Alignement entre sources : cas d'usage des plantes cultivées

F. Michel[3], F. Amardeilh[1], R. Bossy[2], C. Faron[3], C. Roussey[4], C. Noûs[5]

[1] Elzeard R&D, Pessac, France
[2] Université Paris-Saclay, INRAE, UR MAIAGE, 78350 Jouy-en-Josas, France
[3] Université Côte d'Azur, Inria, CNRS, I3S, 06902, Sophia-Antipolis, France
[4] Université Clermont Auvergne, INRAE, UR TSCF, Aubière, France
[5] Laboratoire Cogitamus, Havre de recherche, France.

florence.amardeilh@elzeard.co, robert.bossy@inrae.fr, faron@i3s.unice.fr, fmichel@i3s.unice.fr, catherine.roussey@inrae.fr, camille.nous@cogitamus.fr

Résumé

Dans cet article nous décrivons nos premiers travaux sur l'alignement de deux graphes de connaissances complémentaires utiles dans le domaine de l'agriculture : le thésaurus des usages des plantes cultivées (FCU) et le registre taxonomique national français TAXREF pour la faune, la flore et la fonge. Plusieurs méthodes d'alignement spécifiques à ce cas d'usage ont été implémentées. Les résultats montrent que dans ce domaine il sera nécessaire de nettoyer les alignements produits automatiquement.

Mots-clés

web sémantique, graphes de connaissances, alignement, taxonomie biologique, TAXREF-LD, thesaurus agricole, FrenchCropUsage, SKOS, plantes.

Abstract

In this article we describe our first work on the alignment of two complementary knowledge graphs useful in the agricultural domain. A SKOS thesaurus related to uses of cultivated plants (FCU), a knowledge graph of the biological taxonomy in France (TAXREF-LD). Several alignment methods specific to this use case have been implemented. The results show that for this use case it will be necessary to curate the alignments produced automatically.

Keywords

semantic web, knowledge graph, alignment, biological taxonomy, TAXREF-LD, agricultural thesaurus, FrenchCropUsage, SKOS, plant.

1 Introduction

Le projet ANR *Des Données aux Connaissances en Agronomie et Biodiversité* (D2KAB) illustre comment l'ingénierie des connaissances contribue au développement d'applications innovantes dans le domaine de l'agriculture. L'objectif de D2KAB est de créer un cadre pour transformer les données d'agronomie et de biodiversité en connaissances décrites sémantiquement, interopérables, exploitables et ouvertes. Pour construire un tel cadre, nous nous appuyons sur des ressources sémantiques (terminologies, vocabulaires, ontologies) pour décrire nos données et les publier en tant que données ouvertes liées [1]. Nous utilisons notamment le portail AgroPortal [8] pour trouver, publier et partager ces ressources sémantiques puis nous les exploitons dans des applications dédiées à l'agriculture ou l'environnement.

Alors que le web de données liées met à disposition de plus en plus de graphes de connaissances, leur réutilisation croisée reste souvent un défi. Cet article présente une méthode permettant d'aligner entre eux des graphes de connaissances représentant des points de vue différents sur les mêmes objets d'étude, afin de requêter conjointement ces graphes pour des raisons de complétude d'information. L'agriculture offre un cas d'usage particulier dans ce domaine, lié à la modélisation des plantes cultivées. Plusieurs expertises sont nécessaires pour décrire une plante cultivée : agriculteur versus agronome, agronome versus écologue. Le monde scientifique (les écologues ou agronomes) utilise les noms scientifiques issus de la science taxonomique pour désigner les organismes vivants (plantes, insectes). Ces noms scientifiques sont stockés dans des taxonomies biologiques. Le monde des usagers (les agriculteurs) utilise des noms vernaculaires pour désigner les organismes vivants qui interviennent dans leur pratique. De plus, une plante peut avoir plusieurs rôles en agriculture : (1) une plante cultivée dans un but de production, (2) une adventice (mauvaise herbe), (3) une plante cultivée dite plante de service, pour rendre un service à une autre plante cultivée, dite alors production principale (la plante de service est détruite sans être récoltée, au contraire de la plante de production).

Nous présentons tout d'abord les travaux sur l'alignement des taxonomies dans le domaine agricole. La section 3 décrit en détail des sources utilisées dans notre approche d'alignement. La section 4 présente les algorithmes d'alignements mis en oeuvre. La section 5 présente deux types d'évaluations effectuées sur les résultats de nos algorithmes. Enfin, nous concluons nos travaux en présentant des perspectives d'amélioration.

2 Travaux antérieurs

Il existe déjà plusieurs graphes de connaissances qui essaient de combiner les points de vue des agronomes et des agriculteurs. Le plus connu est le thésaurus Agrovoc de la FAO [2]. Nous débuterons par décrire les travaux sur l'alignement des taxonomies biologiques, qui ont été utilisées pour évaluer les outils d'alignement.

2.1 Les taxonomies biologiques

La taxonomie est la science de la diversité du vivant. Elle consiste à décrire les organismes vivants et à les organiser en groupes appelés *taxons*, selon une hiérarchie reflétant l'histoire de leur évolution [9]. Les taxons se situent à différents niveaux de généralité, appelés *rangs taxonomiques*, parmi lesquels on peut citer l'espèce, la famille, ou la variété.

Il existe de nombreux référentiels taxonomiques, ou taxonomies. La difficulté de leur maintenance par les curateurs, et *in fine* du choix d'un référentiel taxonomique tient, à la fois de leur volume et de la volatilité structurelle de leur contenu. En effet, à ce jour [3] recense près de deux millions d'espèces décrites, dont plus de 300.000 espèces de Magnoliophytes (*"plantes à fleurs"*) auxquelles appartiennent la majorité des plantes cultivées. De plus, le contour des unités conceptuelles des taxonomies est instable puisque les taxons et leur organisation constituent les hypothèses de travail des systématiciens. Cela signifie que les taxonomies sont soumises aux controverses scientifiques ; les taxonomies décrites à une date donnée peuvent être réfutées dans l'avenir.

Ce constat a aussi amené les systématiciens à réguler strictement les conventions de nommage des taxons. Les codes de nomenclature [14] permettent aux systématiciens de stabiliser la nomenclature face à la volatilité des taxonomies. Ces conventions génèrent différents types de recombinaisons pouvant induire des synonymes, voire des homonymes, faisant de la constitution, la maintenance et l'alignement des référentiels taxonomiques une tâche difficile.

2.1.1 Stratégies de curation

Nous mentionnerons trois référentiels parmi les plus complets et utilisés : NCBI Taxonomy, TAXREF, et Catalogue of Life. Ces trois référentiels couvrent bien les différents contours et les différentes politiques de curation.

NCBI Taxonomy [17] est la taxonomie de référence des bases de données maintenues par le NCBI, dont PubMed et GenBank. La taxonomie est complétée au gré des besoins selon les entrées des bases de données du NCBI. L'organisation hiérarchique est assurée par des curateurs volontaires qui se basent sur la littérature en systématique. La taxonomie du NCBI est constitutionnellement biaisée par l'abondance des études et reflète mal la biodiversité. De plus, une politique de stabilité des identifiants de taxons amène quelquefois des inexactitudes. Le principal avantage de NCBI Taxonomy reste les nombreux liens vers des bases de données moléculaires et bibliographiques.

TAXREF [7] est le référentiel taxonomique du Système d'Information de l'Inventaire National du Patrimoine naturel, diffusé et maintenu par le Muséum National d'Histoire Naturelle (MNHN). Il liste toutes les espèces recensées dans les territoires français (métropole et outre-mer), ainsi que plus de 650.000 noms scientifiques associés aux taxons de tous rangs taxonomiques. Les curateurs de TAXREF sont en contact direct avec les spécialistes identifiés pour chaque branche du vivant. TAXREF constitue donc une source primaire d'une taxonomie scientifiquement fondée pour les espèces que l'on trouve en France. TAXREF-LD est la distribution de TAXREF respectant les principes des données liées. Ce graphe de connaissance est décrit en détails dans la section 3.

Catalogue of Life [16] est un projet à l'ambition universelle, soutenu par le Global Biodiversity Information Facility (GBIF). Il s'agit d'une fédération de référentiels, chaque composant couvrant une branche du vivant (e.g. LPSN), un environnement (e.g. WoRMS) ou une zone géographique (e.g. ITIS). Cette stratégie permet d'obtenir un compromis entre exhaustivité et justesse.

2.1.2 Alignement entre taxonomies

L'alignement automatique des taxons issus de différents référentiels taxonomiques reste une question scientifique ouverte qui a notamment motivé plusieurs tâches dans la campagne Ontology Alignment Evaluation Initiative (OAEI) [1]. Les tâches *OAEI Taxon* et *Biodiv* portent sur la détection d'alignements entre taxons biologiques.

La tâche *OAEI Taxon* porte sur la détection d'alignements complexes. Le benchmark de *OAEI Taxon* est composé de quatre référentiels taxonomiques représentés sous forme de graphes de connaissances. En 2021, seulement 3 systèmes automatiques d'alignement sur 11 ont pu proposer des alignements valides : ATM, Fine-TOM et logmap [15]. De plus, la plupart des alignements proposés étaient jugés simples.

La tâche *Biodiv* porte sur la détection d'alignements simples dans le domaine de la biodiversité. L'un des benchmarks de *Biodiv* se compose de TAXREF-LD et NCBI Taxonomy. Malheureusement à cause de la taille de ces deux taxonomies, aucun système n'a été capable de proposer des alignements. A noter que l'outil AgreementMaker-Light (AML) a obtenu les meilleurs résultats sur les deux autres benchmarks de cette tâche [6].

En outre, les référentiels taxonomiques les plus connus et adoptés font rarement le travail de produire des alignements avec d'autres référentiels. Les auteurs de TAXREF ont fait ce travail en alignant TAXREF-LD et plusieurs autres référentiels dont NCBI Taxonomy. Ce calcul a été mis en oeuvre à l'aide de l'outil SILK [18] et d'une extension développée pour implémenter les règles métier d'alignement de noms scientifiques [2].

1. http://oaei.ontologymatching.org/
2. https://github.com/frmichel/taxrefmatch-silk-plugin

2.2 Sources associant des taxonomies biologiques et des noms vernaculaires de plantes cultivées

A notre connaissance, il existe trois sources qui proposent une représentation multiple des plantes cultivées avec une terminologie française : Agrovoc, la base de données mondiale EPPO et le catalogue du GEVES.

2.2.1 Agrovoc

Le thésaurus Agrovoc est publié par l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO) [2]. Il est édité manuellement par une communauté mondiale d'experts et couvre tous les domaines d'intérêt de la FAO, y compris l'agriculture, la sylviculture, la pêche, l'alimentation et les domaines connexes. Il est disponible en 29 langues, avec une moyenne de 35 000 termes par langue et développé en SKOS-XL [12]. La force de ce thésaurus est sa couverture lexicale multilingue. Il est donc souvent utilisé pour annoter ou indexer des documents ou des images relatifs au domaine de l'agriculture. Agrovoc contient plusieurs représentations des plantes décrites dans des branches différentes de sa hiérarchie. Une plante peut être décrite par son nom scientifique. Par exemple, le `skos:Concept` http://aims.fao.org/aos/agrovoc/c_8283 représente l'espèce *"Vitis vinifera"*, qui a pour parent un `skos:Concept` représentant le genre *"Vitis"*. Une plante peut être décrite par son nom vernaculaire associé à la filière agricole qui la cultive. Par exemple le `skos:Concept` http://aims.fao.org/aos/agrovoc/c_3360 représente la vigne et a pour parent les cultures fruitières. Il existe plusieurs types de liens possibles entre ces deux branches d'Agrovoc, par exemple "includes" ou "is used as". La branche représentant les noms vernaculaires des plantes cultivées a plusieurs composants dont il n'est pas évident de comprendre la logique. Par exemple, le concept de vigne est associé :

- aux concepts de "vitis vinifera", "vitis labrusca", "vitis rotundifolia" et "vitis aestivalis" par un lien "includes",
- au concept de "vitis" par un lien "included in",
- au concept "beverage crop" par un lien "is used as",
- au concept de raisin par un lien "produces".

.

2.2.2 La base de données EPPO

La base de données mondiale EPPO [4] est maintenue par le Secrétariat de l'*Organisation Européenne et Méditerranéenne pour la Protection des Plantes* (EPPO) [3]. L'objectif de cette base est de fournir des informations spécifiques aux organismes nuisibles, qui ont été produites ou collectées par l'EPPO. Le contenu de la base de données est constamment mis à jour par le Secrétariat de l'EPPO. Cette base est interrogeable en ligne ou par le biais d'une API. Chaque plante est identifiée par un code de 5 caractères qui sert de référence dans de nombreuses autres bases de données agricoles européennes. Cette base identifie aussi des groupes de plantes (taxon biologique, filière agricole, ...) par des codes à 5 caractères. Par exemple le code pour l'espèce *"Vitis vinifera"* est VITVI. Le code pour le genre *"Vitis"* est 1VITG. EPPO représente aussi d'autres classifications des plantes (crop groups, commodity groups, crop destination,...). Les filières agricoles sont partiellement représentées dans la classification des crop groups (3CRGK). Cette classification contient par exemple les cultures fruitières (fruit crops : 3FRUC) et les légumes (vegetable crops : 3VEGC). Pour chacun de ces groupes sont affichés la liste des taxons associés. Ainsi l'espèce *"Vitis vinifera"* apparaît comme taxon associé des cultures fruitières. La filière vigne n'existe pas dans la classification crop groups.

2.2.3 Catalogue du GEVES

Le Groupe d'Etude et de contrôle des Variétés Et des Semences (GEVES) produit un catalogue officiel des espèces et variétés de plantes cultivées en France [4]. Ce catalogue contient 9 000 variétés pour 190 espèces. Toute variété, produite par un institut agricole est inscrite dans ce catalogue pour être commercialisé. Une variété de plante est décrite entre autre par son nom, le détenteur de la variété, une indication de son type variétal, son espèce biologique, sa filière agricole. Par exemple la variété de raisin de cuve abouriou, produite par Institut Français de la Vigne et du Vin, a comme indication de type variétal "couleur de baie blanche" et comme espèce *"Vitis vinifera"*. Ce catalogue est disponible sous forme de plusieurs fichiers CSV ou d'une API.

Ces trois sources montrent que le rang espèce des taxons biologiques est associé au nom vernaculaire de la plante pour identifier au mieux une plante cultivée. A notre connaissance cet article présente un premier cas d'étude d'alignement automatique de graphes de connaissances complémentaires dans le domaine de l'agriculture : un alignement de taxons biologiques avec des noms d'usage des plantes cultivées.

3 Sources à aligner : TAXREF-LD et FCU

Les deux graphes de connaissances que nous cherchons à aligner sont fondés sur le vocabulaire SKOS ou une extension de SKOS. TAXREF-LD est la publication du référentiel TAXREF sur le web de données liées. FCU est un thésaurus francophone des usages des plantes cultivées. Ces deux graphes complémentaires ont en commun uniquement les noms vernaculaires des plantes cultivées.

3.1 TAXREF et TAXREF-LD

TAXREF [7] est le référentiel taxonomique français pour la faune, la flore et la fonge. Outre un portail Web, un service REST et un ensemble de fichiers CSV téléchargeables, TAXREF est disponible sous forme d'un graphe de connaissances respectant les principes des données liées, nommé TAXREF-LD [11] [5].

3. https://www.eppo.int/

4. https://www.geves.fr/catalogue/

5. TAXREF-LD peut être téléchargé depuis https://doi.org/10.5281/zenodo.5876775. Il est possible de l'interroger par le biais d'un SPARQL endpoint public, et voir les informations disponibles dans le dépôt https://github.

Afin de refléter fidèlement la distinction entre taxonomie et nomenclature, TAXREF-LD comporte deux niveaux distincts de modélisation illustrés par la figure 1. Au niveau taxonomique, chaque taxon biologique est modélisé comme une classe OWL dont les membres sont les individus biologiques de ce taxon. La classe parente est le taxon de rang supérieur (e.g. *"Daucus carota"* est de rang espèce, la classe parente, *"Daucus"*, est de rang genre). Au niveau nomenclatural, les noms scientifiques sont représentés comme les concepts d'un thésaurus SKOS. Chaque nom (concept SKOS) est lié à un taxon (classe OWL) par une propriété indiquant s'il s'agit du nom de référence (nom *accepté* en zoologie ou *valide* en botanique), ou d'un synonyme.

Outre l'information strictement taxonomique, TAXREF-LD représente également d'autres types d'information : noms vernaculaires, habitats, statuts de conservation, statuts biogéographiques, interactions entre espèces, ainsi que les références bibliographiques associées à ces différentes informations. A noter que TAXREF-LD associe parfois le même nom vernaculaire à plusieurs taxons. Ces noms vernaculaires sont issus des publications où sont déclarés les noms scientifiques. Par ailleurs, TAXREF-LD est lié à plusieurs référentiels taxonomiques tierces dont Agrovoc et NCBI Organismal Taxonomy.

3.2 Thésaurus agricole FCU

Le thésaurus intitulé "usages des plantes cultivées en France" ou French Crop Usage (FCU) [6] normalise les noms de plantes cultivées en français. De plus, il les organise dans des catégories représentant des filières agricoles : par exemple, *"fourrage"* et *"grandes cultures"* sont deux exemples de filières agricoles. Ainsi, une hiérarchie est formée par des relations de généralisation/spécialisation entre les filières agricoles et les noms d'usage des plantes cultivées : par exemple, *"grandes cultures"* se spécialise en *"céréales"*). Les termes du thésaurus ont été sélectionnés manuellement à partir de documents de référence. Les documents étudiés pour construire le thésaurus sont :

Les statistiques agricoles annuelles de l'Agreste [7], les métadonnées du registre parcellaire graphique, le classement des plantes cultivées par groupe d'usage proposé par wikipédia France, le catalogue officiel des espèces et variétés de plantes cultivées en France du GEVES, les fiches "les plantes fourragères pour les prairies" du GNIS [8], la base Ephy qui décrit l'usage des produits phytosanitaires sur les plantes, le Larousse Agricole.

Concernant les légumes et leur classification, les points communs entre plusieurs sources ont été cherchés : Wikipedia, Bonduelle, FranceAgriMer, Encyclopedia Universalis, La ferme du Bec Hellouin. Notons qu'il n'existe pas de consensus sur la classification des légumes. Le choix des noms d'usage des plantes cultivées, les définitions associées et leur organisation ont été discutés par au moins un expert de la filière agricole. Ce thésaurus n'est pas complet et évolue en fonction des projets.

Le thésaurus est modélisé à l'aide du vocabulaire SKOS proposé par le W3C [13], la figure 2 en présente un extrait. Il est disponible sur le web de données liées [9]. FCU contient 526 `skos:Concepts`. La profondeur maximale de la hiérarchie est de 6. Chaque concept est défini par un ensemble d'étiquettes (les noms vernaculaires de la plante), des notes, des liens vers d'autres sources d'information et de liens hiérarchiques. Nous présentons une liste succincte des propriétés utilisées pour définir chaque `skos:Concept` :

- `skos:prefLabel` : contient le terme utilisé comme étiquette préférée du concept en français. En général, le terme est le nom vernaculaire de la plante cultivée avec une indication de son usage (ex : *"vigne ornementale"*, *"vigne cultivée"* ou *"vigne de cuve"*). Au besoin, cette étiquette préférée peut être construite de manière artificielle pour bien identifier l'usage agricole de la plante. En générale, cette étiquette est prise dans l'une des sources identifiées. Par exemple, les trois étiquettes (*"vigne ornementale"*, *"vigne cultivée"* ou *"vigne de cuve"*) ont toutes été trouvées dans une des sources.
- `skos:altLabel` : contient les autres termes qui peuvent être utilisés comme étiquettes du concept (ex : *"vigne vierge"* est une autre étiquette de *"vigne ornementale"*). Ces étiquettes peuvent indiquer le produit récolté (ex : "raisin") ou l'activité agricole (ex : viticulture).
- `skos:definition` : contient la définition en français du concept justifiant sa position dans la hiérarchie du thésaurus.
- `rdfs:seeAlso` : contient un lien web vers une définition retenue lors de la construction du thésaurus, comme par exemple les pages wikipédia.
- `skos:note` contient au moins une définition trouvée dans une autre source comme l'Agreste ou wikipédia.

Lorsqu'une plante a plusieurs usages, elle est représentée par plusieurs concepts : un concept pour chacun des usages, plus un concept pour l'ensemble des usages, parent des concepts précédents. Un concept dans la branche "multiusage" porte le nom vernaculaire de la plante sans indication d'usage (par exemple *"carotte"*). Ce concept est ensuite décliné en autant de fils qu'il y a d'usages (*"carotte potagère"* pour l'alimentation humaine et *"carotte fourragère"* pour l'alimentation animale). Chacun des fils est de plus positionné à un seul endroit dans la branche *"usage des plantes cultivées"*. Dans la figure 3 le concept *"carotte potagère"* est positionné comme fils du concept *"légume racine"*.

3.3 Difficultés pour aligner les sources

Bien que les taxonomies et listes de plantes cultivées référencent les mêmes objets du monde (les organismes vivants), leur alignement présente plusieurs difficultés qui rendent nécessaire une validation manuelle en pratique.

La taxonomie est une science difficile à appréhender pour

com/frmichel/taxref-ld/.

6. https://doi.org/10.15454/QHFTMX
7. l'Agreste est le service statistique ministériel de l'agriculture
8. Le GNIS est l'interprofession des semences et plants, il a été renomé SEMAE
9. http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage

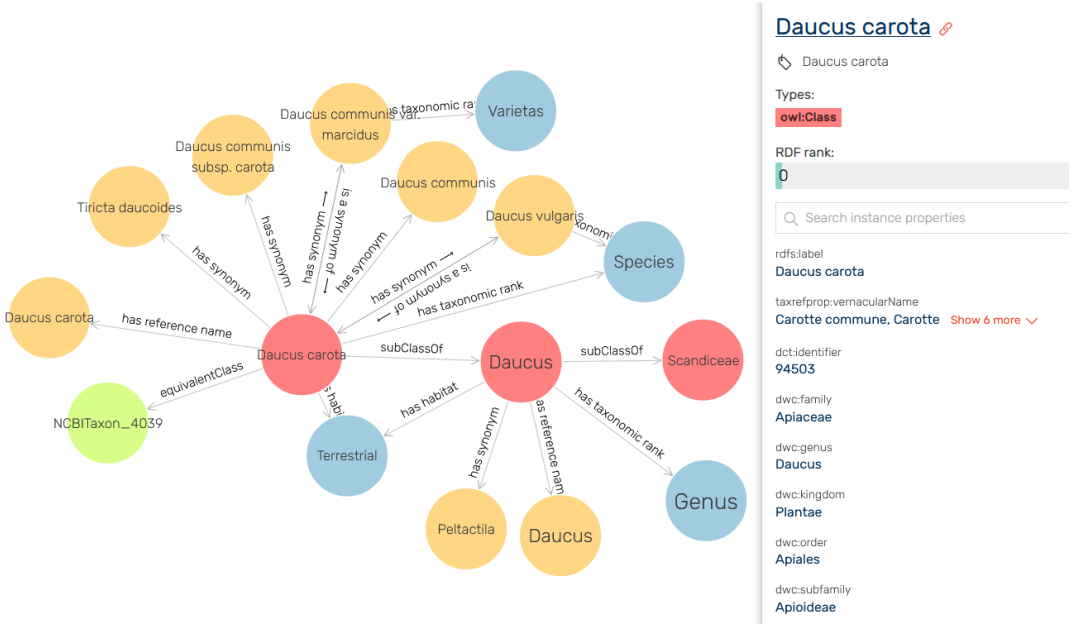

FIGURE 1 – Deux niveaux de modélisation dans TAXREF-LD, exemple de la carotte. Noeuds rouges : taxons modélisés comme des classes OWL. Noeuds oranges : noms scientifiques modélisés comme des concepts SKOS associés aux taxons en tant que nom de référence ou synonyme.

les non-spécialistes. Un taxon est une hypothèse scientifique affirmant qu'un ensemble d'individus biologiques appartiennent au même groupe taxonomique en raison de certaines caractéristiques communes. Dans le cas simple, à un taxon est associé un nom scientifique. Cependant, l'évolution du consensus scientifique entraîne des changements dans la taxonomie, ainsi des recombinaisons peuvent se produire : deux taxons peuvent être fusionnés en un seul, un taxon existant peut être divisé en deux taxons distincts, ou un taxon peut changer de rang taxonomique (espèce vers sous-espèce par exemple). Par conséquent, un taxon peut avoir un nom préféré utilisé pour désigner le taxon, et plusieurs synonymes. Les noms et leurs recombinaisons sont publiés dans la littérature scientifique, toutefois la prise en compte de ces évolutions dans les taxonomies et les liste de plantes cultivées peut se faire à des rythmes différents, ce qui mène fréquemment à des désaccords. Par exemple, une liste de plantes peut utiliser un nom scientifique qui n'est plus le nom de référence du taxon, ou dont le taxon a changé de taxon parent ou de rang.

Les Codes de nomenclature regroupent l'ensemble des règles régissant les noms scientifiques. Si ces règles s'appliquent sans ambiguïté jusqu'aux niveaux espèce et sous-espèce, les noms scientifiques associés aux rangs inférieurs (e.g. variété, cultivar) ne sont pas concernés. Or les listes de plantes cultivées dénotent parfois des taxons appartenant à ces niveaux inférieurs. En outre, il n'existe pas de règle sur le fait qu'un nom de plante cultivée dénote une sous-espèce, une variété, etc., et les noms vernaculaires retenus pour nommer les plantes sont parfois spécifiques à une région donnée, rendant l'alignement encore plus délicat. A ces difficultés s'ajoutent des problèmes de fiabilité. En raison de leur complexité, les règles de nomenclature ne sont pas toujours respectées. Par exemple le catalogue du GEVES donne l'autorité sans la date (e.g. "L." au lieu de "L. 1758") et ne respecte pas la casse. Qui plus est, les listes de plantes cultivées sont construites par agrégation mais ne précisent pas nécessairement leurs sources (publications scientifiques attestant de l'utilisation d'un nom), empêchant d'évaluer la confiance que l'on peut leur accorder.

Enfin, il existe des difficultés plus techniques, liées aux choix de modélisation des taxons, noms scientifiques et cultures. Par exemple, TAXREF-LD sépare strictement taxonomie et nomenclature. D'autres classifications ne font pas cette distinction, représentant à la fois des taxons et leurs noms. Certaines classifications ne représentent que des noms scientifiques, comme Catalog of Life. Les listes de plantes cultivées ne retiennent souvent qu'un nom scientifique en lieu et place d'un taxon, nom qui n'est peut-être plus le nom de référence du taxon. Ces variations de conception et de modélisation posent ainsi des questions récurrentes quant au choix des objets à aligner : aligne-t-on deux taxons, un taxon et un nom, une plante cultivée et un taxon etc. ?

4 Algorithme d'alignement mis en oeuvre

La section 3.3 a souligné les différences de modélisation existant entre les taxonomies biologiques et les listes de plantes cultivées, ainsi que l'écart entre les données représentées (noms vernaculaires, nom scientifique, taxon). Ces différences rendent difficile l'utilisation d'outils classiques d'alignement d'ontologies ou de liage d'entités. Aussi nous avons exploré puis combiné plusieurs méthodes.

Le code implémentant les méthodes décrites dans cette sec-

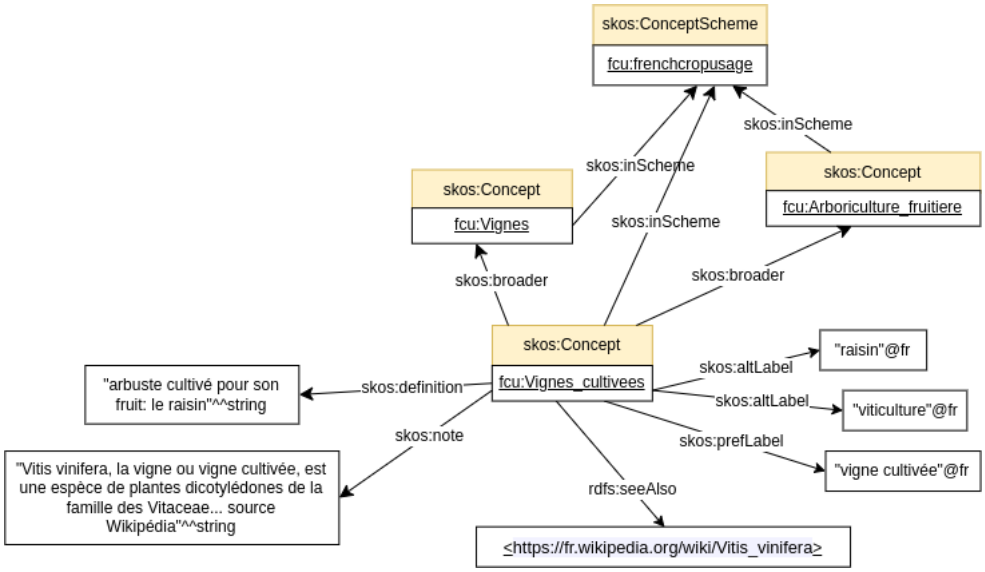

FIGURE 2 – Un extrait du thésaurus FCU présentant le concept de vigne cultivée.

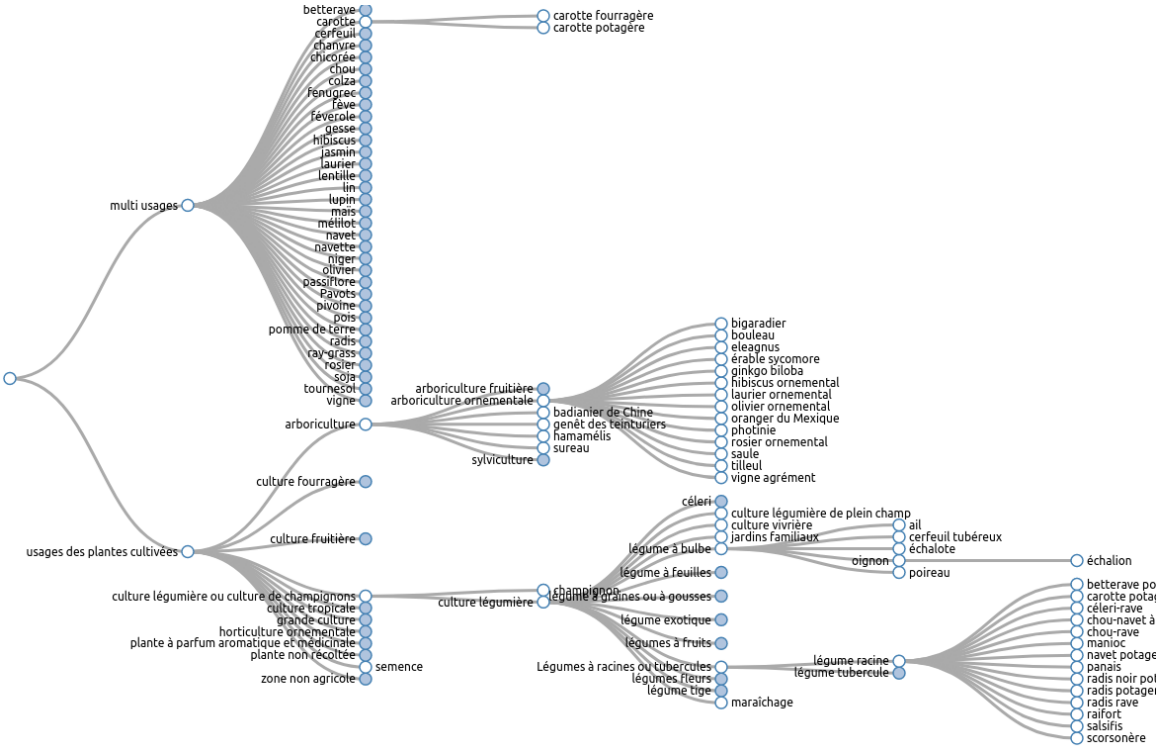

FIGURE 3 – Un exemple de visualisation du thésaurus avec l'outil SKOSPlay

tion ainsi que les données produites sont disponibles sous licence ouverte sur un dépôt public [10].

4.1 Méthodes d'alignement

Une première méthode consiste en un alignement direct entre FCU et TAXREF-LD basé sur la correspondance exacte, insensible à la casse, entre les noms d'usage des plantes cultivées de FCU (étiquettes préférées ou alternatives de concepts) et les noms vernaculaires de TAXREF-LD. Pour ce faire nous avons utilisé la version 2.2 de FCU et la version 15 de TAXREF-LD. En pratique, cette méthode donne des résultats médiocres en raison de la variabilité des noms vernaculaires retenus dans chaque source.

Une autre approche consiste à utiliser une source intermédiaire faisant la correspondance entre les noms vernaculaires des plantes cultivées et leurs noms scientifiques. La section 3 propose plusieurs référentiels faisant autorité, qui font l'objet d'une curation manuelle. Nous en avons retenu deux : la *Base de Données Mondiale* publiée par l'EPPO [4] que nous avons interrogée par son interface programmatique ; le *Catalogue officiel des espèces et variétés de plantes cultivées en France* publié par le GEVES [11] dont nous avons téléchargé les fichiers tabulaires depuis le site web du GEVES. Dans un premier temps, l'algorithme cherche à faire correspondre les noms d'usage des plantes de FCU avec les noms vernaculaires de ces deux référentiels. Dans le cas du GEVES, il s'agit d'une correspondance exacte, insensible à la casse. Dans le cas d'EPPO, il s'agit d'une correspondance approchée implémentée par l'API EPPO [12], toutefois la mesure de distance utilisée n'est pas documentée. L'algorithme retient le nom scientifique correspondant à chaque nom vernaculaire. Dans un deuxième temps, il cherche ce nom scientifique dans TAXREF-LD, puis il retient le taxon dont ce nom scientifique est soit le nom de référence soit un synonyme. La base EPPO et TAXREF-LD respectent strictement le code de nomenclature pour le nommage des noms scientifiques (sous la forme : "nom latin autorité, année", e.g. *"Prunus armeniaca L., 1753"*. La correspondance est donc une simple égalité insensible à la casse. En revanche, le catalogue du GEVES ne fournit pas l'année et ne respecte pas la casse (e.g. *"prunus armeniaca l."*). La correspondance se fait donc en cherchant des noms scientifiques de TAXREF-LD commençant par le nom issus du GEVES du GEVES (comparaison insensible à la casse).

Afin de permettre à des experts de valider les alignements, les résultats des trois méthodes sont conservés (alignement direct, via EPPO, via GEVES) et ordonnés par étiquette de FCU. Un score de confiance est attribué à chaque alignement candidat, calculé en fonction du nombre de méthodes (1, 2 ou 3) ayant proposé cet alignement. Le score peut donc valoir 1/3, 2/3 ou 1. Notons que le score 1 signifie simplement l'accord entre les trois méthodes, mais ne garantit pas sa justesse qui doit être vérifiée par un expert.

10. https://github.com/Wimmics/d2kab-alignments
11. https://www.geves.fr/catalogue/
12. Service /tools/names2codes : https://data.eppo.int/documentation/rest

4.2 Choix des entités à aligner

Dans les trois méthodes ci-dessus, on cherche à aligner les concepts de FCU avec des taxons de TAXREF-LD. Côté TAXREF-LD, on restreint les taxons candidats aux rangs espèce ou infra-spécifiques (sous-espèce, variété, etc.). En effet, le nom scientifique d'une plante cultivée se caractérise au moins par son espèce.

Côté FCU, on considère deux groupes de concepts de FCU à aligner. Dans le groupe *plantes cultivées*, on ne considère que les plantes de la branche "usages des plantes cultivées" et seulement celles des deux derniers niveaux de la hiérarchie (les feuilles ou leurs parents immédiats). Nous faisons donc l'hypothèse que l'unité d'alignement avec TAXREF-LD est un usage précis de plantes cultivées et non un regroupement de plantes (comme céréales). Dans le groupe *tous concepts*, on considère tous les concepts des branches "usages des plantes cultivées" et "multiusage" quel que soit leur niveau dans la hiérarchie. On ne fait donc aucune hypothèse sur l'unité d'alignement entre FCU et TAXREF-LD. Le groupe *plantes cultivées* est donc un sous-ensemble du groupe *tous concepts*.

5 Évaluation des alignements

Les résultats des méthodes d'alignement ont été évalués de deux manières.

5.1 Évaluation quantitative

Les statistiques données dans cette section ont été calculées par des requêtes SPARQL soumises depuis deux Jupyter Notebooks disponibles sur le dépôt du projet [13].

L'algorithme d'alignement a été exécuté pour les deux groupes de concepts décrits en section 4.2. Dans le groupe *plantes cultivées* qui contient 447 concepts, l'algorithme a proposé 651 alignements pour 300 de ces concepts (67% des concepts alignés) vers 579 taxons. Aucun alignement n'a été proposé pour 147 concepts (33%). Ces 147 concepts non alignés ont été évalués par un expert qui a indiqué que 118 concepts auraient dû être alignés car ils correspondent bien à des plantes cultivées et non à des groupes de plantes. Dans le groupe *tous concepts* qui considère 526 concepts, l'algorithme a proposé 710 alignements pour 337 concepts (64% des concepts alignés) vers 609 taxons. Aucun alignement n'a été proposé pour 189 concepts (36% des concepts non alignés). Le détail des nombres d'alignements proposés par méthode et par groupe est donné dans la table 1.

On remarque que la méthode d'alignement direct fournit de nombreux alignements mais est peu discriminante : elle génère en moyenne 2,39 alignements/concept pour 86% des concepts dans le groupe *plantes cultivées*, et 2,33 alignement/concept pour 77% des concepts dans le groupe *tous concepts*. A l'inverse, la méthode utilisant le catalogue du GEVES est plus discriminante - environ 1 alignement/concept - mais pour seulement 14% et 15% des concepts respectivement. La méthode utilisant EPPO semble la plus équilibrée : en moyenne 1,36 alignements/concept pour

13. Notebooks query-alignments.ipynb disponibles sur https://github.com/Wimmics/d2kab-alignments

TABLE 1 – Nombre d'alignements produits, et nombre de concepts et taxons impliqués dans ces alignements. La ligne "Total dédoubl." donne le total dans chaque groupe après suppression des alignements proposés par plusieurs méthodes.

Méthode	Nb. total d'alignements	Nb. de concepts FCU	Nb. de taxons
Groupe *plantes cultivées*			
Align. direct	385	161	369
via cat. GEVES	67	64	57
via BD EPPO	362	266	315
Total dédoubl.	**651**	**300**	**579**
Groupe *tous concepts*			
Align. direct	406	174	385
via cat. GEVES	84	81	70
via BD EPPO	404	300	336
Total dédoubl.	**710**	**337**	**609**

TABLE 2 – Nombre d'alignements proposés par 2 ou 3 méthodes à la fois.

Communs aux 3 méthodes	direct & GEVES	direct & EPPO	GEVES & EPPO
Groupe *plantes cultivées*			
15	17	115	46
Groupe *tous concepts*			
18	20	123	59

67% des concepts du groupe *plantes cultivées*, et 1,34 alignement/concept pour 64% des concepts du groupe *tous concepts*.

En outre, une analyse détaillée indique que les trois méthodes sont fortement complémentaires. En effet, quel que soit le groupe, environ 77% des alignements ne sont proposés que par une seule méthode (503 alignements dans le groupe *plantes cultivées*, 544 dans le groupe *tous concepts*). La table 2 montre que les trois méthodes ne s'accordent que sur 15 alignements (2,3%) dans le groupe *plantes cultivées*, et 18 alignements (2,5%) dans le groupe *tous concepts*. L'accord le plus fort apparaît entre les méthodes d'alignement direct et via EPPO avec seulement 115 alignements (17%), 123 alignements (17%) respectivement.

5.2 Évaluation qualitative

L'évaluation qualitative porte sur un sous-ensemble de plantes : la vigne, la carotte, les salades, la tomate. Pour chaque plante, deux experts ont évalué les couples (concept FCU, taxon TAXREF-LD) existants et détecté les couples manquants. Sur ce faible nombre d'alignements, les experts étaient majoritairement d'accord. Aucune avis contradictoire entre experts n'a été noté. La différence vient de couples manquants proposé par un des experts. Les alignements ont été qualifiés à l'aide de propriétés skos match. Ce choix est uniquement pragmatique et nous avons précisé la signification de ces propriétés de la manière suivante :

- `skos:exactMatch` signifie dans notre cas que le groupe de plantes représenté par le taxon est utilisé pour remplir cet usage en agriculture. Par exemple, la sous-espèce *"Vitis vinifera subsp. vinifera"* a pour usage "vigne cultivée".
- `skos:broadMatch` signifie que le groupe de plantes représenté par l'usage est inclus dans le groupe de plantes représenté par le taxon. Par exemple, les plantes qui ont pour usage "vigne cultivée" sont toutes de l'espèce *"Vitis vinifera"*.
- `skos:narrowMatch` signifie que le groupe de plantes représenté par l'usage inclut l'ensemble des plantes représenté par le taxon. Par exemple, les chicorées potagères incluent la variété *"Cichorium intybus var. sativum"*.
- `skos:closeMatch` signifie qu'il existe un lien entre le groupe de plantes représentées par l'usage et celui du taxon mais que la signification de ce lien est inconnue de l'expert.

Dans un futur proche nous définirons un ensemble de propriétés spécifiques à l'alignement entre un taxon biologique et un usage de plante en agriculture.

Le tableau 3 présente l'évaluation des alignements de l'ensemble des méthodes sur le groupe *plantes cultivées*. Pour le sous-ensemble de plantes considéré pour l'évaluation, la base EPPO a produit 15 alignements (dont 2 faux), et le catalogue du GEVES a produit 2 alignements. 9 alignements ont été trouvés en direct (dont 2 faux). 7 alignements sont communs à EPPO et en direct (dont 1 faux). 2 alignements sont communs au catalogue du GEVES et à EPPO.

Le tableau 4 présente les évaluations des alignements de l'ensemble des méthodes sur le groupe *tous concepts*. Sur ce groupe, la base EPPO a produit 17 alignements (dont 2 faux), et le catalogue du GEVES a produit 3 alignements. 10 alignements ont été trouvés en direct (dont 2 faux). 7 alignements sont communs à EPPO et en direct (dont 1 faux). 3 alignements sont communs au catalogue du GEVES et à EPPO.

Dans le cas de la vigne, il existe plusieurs alignements de type exact match entre un même concept FCU et plusieurs taxons. Cela signifie qu'une plante cultivée correspond à plusieurs espèces ou que certains taxons sont référencés plusieurs fois par des noms synonymes. Dans le cas des salades nous avons noté l'inverse, il existe plusieurs alignements de type exact match entre un même taxon et plusieurs concepts FCU. Cela signifie que la même espèce est utilisée pour différents usages.

EPPO est la source qui produit le plus d'alignements mais certains d'entre eux sont jugés erronés par les experts. Le catalogue du GEVES produit peu d'alignements mais ils sont tous justes. L'accord entre deux sources n'est pas un bon critère pour nettoyer les alignements étant donné qu'un des alignements jugés faux a été détecté par EPPO et en direct. Cette évaluation qualitative n'a pas identifié d'accord entre les 3 sources. Nous avons besoin de procéder à plus d'évaluation pour identifier si le catalogue du GEVES est bien la source de référence à utiliser. Nous aurons aussi besoin d'étudier pourquoi cette source, qui recense toutes les

TABLE 3 – Résultat de l'évaluation qualitative pour le groupe *plantes cultivées*

nom de cultures	nb alig. détectés	nb alig. exact	nb alig. broad	nb alig. narrow	nb alig. close	nb alig. faux	nb alig. manquants
salade	22	11	5	2	0	4	0
tomate	2	2	0	0	0	0	0
carotte	0	0	0	0	0	0	2
vigne	3	2	1	0	0	0	4

TABLE 4 – Résultat de l'évaluation qualitative pour le groupe *tous concepts*

nom de cultures	nb alig. détectés	nb alig. exact	nb alig. broad	nb alig. narrow	nb alig. close	nb alig. faux	nb alig. manquants
salade	24	11	6	2	0	5	0
tomate	2	2	0	0	0	0	0
carotte	3	1	2	0	0	0	0
vigne	3	2	1	0	0	0	4

variétés cultivées, produit si peu d'alignements.

6 Conclusion et Perspectives

Les méthodes d'alignements automatiques que nous avons produites ne donnent pas entièrement satisfaction. Plusieurs causes peuvent être identifiées : la variabilité des noms vernaculaires qui ne suivent aucune convention, le manque de couverture en noms vernaculaires des taxonomies biologiques, et la simplicité des techniques de comparaison mises en oeuvre actuellement dans notre algorithme. Concernant ce dernier point, nous envisageons d'améliorer l'algorithme en utilisant des mesures de similarité plus adaptées. Par exemple, en utilisant une distance de Levenstein pour la correspondance entre noms vernaculaires, ou les règles métier d'alignement de noms scientifiques implémentées pour aligner TAXREF-LD avec d'autres référentiels taxonomiques (voir section 2.1.2). Ainsi, les alignements produits ont besoin d'être améliorés et nettoyés par des experts. Le catalogue du GEVES est la source qui a produit les alignements les plus fiables (validés par les experts) mais en nombre insuffisant.

Nos travaux montrent que les alignements entre des classifications agricoles et des taxonomies biologiques sont plus complexes que de simples correspondances 1:1. Nous avons besoin d'exprimer le fait qu'une plante cultivée correspond à plusieurs espèces, voire à un ensemble d'espèces et de sous-espèces, et inversement. Il s'agit donc d'alignements N:N pouvant impliquer différents types de relations.

Nous avons étudié à ce jour deux schémas permettant de stocker les alignements : le langage EDOAL et le schéma "A Simple Standard for Sharing Ontology Mappings" (SSSOM). EDOAL permet de représenter les alignements complexes [5] mais est difficile d'accès pour les non-spécialistes de ce langage. Il faudra donc réfléchir à des modalités de validation des alignements pour les agronomes. SSSOM est un standard en cours d'évolution qui pour le moment se limite aux alignements simples 1:1 [10]. La prochaine étape de notre travail est de définir un schéma permettant d'exprimer nos alignements automatiques et leurs validations par des experts.

Références

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Semantic Web and Information Systems*, 5(3) :1–22, 2009.

[2] Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. The AGROVOC linked dataset. *Semantic Web - Interoperability, Usability, Applicability*, 4(3) :341–348, 2013. http://content.iospress.com/articles/semantic-web/sw106.

[3] Arthur D Chapman. Numbers of living species in Australia and the world. https://www.awe.gov.au/science-research/abrs/publications/other/numbers-living-species, Canberra, Australia, september 2009.

[4] EPPO. EPPO Global Database (available online). https://gd.eppo.int, 2022.

[5] Jérôme Euzenat, François Scharffe, and Antoine Zimmermann. Expressive alignment language and implementation. Contract, June 2007. https://hal.inria.fr/hal-00822892.

[6] Daniel Faria, Beatriz Lima, Marta Contreiras Silva, Francisco M Couto, and Catia Pesquita. AML and AMLC results for OAEI 2021. In *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference*, pages 131–136, 2021.

[7] Olivier Gargominy, Sandrine Tercerie, C Régnier, T Ramage, P Dupont, P Daszkiewicz, and L Pon-

cet. TAXREF v15, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion. https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref, 2021.

[8] Clement Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, and Pierre Larmande. AgroPortal : a vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144 :126–143, January 2018.

[9] Guillaume Lecointre and Hervé Le Guyader. *Classification phylogénétique du vivant : tome 2*. Belin, 05 2017.

[10] Nicolas Matentzoglu, James P. Balhoff, Susan M. Bello, Chris Bizon, Matthew Brush, Tiffany J. Callahan, Christopher G Chute, William D. Duncan, Chris T. Evelo, Davera Gabriel, John Graybeal, Alasdair Gray, Benjamin M. Gyori, Melissa Haendel, Henriette Harmse, Nomi L. Harris, Ian Harrow, Harshad Hegde, Amelia L. Hoyt, Charles T. Hoyt, Dazhi Jiao, Ernesto Jiménez-Ruiz, Simon Jupp, Hyeongsik Kim, Sebastian Koehler, Thomas Liener, Qinqin Long, James Malone, James A. McLaughlin, Julie A. McMurry, Sierra Moxon, Monica C. Munoz-Torres, David Osumi-Sutherland, James A. Overton, Bjoern Peters, Tim Putman, Núria Queralt-Rosinach, Kent Shefchek, Harold Solbrig, Anne Thessen, Tania Tudorache, Nicole Vasilevsky, Alex H. Wagner, and Christopher J. Mungall. A Simple Standard for Sharing Ontological Mappings (SSSOM). dec 2021. http://arxiv.org/abs/2112.07051.

[11] Franck Michel, Olivier Gargominy, Sandrine Tercerie, and Catherine Faron-Zucker. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. In *Proceedings of the ISWC2017 workshop on Semantics for Biodiversity (S4BioDiv)*, volume 1933, Vienna, Austria, 2017. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1933/paper-3.pdf.

[12] Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). https://www.w3.org/TR/skos-reference/skos-xl.html, 2009.

[13] Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. http://www.w3.org/TR/skos-reference/, 2009.

[14] Dan H. Nicolson. A history of botanical nomenclature. *Annals of the Missouri Botanical Garden*, 78(1) :33–56, 1991. http://www.jstor.org/stable/2399589.

[15] Mina Abd Nikooie Pour, Alsayed Algergawy, Florence Amardeilh, Reihaneh Amini, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Pascal Hitzler, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, Engy Nasr, Heiko Paulheim, Catia Pesquita, Jan Portisch, Catherine Roussey, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Cássia Trojahn, Jana Vatascinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the Ontology Alignment Evaluation Initiative 2021. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021*, volume 3063 of *CEUR Workshop Proceedings*, pages 62–108. CEUR Workshop Proceedings, 2021. http://ceur-ws.org/Vol-3063/oaei21_paper0.pdf.

[16] Y. Roskov, G. Ower, T. Orrell, D. Nicolson, N. Bailly, and et al. (eds) Kirk, P. M. Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. http://www.catalogueoflife.org/annual-checklist/2019/, 2019.

[17] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. NCBI Taxonomy : a comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation*, 2020 :21, 2020. https://doi.org/10.1093/database/baaa062.

[18] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk - A Link Discovery Framework for the Web of Data. In *2nd Workshop about Linked Data on the Web*, volume 538, Madrid, Spain, 2009. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf.

Remerciements

Ce travail a été réalisé avec le soutien du projet "Des Données aux Connaissances en Agronomie et Biodiversité (D2KAB–www.d2kab.org) financé par l'Agence Nationale de la Recherche (ANR-18-CE23-0017) et du projet "Partage de Connaissances" (PACON) du programme transverse MetaBio financé par INRAE ; ainsi qu'avec l'aide de l'entreprise Elzeard https://www.elzeard.co. Nous remercions également les membres de la tâche 4.3 du projet D2KAB : Sophie Aubin, Stephan Bernard, Sonia Bravo, Anna Chepaikina, Baptiste Darnala, Matthieu Hirschy, Clement Jonquet et Nadia Yacoubi. Un remerciement particulier pour Juliette Raphel, ingénieur agronome chez Elzeard et les deux agronomes spécialistes de la vigne qui ont participé à l'évaluation : Thierry Lacombe de SupAgro et Olivier Yobregat de l'Institut Français de la Vigne et du Vin (IFV).