des espèces et variétés de plantes cultivées en France* (GEVES), le *catalogue officiel des variétés de vigne* (FranceAgriMer). Dans ce travail nous nous intéressons au *catalogue des produits phytopharmaceutiques et de leurs usages, des Matières Fertilisantes et des Supports de Culture autorisés en France* (ANSES)[1] et à son sous ensemble le *guide des produits de protection des cultures utilisables en agriculture biologique en France* (ITAB)[2] qui décrivent des intrants dont les usages sont eux même décrits dans le *catalogue des usages phytopharmaceutiques* (Ministère de l'Agriculture). Ces référentiels sont indispensables en agriculture mais ils sont pratiquement inutilisables lorsqu'il s'agit de les utiliser pour décrire ou structurer des données : localisation variées, formats hétérogènes et pas exploitables par une machine, pas d'identifiant, maintenance, etc. Ils contiennent une connaissance très riche qui pourrait être mieux représentée pour ensuite être mieux exploitée dans le cadre d'applications ou de recherches en agriculture.

Dans le cas du catalogue des produits phytopharmaceutiques, une première étape a été franchie avec la mise à disposition, depuis 2016, des référentiels produits par l'ANSES[3] sur la plateforme nationale d'ouverture et de partage des données publiques data.gouv.fr. On y trouve le catalogue sous forme d'export de données de l'application web E-Phy créé par l'ANSES pour accéder/rechercher le catalogue en ligne. Ce

[1] www.anses.fr/fr/content/registre-des-amm-de-produits-phyto-et-mfsc
[2] www.itab.asso.fr/downloads/com-intrants/guide-protection-plantes6.pdf
[3] Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

catalogue est proposé sur la plateforme sous deux formats CSV ou XML qui sont des formats libres largement utilisés sur le web et qui ont l'avantage d'être très simples à utiliser, qui structurent un minimum les données, mais qui souffrent aussi terriblement d'un manque de sémantique (pas d'identifiant, pas de hiérarchie ou de typage). L'absence de contraintes d'intégrité engendre également de nombreuses erreurs (de saisie, de valeurs, ou de valeur nulle).

C'est avec l'objectif d'avoir une meilleure structuration des données et pour permettre une meilleure réutilisation et interopérabilité que la prochaine étape serait maintenant de construire une base de connaissances de la base de données E-Phy en l'encodant avec les technologies du web sémantique (e.g., SKOS, OWL) et en l'enrichissant d'alignements avec d'autres ontologies ou vocabulaires standards en agronomie. Ainsi, il s'agit de faire passer les données d'E-Phy du niveau 3 étoiles du fameux modèle des données ouvertes et liées de Berners-Lee [1], au niveau 5 étoiles où les données sont identifiées par des URI et publiées sur le web dans un format ouvert, non-propriétaire, riche et standard tel que RDF.

La démarche de création de ressource sémantique RDF a été largement décrite dans la littérature que ça soit dans le domaine de l'agronomie –par exemple, la définition d'une ontologie pour les stades phénologiques des plantes cultivées [2] –ou dans d'autres domaines comme le médical –par exemple MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein extrait de données de forum [3]. Parfois, des méthodes spécifiques sont utilisées, comme la méthode Linked Open Terms [4] ou DataLift[1]. Mais dans le cas des démarches de "RDFisation" de données préexistantes, le plus compliqué est l'instanciation des éléments des ressources d'origine et leur alignement à d'autres ressources. Pour cela divers outils et projets existent pour faciliter la transformation des données tabulaires ou XML en données sémantiquement structurées et interconnectées. Comme exemples nous pouvons citer Any23[4], Triplify / Sparqlify [5], Open Refine[5] ou Cellfie.[6] Ces méthodes et outils permettent de faire le mapping entre les éléments des fichiers d'origine vers un modèle défini par une étude de l'existant, cependant leur utilisation n'évite pas un prétraitement pour s'adapter à la structure de la ressource et préparer le dit mapping.

Nous nous sommes alors orientés vers une méthodologie en trois étapes: étude de l'existant en créant un diagramme de classe UML de la ressource d'origine, création de la structure de l'ontologie avec Protégé [6] et instanciation/alignement avec une phase de prétraitement avec Talend,[7] puis le plug-in Cellfie de Protégé.

Dans cette article, nous présentons la démarche que nous avons suivi pour transformer le catalogue des produits phytopharmaceutiques de l'ANSES en une base de connaissances OWL dont le modèle ontologique est celui de la base de données E-Phy et les instances sont les produits qui y sont listés, alignées avec d'autres ressources sémantiques tel que le thésaurus French Crop Usage pour les cultures et l'ontologie CHEBI pour les familles chimiques. Cependant considérant que nos organismes ne sont pas des autorités pour ce catalogue, notre modèle de données est volontairement "bridé"[8] pour être complètement rétro-compatible avec la base d'origine de façon à facilement mettre à jour notre ontologie à partir de nouveaux exports de l'ANSES.

Nous illustrons, à travers des requêtes SPARQL, comment l'ontologie E-Phy permet de répondre à des requêtes impossibles sur les données d'origines car elles valorisent la sémantique des ontologies alignées. Par exemple, avec l'utilisation de la hiérarchie des cultures du thésaurus French Crop Usage pour obtenir tous les produits utilisables pour une famille de culture donnée. Nous pouvons également obtenir avec une requête SPARQL une vue (i.e., un sous-ensemble de triplets) de l'ontologie E-Phy qui représente le *catalogue des usages phytopharmaceutiques* produit par l'ITAB à partir des données de l'ANSES.

L'article est organisé comme suit : Section 2, nous commençons par une présentation détaillée du catalogue avec une description des fichiers sources, de ses contenus et de sa structure. Section 3, nous présentons la méthodologie suivie. Section 4, nous présentons nos résultats avec le détail technique, l'ontologie produite et les difficultés rencontrées et les requêtes SPARQL. Finalement, la section 5 conclut et donne quelques perspectives.

2 Présentation du Catalogue E-Phy

Le catalogue E-Phy contient l'ensemble des données des produits (produits phytopharmaceutiques, matières fertilisantes et supports de culture, adjuvants, produits mixtes et mélanges) couverts par une Autorisation de Mise sur le Marché (AMM) ou un permis de commerce parallèle. En France, les décisions d'AMM et de permis sont délivrées par l'ANSES depuis juillet 2015. L'agence partage ce catalogue via l'application web E-Phy (www.ephy.anses.fr) qui est un reflet de l'état actuel des autorisations de produits et qui permet de faire des recherches pour retrouver un produit par numéro AMM, usage ou composition. Ces données servent principalement aux professionnels du secteur pour savoir si un produit est autorisé ou pour connaître les substances actives, le titulaire des autorisations, et les usages possibles d'un produit.

2.1 Description des fichiers sources

Les fichiers publiés mensuellement sur la plateforme data.gouv.fr[9] sont un export du catalogue en CSV et XML. Les fichiers proposés au format CSV, sont vérifiés et offrent une équivalence complète avec les données publiées sur le site E-Phy et sont au nombre de neuf :

— La liste des produits autorisés ou retirés,
— La liste des usages des produits (hors MFSC),

[4] http://any23.apache.org
[5] https://openrefine.org
[6] https://github.com/protegeproject/cellfie-plugin
[7] https://www.talend.com/fr/

[8] C'est-à-dire que si nous avions dû concevoir un modèle de données "de zéro", nous n'aurions pas forcément modélisé cela. Par exemple, dans une ontologie on évite les classes qui sont des unions de concepts (ici MFSC).
[9] www.data.gouv.fr/fr/datasets/donnees-ouvertes-du-catalogue-e-phy-des-produits-phytopharmaceutiques-matieres-fertilisantes-et-supports-de-culture-adjuvants-produits-mixtes-et-melanges

— La liste des usages des produits (hors MFSC) autorisés,
— La liste des phrases de risque des produits,
— La liste des substances actives,
— La liste des conditions d'emploi des produits,
— La liste des classes et des mentions danger des produits (hors MFSC),
— La liste des usages des MFSC et produits mixtes,
— La liste des compositions des MFSC et produits mixtes.

Quant aux fichiers XML, ils sont au nombre de six : un fichier XML contenant les données et cinq autres de type XSD faisant office de description de la structure des données.

2.2 Description des caractéristiques du catalogue

Le catalogue E-Phy contient principalement des intrants. Les intrants sont tous *les différents produits apportés aux terres et aux cultures, qui ne proviennent ni de l'exploitation agricole, ni de sa proximité. Les intrants ne sont pas naturellement présents dans le sol, ils y sont rajoutés pour améliorer le rendement des cultures*.[10] Parmi ces produits :

— Les *PPPs* (Produit PhytoPharmaceutiques) sont des préparations destinées à protéger les végétaux et les produits de culture.
— Les *MFSCs* (Matières Fertilisantes et Supports de Culture) sont des produits destinés à assurer ou à améliorer la nutrition des végétaux, ainsi que les propriétés des sols. Les supports de culture sont destinés à servir de milieu de culture à certains végétaux.
— Les *adjuvants* sont des substances qui renforcent l'action des produits phytosanitaires en augmentant le pouvoir d'absorption du produit par la plante, un insecte, le bois, etc.
— Les *produits mixtes* sont composés soit d'une matière fertilisante ou d'un support de culture et d'un produit phytopharmaceutique, de façon à avoir un double effet.
— Les *mélanges* se composent de plusieurs produits phytopharmaceutiques rendus solubles et bénéficiant chacun d'une autorisation de mise sur le marché à titre individuel.

Le catalogue contient, comme indiqué sur la Figure 1, majoritairement des PPPs. Dans sa version d'octobre 2020, que nous avons utilisé, il en compte 13087 pour un total de 14093 intrants.

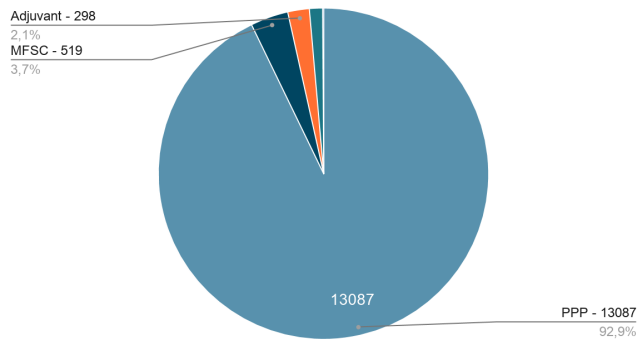

Fig. 1. Répartition des types d'intrants dans E-Phy.

[10] https://www.agriculture-nouvelle.fr/qu-est-ce-qu-un-intrant

En plus des intrants le catalogue contient d'autres informations utiles telles que :

— Les *substances actives*, qui constituent le principe actif des produits, ce sont elles qui agissent sur les nuisibles. Les substances actives sont homologuées au niveau européen; la version que nous avons utilisée en compte 1247.
— *Le catalogue des usages*, auquel E-Phy fait référence, est publié sous forme de note de service au Bulletin officiel du ministère chargé de l'agriculture.[11] Ce dernier n'existe pas non plus sous forme RDF. La version que nous avons utilisée compte 1540 usages.
— Les *cultures préconisées* pour lesquelles un MFSC ou un produit mixte peut être utilisé ; ils sont au nombre de 144 dans la version actuelle.

3 Méthodologie

Afin d'atteindre notre objectif d'avoir une ressource RDF représentant le catalogue E-Phy et dont le processus de création soit automatisé et reproductible, nous avons analysé sa structure, son contenu, essayé de comprendre le sens de chacun de ces éléments afin d'en avoir une vue d'ensemble. L'objectif était de pouvoir détecter des axes d'amélioration que ça soit côté modélisation, OWL nous permettant d'exprimer des aspects non envisageables dans le modèle relationnel de base, ou côté alignement avec d'autres ressources existantes. Durant cette première étape nous sommes parti d'un diagramme de classe UML strictement identique au modèle des données sources auquel nous avons ajouté petit à petit des modifications afin de le rendre compatible avec une approche orientée ontologie. Le tout en gardant une rétrocompatibilité avec le modèle original pour ne pas trop éloigner les données et également afin de pouvoir automatiser la re-création de l'ontologie et de la base de connaissance à chaque mise à jour des données source. Dans une deuxième étape, nous avons construit une ontologie OWL qui implémente le modèle de données et qui servira de structure d'accueil pour les instances du catalogue.

Après finalisation du modèle ontologique, la prochaine étape a été de préparer l'alignement. Cela consiste à rechercher les ressources cibles candidates pour établir des correspondances avec les classes et les instances de l'ontologie E-Phy et à évaluer pour chacune leurs taux de couverture des données afin de décider de la pertinence de l'alignement et des stratégies à suivre en cas d'éléments manquants.[12] Ces correspondances/alignements peuvent ensuite être utilisés pour diverses tâches d'intégration de données. Nous avons mis en place une méthode semi-automatique avec une validation humaine –faites par les auteurs– pour certain des termes à aligner (éléments chimique, substances, etc) pour lesquels de nombreuses correspondances possibles ont été identifiées en utilisant des portails d'ontologies / vocabulaires de références comme BioPortal et AgroPortal. Il fallait alors sélectionner les plus pertinentes et veiller à garder la cohérence sémantique

[11] https://daaf.reunion.agriculture.gouv.fr/Catalogue-national-des-usages

[12] Pour favoriser la valorisation des alignements, il est fréquent et logique de chercher à identifier la ressource sémantique la plus répandue/standard qui recouvre au mieux l'ontologie à aligner.

(e.g., liée à la hiérarchie d'origine). Nous avons utilisé `skos:exactMatch` ou `skos:closeMatch` pour encoder les alignements des classes de l'ontologie.
Finalement, l'étape d'instanciation (c.-à-d l'importation des données dans l'ontologie) clôture le processus, comme indiqué sur la Figure 3 et détaillée ci-après.

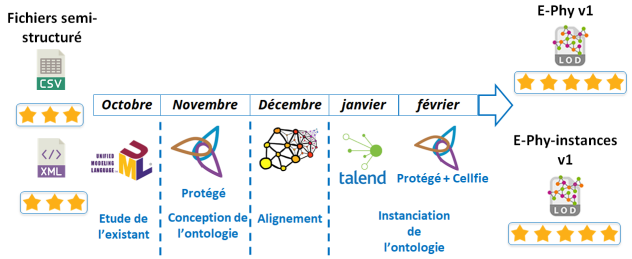

Fig. 2. Organisation dans le temps de la démarche de RDFisation.

La première phase est la préparation des données en fichiers Excel prêts à l'import à partir des fichiers CSV d'origine. Pour ces opérations de prétraitement, nous avons utilisé l'outil d'ETL Talend, en exécutant des scripts automatisés et reproductibles. Nous avons nettoyé les doublons et réorganisé les fichiers pour utiliser Cellfie, le plugin de Protégé qui crée les instances RDF en se basant sur des règles de mapping. Nous avons également créé le patron pour les URI en réutilisant les identifiants existants ou, lorsque nécessaire, en créant certains identifiants à partir d'un label (si unique) (e.g., *"ColorantBleuBrillant(acideBlue9)"* pour une substance) ou avec un numéro unique (e.g., *"usage_10"* pour un usage)). Nous avons également manipulé les données comme suit :

- Création des doses (e.g., *"dose min par apport : 2.0 L/ha"* pour un MFSC) en éclatant le champ texte d'origine en deux parties la valeur et l'unité;
- Scinder les colonnes de type liste en plusieurs lignes avec une colonne clé commune (e.g., la colonne "Variant" qui a comme valeur une liste de variants séparés par le caractère |: "bromoxynil | bromoxynil octanoate | bromoxynil butyrate | bromoxil (octanoate, heptanoate)");
- Éclater les identifiants des usages en trois parties, pour obtenir la portée d'usage, la méthode d'application et le groupe de nuisible (e.g., scinder "Ananas*Trt Part.Aer.*Act. Floraison" en "Ananas" , "Trt Part.Aer." et "Act. Floraison");
- Ajouter les alignements (détails section 4.1.3).

La deuxième phase de l'instanciation utilise les fichiers produits précédemment pour créer des instances dans l'ontologie (modèle) et la transformer en base de connaissances. Pour cela nous avons utilisé Protégé et son plugin Cellfie qui utilise le language *Mapping Master*[13] pour définir les mappings du contenu d'une feuille de calcul vers un triplet RDF.
Par exemple, la règle suivante permet de créer un individu de la classe *"Substance"* avec la valeur de la colonne B comme URI et label.

```
Individual: @B*
types: Substance
annotations:
        rdfs:label @B*
```

[13] www.github.com/protegeproject/mapping-master

Une phase de validation/contrôle de l'instanciation a été faite à postériori en comparant le nombre d'éléments de chaque type dans les fichiers sources et dans l'ontologie avec respectivement Talend et des requêtes SPARQL.

4 Résultats

4.1 La base de connaissances E-Phy

4.1.1 Etude de l'existant

Le but était de construire un diagramme de classe UML représentant la donnée d'origine, au total nous avons obtenu 18 classes (Table 2 et Figure 2).

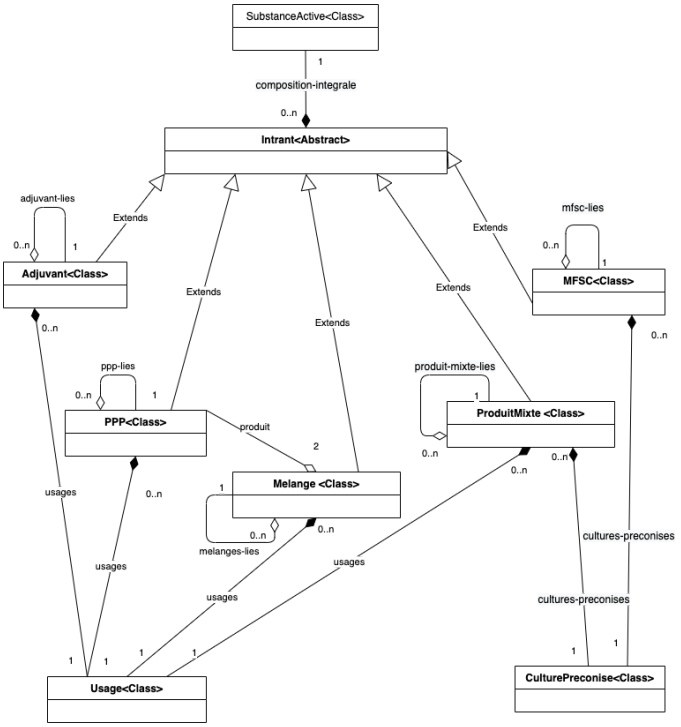

Fig. 3. Schéma UML simplifié d'une partie de la structure de la base de données du catalogue E-Phy.

Table 1. Description des principales classes du modèle UML.

Nom de la classe	*Exemple de propriétés*
Intrant (parent des classes PPP, MFSC, Adjuvant, Mélanges et Produit-mixte)	titulaire, type-produit, etat-produit, numero-AMM, nom-produit, type-commercial, mentions-autorisees, date-première-autorisation
Usage (utilisé par les classes PPP, Adjuvant, Mélanges et Produit-mixte)	id, identifiant-usage, stade-cultural-min, stade-cultural-max, etat-usage, dose-retenue
culture-preconise (utilisé par les classes MFSC et Produit-mixte)	id, type-culture, culture-commentaire , etat, date-decision

Substance (utilisé par intrant)	identifiant, famille-chimique, nom-produit, etat-produit, variants, numero-cas

A noter que beaucoup de propriétés définies dans les fichiers XSD ne sont jamais utilisées dans les données, tel que *produit-importé* d'Intrant, identifiant-*usage-groupe -organisme-nuisible, identifiant-usage-portée-usage*, *identifiant-usage-methode-application* d'Usage ou encore *autres-noms* de Substance. Nous avons noté aussi des éléments existants seulement dans le XML tel que les familles chimiques et ou a contrario seulement dans les CSV tel que les numéro-cas des substances.

4.1.2 Construction de l'ontologie

L'ontologie (ephy-v1.owl) obtenue à cette étape représente seulement la structure du catalogue E-Phy obtenu à partir des fichiers CSV/XML d'octobre 2020, les données en elle-même seront ajoutées après comme instances dans une base de connaissances séparée (ephy-full-v1.owl) qui importera (`owl:imports`) la structure. Nous avons explicitement séparé la structure des individus pour qu'elle puisse être utilisée individuellement au besoin, entre autres pour la vue des données ITAB (section 4.3). Pour chaque classe de l'ontologie, nous avons déclaré un label français (`rdfs:label`) et tant que possible une définition (`skos:definition`). La métadonnée de l'ontologie a été également décrite en utilisant les propriétés recensées dans MOD [7]. MOD intègre au total 23 vocabulaires de métadonnées existants standards (e.g, Dublin Core, OMV, DCAT, VoID) pour la description et la publication d'une ontologie. La Table 3 détaille quelques métriques sur l'ontologie créée.

Table 2. Métriques de l'ontologie E-Phy de structure

Métrique	***Valeur***
Nombre de classes	22
Nombre de propriétés objets (object properties) utilisées	28
Nombre de propriétés de données (data properties) utilisées	31
Nombre de propriétés d'annotations (annotation properties) utilisées	50
Nombre d'axiome au total	364

4.1.3 Alignement

La liste des ressources sémantiques que nous avons identifiées pour établir des correspondances avec l'ontologie E-Phy est décrite dans la Table 4 pour les instances et dans la Table 5 pour les classes.

Table 3. Ontologies cibles pour les instances.

Classe et ressource sémantique candidate	***Taux de couverture***	***Décision***
Alignement de famille chimique avec CHEBI *Chemical Entities of Biological Interest Ontology* est une classification structurée des composés chimiques d'importance biologique développée dans le cadre de l'OBO Foundry.	**83.33%** (115 familles chimiques trouvées et 23 non trouvées)	Aligner avec CHEBI lorsque possible en utilisant directement les URI des instances de la classe 'Chemical Entity' de CHEBI, sinon créer des nouvelles instances de cette classe dans notre espace de nom.
Alignement des unités avec UO *Units of Measurement Ontology* est un ensemble de d'unités métriques à utiliser avec entre autre avec PATO (Phenotypic Quality Ontology)	**18.35%** (29 unités trouvées et 129 non trouvées)	Abandonner l'alignement à UO car trop spécifique à la ressource E-Phy. Un travail plus significatif d'encodage des unités serait nécessaire.
Alignement des portées des usages et des types de culture avec FCU French Crop Usage est un thesaurus des types de cultures, organisés en fonction de leur usage, développé par INRAE et basé sur les définitions du Larousse Agricole.	**98,8%** (171 cultures trouvées et 2 cultures non trouvées). Les non trouvées sont en fait des cas spéciaux (e.g. "autres cultures").	Aligner avec FCU lorsque possible en utilisant les URI des instances (`skos:Concept`). Solliciter les auteurs de FCU pour leur proposer les cultures manquantes et pertinentes (ce qui nous a fait passer d'un taux de 63% à 98%).

Table 4. Ontologies cibles pour les classes.

Classes candidate	***Équivalents***
Culture (`skos:exactMatch`)	• crop - AnaEE Thesaurus • crop - Agronomy Ontology • crops - AGROVOC • crops - NALT
Saison d'application (`skos:closeMatch`)	• seasons - NALT, • season - ENVO, • seasons - AGROVOC
Intrant (`skos:exactMatch`)	• farm inputs - AGROVOC • farm inputs - NALT
Adjuvant (`skos:exactMatch`)	• adjuvants (AGROVOC) • Adjuvants (CHEBI) • adjuvants (NALT)
MFSC (`skos:exactMatch`)	• fertilizer - Agronomy Ontology • fertilizers (NALT) • fertilizers (AGROVOC)
PPP (`skos:exactMatch`)	• phytosanitary - AnaEE Thesaurus • pesticides (AGROVOC) • pesticides (Nalt) • pesticide (CHEBI)
Substance (`skos:exactMa`	• chemical substance - chebi • chemical substances -NALT)

tch)	• chemical substances (AGROVOC)
usage (skos:closeMatch)	• uses (AGROVOC)
restrictions (skos:closeMatch)	• use restrictions (AGROVOC)
dose (skos:exactMatch)	• dose specification (OBI) • dose (GEMET) • dosage (AGROVOC)

La recherche des ressources sémantiques candidates a été faite en utilisant les outils de recommandations (Recommender Service) des plateformes NCBO BioPortal [8] et AgroPortal [9]. Cette étape a été compliquée pour plusieurs raisons :

- Il fallait trouver les synonymes des labels qui ne matche pas directement.
- Les termes d'origine étant en français, il a fallu passer par une phase de traduction en anglais, non triviale pour les noms scientifiques.
- Le cas "Autres", utilisé très fréquemment dans les cultures (684 fois) est impossible à aligner.
- Le manque de cohérence des données d'origine, e.g., dans les cultures on peut trouver des traitements ou des types de terrains.
- Des granularités différentes dans les termes utilisés, avec par exemple des familles de culture et des cultures dans les mêmes champs/propriétés.
- Le cas des éléments trop spécifiques à la ressource d'origine, tels que les unités, qui font qu'il n'y a plus trop d'intérêt à faire l'alignement, vu le faible taux d'équivalence.

A terme, il nous faudrait une validation des alignements par des experts du sujets, entre autres pour les familles chimiques dont les noms peuvent avoir plusieurs écritures et variantes. Également, les Substances pourraient être également alignées à CHEBI mais leur nombre élevé (1274) nécessiterait la mise en place d'une méthode automatique plus fiable.

4.1.4 Instanciation

La difficulté de cette étape vient principalement du fait que les fichiers CSV ne contenaient pas toutes les données et que certaines ne se trouvaient que dans le XML (e.g., familles chimique d'une substance, nombre d'apport min et max de culture préconisé). Ces données ont dû être extraites séparément à partir du fichier XML. Les erreurs de saisie ont aussi représenté un problème significatif, tels que le nom d'un même élément écrit différemment (e.g., *"Pois écossées frais"* et "*Pois écossés frais*" ou *"Trt Part.Aer. "* sans point final et "*Trt Part.Aer*" avec).

L'utilisation de Cellfie était ensuite relativement simple une fois les fichiers préparés en amont. La Table 5 donne des exemples de règles Cellfie utilisées pour l'instanciation des substances.

Table 5. Instanciation des substances.

Fichier source		"substance_active_v3_Windows-1252.csv" et "famille_chimique.csv"
Nom du fichier produit	**Nom de ces colonnes**	**Règle Cellfie utilisée**
subtance_active_simple.xlsx	A famille_chimique B Nom_substance_active C Etat_d_autorisation	`Individual:@B* types: Substance annotations: rdfs:label @B* facts: etat @C*, familleChimique @A*`
	Explication	Créer un individus si non existant de la classe "Substance" avec la valeur de la colonne "B" comme URI et label et la colonne" C","A" comme valeur des propriétés etat et familleChimique respectivement
substances_active_variants.xlsx	A Nom_substance_active B Variant	`Individual: @A* facts: variant @B*`
	Explication	Créer un individus de la classe "Substance" si non existant avec la valeur de la colonne "A" comme URIet la colonne" B" comme valeur de la propriété variant

La base de connaissances obtenue contient l'intégralité des données du catalogue E-Phy (CSV et XML), apporte en plus un alignement avec CHEBI et FCU et enrichissent les données existantes d'informations extraites des champs textuels comme : les portées d'usage, les méthodes d'application, les groupes de nuisibles, les valeurs et unités des doses, les substances et les teneurs des substances actives. La Table 6 synthétise le nombre d'instances créées dans la base de connaissances.

Table 6. Métriques sur la base de connaissances E-Phy.

Métrique	***Valeur***
Nombre total d'individus	127 280
Nombre de propriétés d'objets (object properties) utilisées	291 162
Nombre de propriété de données (data properties) utilisées	377 764
Nombre de propriétés d'annotations (annotations properties) utilisées	72 880
Nombre total d'axiomes	996 702

Le code des scripts de traitement et de transformation ainsi que la documentation nécessaire pour reproduire une nouvelle version de l'ontologie E-Phy est publiquement disponible : https://github.com/d2kab/E-Phy-Ontology

Le dépôt GitHub contient également les fichiers intermédiaires ainsi que les diagrammes UML des données sources. L'ontologie E-Phy (plus exactement la base de connaissances) est mise à disposition[14] sur AgroPortal à l'URL : http://agroportal.lirmm.fr/ontologies/E-PHY [15]

4.2 Requêtes SPARQL

Pour illustrer la valeur ajoutée de l'ontologie E-Phy nous avons préparé quelques requêtes SPARQL qui valorisent la sémantique et les alignements et permettent des interrogations impossibles sur les données d'origine. Elles correspondent à des cas d'usage pressenti pour l'ontologie E-Phy.

Par exemple, la requête suivante exploite l'héritage et les synonymes du thesaurus FCU pour obtenir tous les produits utilisables pour la famille de culture "Arboriculture fruitière" à partir de son synonyme "verger" :

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX cropusage: <http://ontology.irstea.fr/cropusage/>
PREFIX ephy: <http://www.d2kab.org/ontologies/ephy#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT distinct ?pl ?ptypel ?fcuL ?fcud ?fcuInfo
WHERE {
 SERVICE <http://ontology.irstea.fr/cropusage/sparql>
 {
        ?fcuP a skos:Concept.
        {
          ?fucP skos:altLabel ?fcuPsyn.
          filter(?fcuPsyn = "verger"@fr)
        }
        UNION
        {
          ?fucP skos:prefLabel ?fcuPL.
          filter(?fcuPL = "verger"@fr)
        }
        ?fcu skos:broader* ?fcuP ;
             skos:definition ?fcud;
             skos:prefLabel ?fcuL;
             rdfs:seeAlso ?fcuInfo
 }
 {
    ?p rdfs:label ?pl.
    ?p a ?ptype.
    ?ptype rdfs:label ?ptypel.

   {
       ?iu ephy:porteeUsage ?fcu.
       ?u ephy:identifiantUsage ?iu.
       ?p ephy:usage ?u.
   }
   UNION
   {
       ?cp ephy:typeCulture ?fcu.
       ?p ephy:culturePreconise ?cp.
   }
}}
```

Exemple de résultats :

Table 7. Exemple d'un résultat de la requête utilisant l'alignement avec FCU .

pl (Label du produit)	"KORI FEUILLE"
ptypel (Type du produit)	"Matières fertilisantes et supports de culture" @fr
fcuL (Label de la culture dans FCU)	"arboriculture fruitière" @fr
fcud (Définition de la culture selon FCU)	"cultures des arbres et arbustes qui produisent des fruits comestibles; Les fruits peuvent subir une transformation et être mangés sous forme de jus de fruits alcoolisés ou non." @fr
fcuInfo (URL extérieure pour avoir plus d'information)	https://fr.wikipedia.org/wiki/Arbre_fruitier @fr

D'autres exemples de requêtes permettent d':

- Exploiter les informations de CHEBI pour obtenir la définition de toutes les familles chimiques qui composent un produit.
- Exploiter la structure de l'ontologie pour obtenir tous les produits utilisables pour une culture donnée ou produit par une société spécifique (ou les deux).
- Exploiter la structure de l'ontologie pour obtenir tous les produits efficaces contre un groupe d'organismes nuisibles (à améliorer à l'avenir en liant les agresseurs à une base de référence).

Ces requêtes sont disponibles dans le dépôt GitHub.

4.3 Une vue pour l'AB: l'ontologie ITAB

La ressource ITAB est un autre 'catalogue' sous le nom de *Guide des produits de protection des cultures utilisables en Agriculture Biologique en France* produit par L'Institut Technique de l'Agriculture Biologique (ITAB)[16] qui maintient, de manière distincte de l'ANSES, une liste des produits de protection des cultures, utilisables dans le cadre de la production biologique. Cependant, tous les produits du guide ITAB (version de septembre 2014 en ligne) sont inclus dans la base de données E-Phy dénotés avec la propriété "mentions-autorisées=utilisable en agriculture biologique". Etant donné que le guide ITAB est une ressource en soi, d'intérêt pour les professionnels de l'agriculture biologique, nous avons produit une autre base de connaissances, sous la forme d'une vue (i.e., un sous ensemble de triplets) de la base de connaissances E-Phy. L'ontologie ITAB est ainsi une base de connaissances qui contient le même modèle ontologique que l'ontologie E-Phy mais ne contient que les instances explicitement déclarées pour l'agriculture biologique (ou les instances reliées).

En RDF, pour obtenir la vue ITAB, il suffit de faire une requête SPARQL qui extrait les triplets pertinents de la base de connaissances E-Phy :

```
SELECT ?x ?pp ?v
WHERE { ?p rdf:type/rdfs:subClassOf ephy:Intrant;
```

[14] Les URIs proposées ne sont pas déréférençables et seront sans doute amenées à changer.

[15] La version actuelle d'AgroPortal ne permet pas de visualiser les instances, même si elles sont bien stockées dans le portail et accessibles globalement ou pour une classe via l'API REST : http://data.agroportal.lirmm.fr/ontologies/E-PHY/instances

[16] www.itab.asso.fr/downloads/com-intrants/guide-protection-plantes6.pdf

```
        ephy:mentionAutoriser ?ma;
        (!rdf:null)* ?x.
    ?x ?pp ?v
    filter(?ma = "Utilisable en agriculture biologique")
}
```

Il ne reste plus qu'à ajouter ces triplets dans une nouvelle ontologie (itab-v1.owl) qui importe l'ontologie E-Phy et lui attribuer un identifiant et des métadonnées propres. La base de connaissance ITAB est disponible sur AgroPortal comme une vue de l'ontologie E-PHY à http://agroportal.lirmm.fr/ontologies/ITAB

5 Conclusions et perspectives

A partir des données d'origine (CSV et XML) nous avons pu construire l'ontologie E-Phy (version 1), une base de connaissances (modèle ontologique et instances) du catalogue des produits phytopharmaceutiques et de leurs usages, des matières fertilisantes et des supports de culture autorisés en France. Notre méthodologie a été automatisée (sauf l'étape alignement) et est facilement reproductible pour reconstruire l'ontologie quand les données originales seront mises à jour. Elle a permis d'identifier plusieurs erreurs dans les données : erreurs structurelles (e.g., un produit sans AMM ou deux produits avec le même AMM, ce qui devrait être totalement impossible), erreurs de saisie, des erreurs de valeurs, valeurs générique (e.g., "Autres", "Traitement généraux", "Tous"), des champs vides ou encore des éléments qui n'ont pas d'identifiant prédéfini. Nous travaillons actuellement sur une synthèse que nous fournirons à l'ANSES dans une démarche de qualité. Actuellement, en restant totalement rétrocompatible avec les données d'origine, nous n'avions pas l'autorité (ou l'expertise) pour changer ou corriger les erreurs, ou pour changer de manière significative le modèle de données. La RDFization que nous avons faite permet à minima de détecter plus facilement les problèmes et de les corriger. Cependant, notre démarche se veut également incitative pour encourager l'ANSES à adopter les technologies du web sémantique de manière native dans le développement du catalogue. Ainsi, l'agence pourrait assigner des URIs pérennes dont elle serait responsable.

Parmi les améliorations envisageables, dans la version actuelle, il s'agirait d'aller plus loin dans l'alignement en ajoutant les substances avec CHEBI et restructurer l'utilisation des unités dans les doses avec une ontologie des unités de mesures (UO, QUDT, etc.). Un autre axe d'amélioration serait de RDFizer le *catalogue national des usages phytopharmaceutiques* (actuellement disponible sous forme d'un document PDF seulement) produit par la direction générale de l'alimentation. Ainsi l'ontologie E-Phy pourrait se reposer formellement sur ce catalogue pour aller plus loin dans la représentation des usages.

Dans le cadre du projet ANR "Des Données aux Connaissances en Agronomie et Biodiversité (D2KAB – www.d2kab.org) nous prévoyons d'utiliser l'ontologie E-Phy pour annoter sémantiquement des occurrences de produits ou d'intrants dans les Bulletins de Santé du Végétal qui sont utilisés par les agriculteurs pour faire de la veille sanitaire pour leurs cultures. En partenariat avec la SME Elzeard, l'ontologie E-Phy sera utilisée pour construire un graphe de connaissances exploité dans une application pour la gestion des itinéraires culturaux à destination des maraîchers.

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR D2KAB (ANR-18-CE23-0017) lors d'un stage soutenu par l'Institut de Convergence en Agriculture Numérique, #DigitAg (ANR-16-CONV-0004). Nous remercions le *service des systèmes d'informations des produits réglementés* de l'ANSES pour les renseignements sur la base E-Phy. Nous remercions également C. Roussey (INRAE) pour son aide sur les alignements avec FCU, S. Aubin (INRAE) et O. Corby (INRIA) pour son aide avec l'extraction de la vue ITAB.

Références

[1] T. Berners-lee, "Linked Data - Design Issues," *Design Issues*, 2006. https://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Roussey, X. Delpuech, F. Amardeilh, S. Bernard, C. Jonquet. Semantic Description of Plant Phenological Development Stages, starting with Grapevine. *14th international conference on Metadata and Semantics Research Conference (MTSR)*, Dec 2020, Madrid, Spain. pp.257-268, ff10.1007/978-3-030-71903-6_25ff..

[3] S. Eholié, M. D. T. Nzali, S. Bringay, and C. Jonquet, "MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein," Jun. 2016.

[4] M. Poveda-Villalón, "A reuse-based lightweight method for developing linked data ontologies and vocabularies," in *Lecture Notes in Computer Science*, 2012, vol. 7295 LNCS, pp. 833–837, doi: 10.1007/978-3-642-30284-8_66.

[5] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller, "Triplify - Light-weight Linked Data publication from relational databases," in *WWW'09 - Proceedings of the 18th International World Wide Web Conference*, 2009, pp. 621–630, doi: 10.1145/1526709.1526793.

[6] J. H. Gennari *et al.*, "The evolution of Protégé: An environment for knowledge-based systems development," *Int. J. Hum. Comput. Stud.*, vol. 58, no. 1, pp. 89–123, Jan. 2003, doi: 10.1016/S1071-5819(02)00127-1.

[7] B. Dutta, A. Toulet, V. Emonet, and C. Jonquet, "New generation metadata vocabulary for ontology description and publication," in *Communications in Computer and Information Science*, Nov. 2017, vol. 755, no. 755, pp. 173–185, doi: 10.1007/978-3-319-70863-8_17.

[8] N. F. Noy *et al.*, "BioPortal: Ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 2, 2009, doi: 10.1093/nar/gkp440.

[9] C. Jonquet *et al.*, "AgroPortal: A vocabulary and ontology repository for agronomy," *Comput. Electron. Agric.*, vol. 144, pp. 126–143, Jan. 2018, doi: 10.1016/j.compag.2017.10.012.