

ChampySeed Exploration des données et rapport de Dataviz

Benoit Maxence Bastien

September 3, 2021

0.1 Mise en contexte

Nous avons le sujet sur les champignons et notre premier rapport évoquait les différentes bases de données que l'on avait à notre disposition. Il fallait faire un choix. C'est l'objet de la première partie de ce rapport. Ensuite nous explicitons nos démarches pour découvrir la dataset.

1 Mushroom Observer

La fiche du projet proposait 2 sources d'informations:

- Le site Mushroom Observer (MO) qui est un site participatif où les contributeurs postent une photo, indiquent le lieu d'observation et proposent un nom de champignon validé par la communauté après consensus. Les contributions proviennent presque uniquement des US.
- Un Github Mushroom Observer Dataset dont les photos (sur un Dropbox) proviennent du site précédent. Des métadonnées sont disponibles. Après examen du code il s'avère que ces informations proviennent d'un autre site GBIF.org

Le site MO propose une API mais celle-ci s'avère très contraignante en raison des limites sur le nombre de requêtes mais aussi la nécessité de gérer la pagination. (limite de 100 résultats par requête) Le site souffre aussi de performances dégradées en ce moment. Finalement le fichier le plus pertinent sur le site MO est un fichier contenant 3 types d'infos (figure 1):

1. URL d'une photo
2. Nom du champignon
3. Date de l'observation et copyrights

Il est alors possible d'utiliser la colonne "Name" pour chercher les informations sur GBIF.org. Deux méthodes ont été explorées:

1. En utilisant l'API de gbif.org pour ajouter les métadonnées ligne par ligne dans le fichier MO

E3	A	B	C	D	E
1	image	name	created	license	rightsHolder
2	images/640/2.jpg	Xylaria magnolia	2004-07-17	http://creativecommons.org/licenses/by-sa/3.0/	Nathan Wilson
3	images/640/16.jpg	Volvopluteus glo	2005-01-07	http://creativecommons.org/licenses/by-sa/3.0/	Nathan Wilson
4	images/640/26.jpg	Panellus stipticus	2004-11-26	http://creativecommons.org/licenses/by-sa/3.0/	Nathan Wilson
5	images/640/36.jpg	Sutorius eximius	2004-07-14	http://creativecommons.org/licenses/by-sa/3.0/	Nathan Wilson

Figure 1: illustration des données de MO

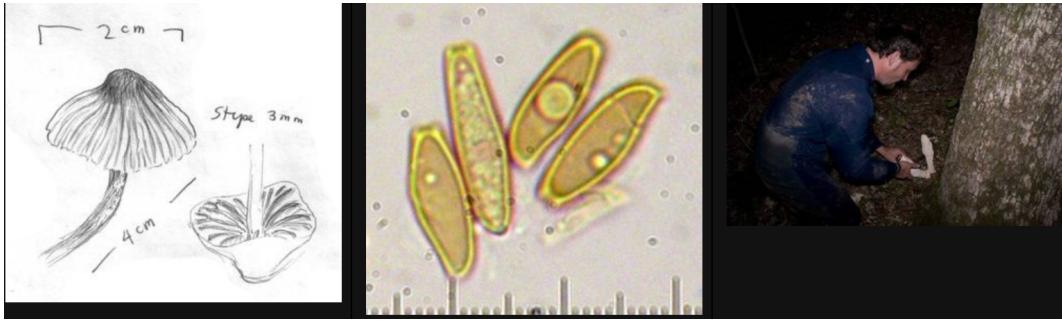


Figure 2: exemples 1,2,3



Figure 3: exemples 4,5

2. La récupération d'un Dataset Herbarium GB, University of Gothenburg (gbif.org) puis un merge avec le fichier MO

Nous avons ensuite visualisé quelques photos : Un examen partiel des images contenues sur le site a mis en lumière la qualité très médiocre des photos. Voici quelques exemples d'images plus ou moins faciles à détecter et éliminer (figure 2 et 3) :

1. Dessins
2. Coupes microscopiques
3. Avec le cueilleur
4. Paysage d'un site de cueillette
5. Plan trop large

De plus il y a très peu de photos par champignon, seulement 1 photo pour la grande majorité, 6 pour quelques autres et la seule exception est 70 photos pour l'Amanita.

En conclusion, le contenu du site MO s'est révélé très décevant mais nous a permis de découvrir un site plus qualitatif et professionnel, gbif.org. Les photos de MO feront toutefois une bonne source d'images pour nos démos dans la mesure où nous pourrons cette fois sélectionner des photos pertinentes.

2 GBIF.org

Le GBIF - Global Biodiversity Information Facility (Système mondial d'information sur la biodiversité) - est un réseau international et une infrastructure de données financés par les gouvernements mondiaux ayant pour but de fournir à tous et partout un accès libre aux données sur toutes les formes de vie sur Terre. Il est coordonné via son Secrétariat de Copenhagen.

Dans un premier temps, nous sommes partis du dataset suivant : Herbarium GB, University of Gothenburg (gbif.org) pour n'utiliser que le sous-ensemble "fungi" pour merger les métadonnées GBIF avec les images de MO. Il s'est avéré que cette base n'était pas très complète sur les Agaricomycètes.

Finalement nous avons fait une recherche sur l'ensemble des occurrences du site avec les 2 filtres suivants:

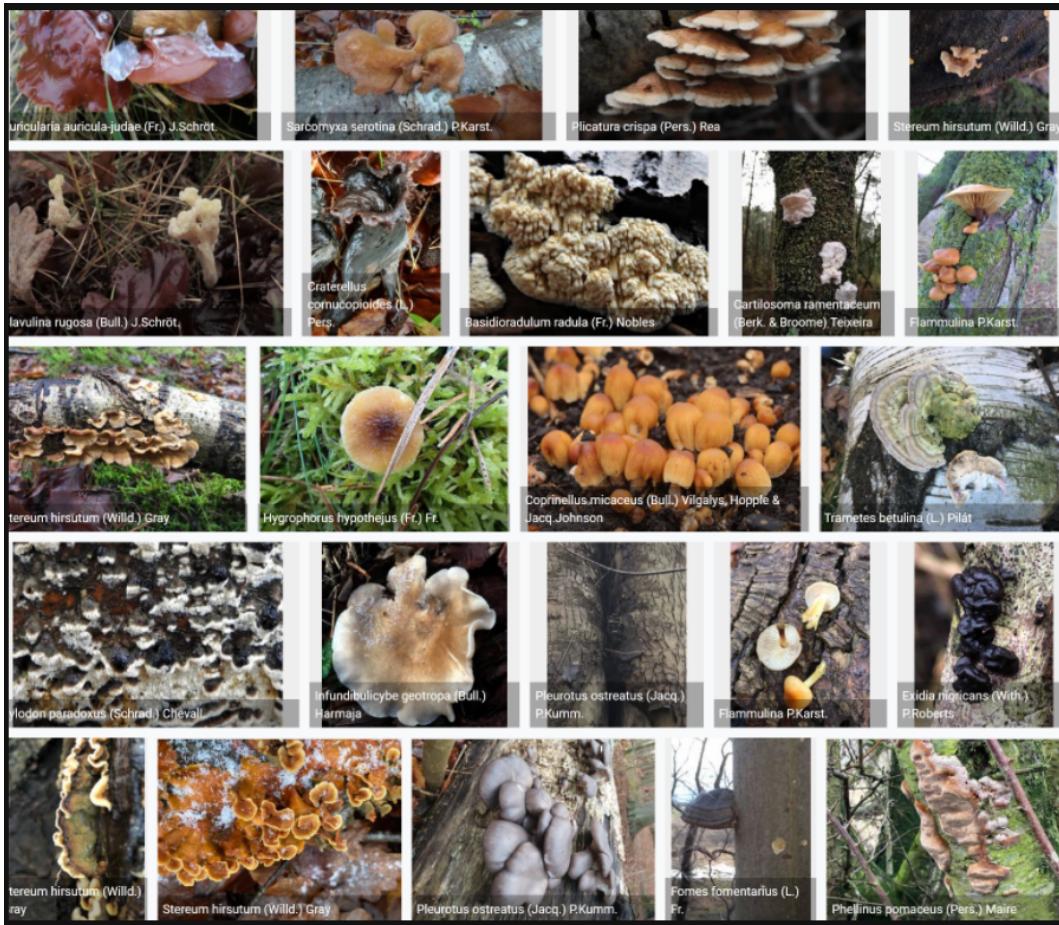


Figure 4: GBIF exemple de photos

1. Scientific Name = “Agaricomycetes”
2. Media Type : Image (1 355 857 photos)

Ceci nous a permis de télécharger un zip (format Darwin Core Archive) contenant notamment les fichiers :

1. Occurrences.txt (Inclut les métadonnées)
2. Multimedia.txt (Inclut les url des photos)

Le chargement en utilisant la commande `read csv` habituelle s'est avéré compliqué à cause de certains caractères parasites mais nous avons pu trouver un module Python-dwca-reader dédié à l'exploitation des fichiers Darwin Core Archive.

Nous avons alors pu explorer ces fichiers et sélectionner les variables pertinentes pour notre projet. Une fois ces deux datasets épurés, ils ont été mergés suivant la variable commune dans les 2 fichiers (ID et CoreID)

Il nous semble que ce genre d'échantillon de photographie est bien plus profitable pour l'apprentissage.(figure 4)

Il est possible de se faire une première idée des photos disponibles sur GBIF.org directement depuis ce lien : https://www.gbif.org/occurrence/gallery?taxon_key=186

Nous avons également téléchargé aléatoirement des photos en utilisant les url du dataset. Les photos sont clairement plus qualitatives que MO, il reste néanmoins des photos que nous devrons écarter comme sur ces exemples(figure 5):

3 Exploration des datasets

Fichiers GBIF utilisés:

Nom de fichier	Nombre de ligne	Nombre de variable	Taille disque
occurrences.txt	1354629	251	1717 Mo
multimedia.txt	2505448	15	533 Mo



Figure 5: Exemples encore à écarter

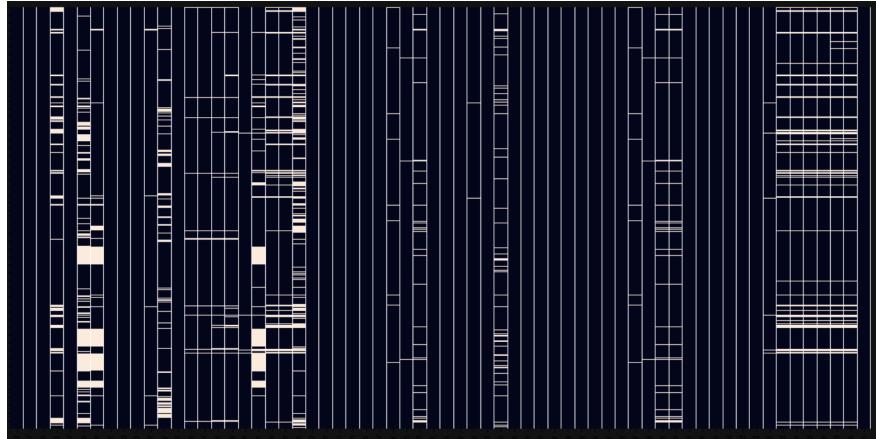


Figure 6: Aperçu des valeurs manquantes du tableau

Variables conservées:

Variable	Pourquoi la garder
countryCode	En cas d'affichage par pays
decimalLatitude	explicite
decimalLongitude	explicite
hasCoordinate	booléenne explicite

Variables conservées dans l'optique de l'apprentissage du modèle:

Fungi	Le royaume des champi	En cas d'erreur
phylum	Division en deux classes	
class		La classe probable a garder
family	Familles	Variable candidate
genus	Type (lamelle, autre)	Variable candidate
verbatimScientificName	Nom scientifique	Nom certain du champignon
species	Race du champignon	
identifier_y	Lien d'images	

La taxinomie des champignons étant généreuse nous n'avons pas encore cerné les différentes nécessité de toutes ces espèces sous espèces.

En résumé, nous gardons les infos liées à la taxonomie, à la localisation et bien entendu l'url de la photo en vue du téléchargement.

Le Dataset final contient très peu de données manquantes comme on peut le voir sur la visualisation suivante:(Figure 6 et 7)

3.1 Champignons vénéneux (infos Bonus)

De même que les coordonnées de localisations seront affichées une fois le champignon identifié, nous gardons la possibilité d'ajouter s'il est vénéneux. Nous pourrons l'ajouter à la fin si le temps le permet dans la mesure où ça n'impactera pas l'apprentissage.

Voici un graphe du nombre de photos référencées pour la comestibilité.
(environ 73% du total des familles)
(figure 9)

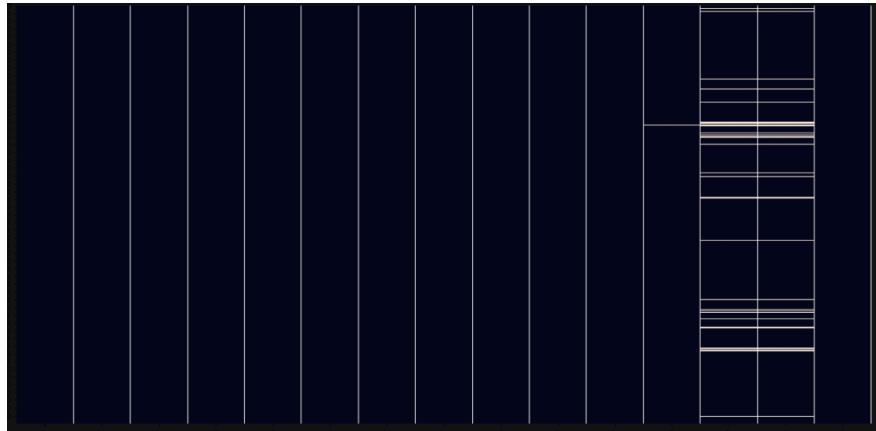


Figure 7: Aperçu des valeurs manquantes du tableau final

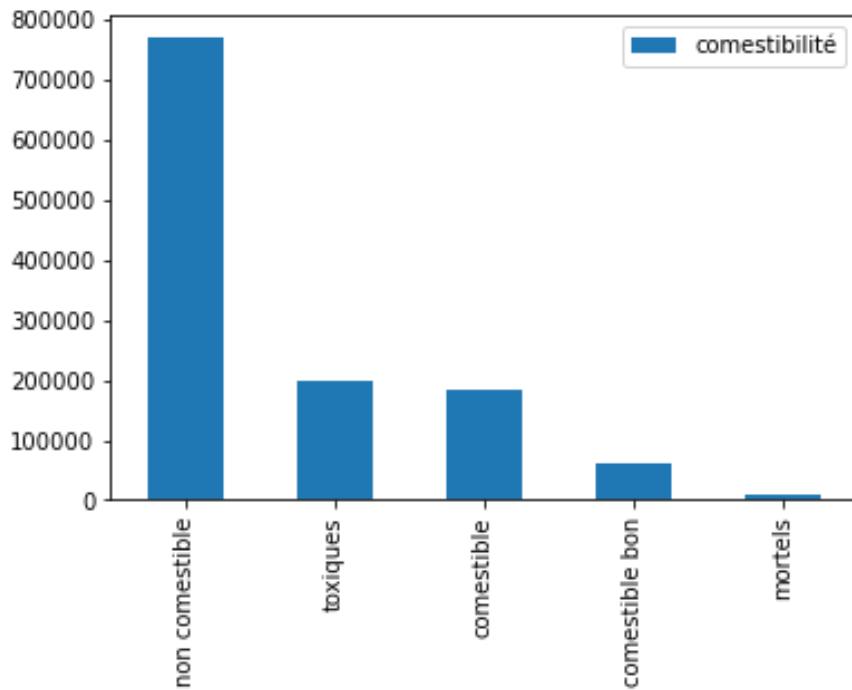


Figure 8: Nombre de photos par type de comestibilité

Nous avons aussi un graphe détaillant le nombre de photos pour chaque type de comestibilité possible.
(Après avoir supprimé ceux sans label correspondant)
(figure 8)

4 DataViz

4.1 Données géographiques (infos Bonus)

Contrairement à MO, plus de 90% des occurrences du dataset de gbif contient des coordonnées géographiques. De plus, on trouve des coordonnées sur tout le globe comme le montre la visualisation Bokeh (figure 8). Pour ce graphique, il a fallu convertir les coordonnées de Longitude/Latitude (WGS84) en coordonnées “web Mercator”.

4.2 Taxonomie

Voici un tableau récapitulatif : la classe Agaricomycetes contient plusieurs ordres qui contiennent plusieurs sous-familles qui contiennent elles-mêmes plusieurs genres. Cette classe contenant encore énormément de contenu, il n'est pas encore impossible que notre recherche se spécialise un peu plus par la suite.

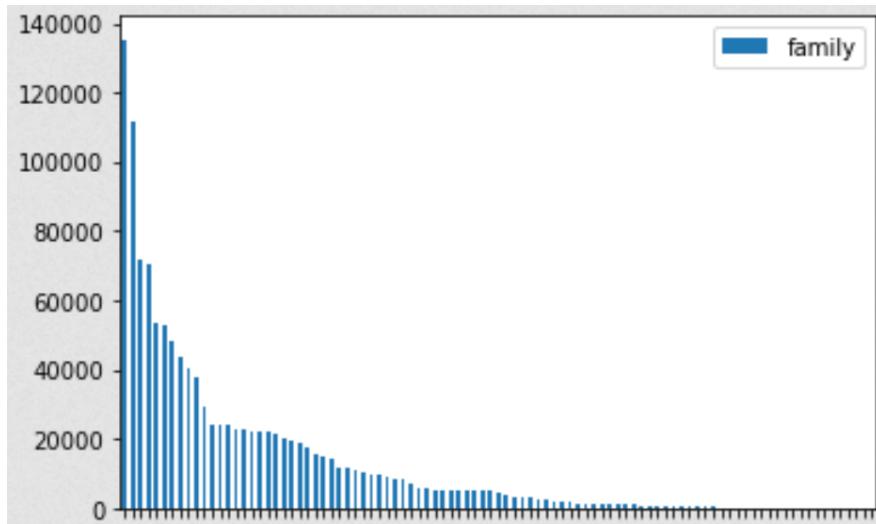


Figure 9: Nombre de photos par famille référencées pour la comestibilité

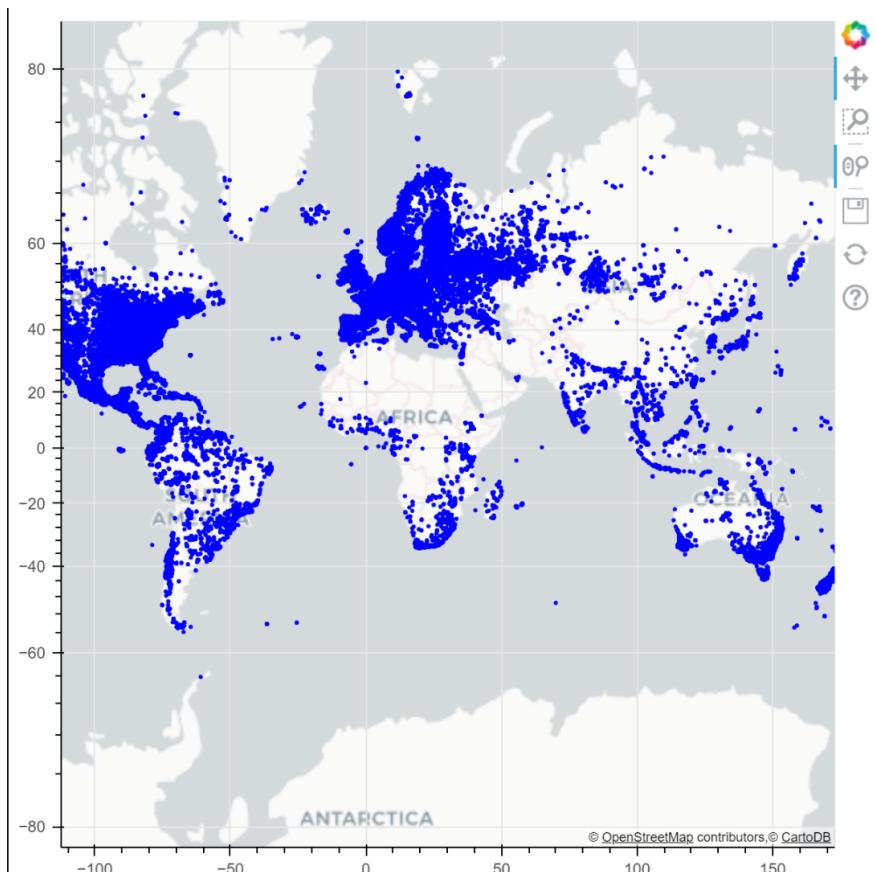


Figure 10: Coordonnées géographiques des données

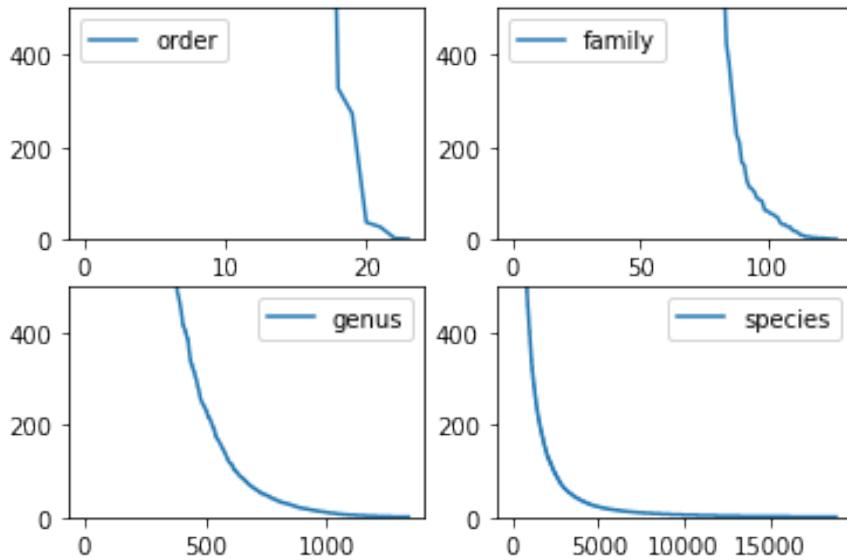


Figure 11: Nombre de photos disponibles selon les éléments de chaque classes(axes des x numéro de chaque sous classe de la classe)

Kingdom	Phylum	Class	Order	Family	Genus
Fungi	Basidiomycota	Agaricomycetes			

Vu la taxonomie nous avons essayé de voir quel était le nombre de photos pour chaque "genre" de champignons. Si au bout d'un moment nous apercevions qu' il n'y avait plus assez de nombre de photos pour les différents "genres", la question se posait alors de garder ou non une classification aussi spécifique. La figure 9 montre le nombre de photos disponible pour chaque éléments dans la classe identifiée. Les différentes classes du genre sont triées selon le nombre de photos (décroissance de chaque courbe). Il est normal de voir que le nombre de sous classes augmente très vite selon la profondeur de la taxonomie

5 Choix de notre target

Nous n'avons pas encore parfaitement choisi notre target. Nous devons évaluer notre capacité à faire le but premier qui serait la reconnaissance d'un champignon complet, soit son nom scientifique.(verbatimScientificName) Si nous devions abandonner cette voie nous avons l'options de remonter dans la taxonomie afin de reconnaître plus grossièrement le champignon:

- Reconnaître par Species
- Reconnaître par Genus
- Reconnaître par Family

Nous envisageons aussi de se restreindre à un nombre minimal de photos par feuille de notre arbre. Cela aurait deux mérites principaux:

- Obtenir un dataset plus équilibré
- réduire la taille de notre dataset

6 Conclusion

Nous avons maintenant un dataset riche et de qualité qui nous permet de choisir quel sera la target pour notre projet. Il nous manque néanmoins une information sur **le nombre de photos raisonnables** à utiliser dans le temps imparti pour notre projet.

MycoDB : Fiche de Simocybe sumptuosa - Profil 1 – Microsoft Edge

<https://www.mycodb.fr/fiche.php?genre=Simocybe&espece=sumptuosa&source=search>

- **Division - Classe - Ordre - Famille**
Basidiomycota / Agaricomycetes / Agaricales / Crustaceae
- **Synonymes** [IPNI](#) [Monograph](#)
 - Naucoria centunculus f. luxurians* Romagnesi (1944) [1942], Bulletin de la Société mycologique de France, 59(3-4), p. 129. 149
 - Agrocybe centunculus* var. *luxurians*(Romagnesi) Romagnesi (1950). Bulletin de la Société mycologique de France, 66, p. 348
 - Naucoria sumptuosa* P.D. Orton (1960). Transactions of the British mycological Society, 43(2), p. 324 (Basionyme)
 - Simocybe sumptuosa* (P.D. Orton) Singer (1962) [1961]. Sydowia : Annales mycologici, editi in notitiam scientiarum mycologicas universali, series II, 15(1-6), p. 74 (nom actuel)
 - Agrocybe sumptuosa* (P.D. Orton) Romagnesi (1963) [1962]. Bulletin de la Société mycologique de France, 78(4), p. 348
 - Ramcole sumptuosa* (P.D. Orton) Watling (1989) [1988]. Notes from the royal botanic Garden, Edinburgh, 45(3), p. 556
- **Chapeau**
 - **Couleur :** Brun, Crème
 - **Description :** Diamètre 10-40 (50) mm, hémisphérique à convexe au début, puis convexe aplati à étalé, souvent un peu déprimé au centre ; surface finement granuleuse à veloutée et mate, hygrophane, brun foncé ou brun sépia par temps humide, aussi brun olivâtre par le sec. Marge longtemps inférieure, lisse et aiguë.
- **Lames**
 - **Couleur :** Brun, Crème
 - **Description :** Au début brun beige, puis brun foncé, larges, nettement adnées, arêtes plus pâles que les faces et finement ciliées.
- **Chair**
 - Brun clair, mince. Odeur agréable, saveur douce.
- **Stipe**
 - **Couleur :** Brun, Crème
 - **Anneau :** Non
 - **Description :** 15-40 (50) x 1,5-4 mm, cylindrique, souvent arqué, parfois un peu épaissi à la base, flexible, plein au début, tubuleux avec l'âge, surface pruineuse veloutée dans le jeune âge sur toute la hauteur, ensuite glabrescente, longitudinalement fibrillaire, pruineuse au sommet, brun ocreâtre à brun olivâtre, brun plus foncé vers la base.
- **Spores**
 - Ellipsoïdales à phaséoliformes, faiblement verrueuses, jaune pâle, à parois épaisses, 7-9,5 x 4,5-5,5 µm. Sporée brun olivâtre. Basides clavées à ventrues, tétrasporiques, 22-30 x 7-10 µm, bouclées. Cheliocystides sinuées, clavées à capitées, 45-75 x 7-15 µm. Hyphes de la cuticule, couchées, parallèles, larges de 4-17 µm, pigmentées et partiellement incrustées de jaune; cloisons bouclées; nombreuses pleurocystides émergentes de 40-75 x 9-15 µm, rangées en structure presque palissadique au centre du chapeau. Caulocystides au sommet du pied, en partie pigmentées et incrustées de jaune.
- **Ecologie**
 - Sur bois dégradé de feuillus, troncs, souches, branches tombées, mais aussi sur parties mortes d'arbres encore vivants tels que *Fagus*, *Fraxinus*, etc.
 - De l'étage collinien à submontagnard.
 - Grégaires à fascicules, rarement isolés. Assez rare. Été-automne.
- **Comestibilité**
 - Sans intérêt

Figure 12: Exemple d'une fiche du site www.mycodb.fr

6.1 Questions et réflexions

6.1.1 Récupération des photos

Nous téléchargerons les photos à partir des url avec un script Python. Nous avons procédé à quelques tests pour récupérer des photos aléatoirement dans la base et également pour estimer la taille nécessaire en local sur nos disques durs. Nous avons constaté une taille moyenne de 1Mo par photo, ce qui veut dire environ 2,5To au total ! Il va de soi que c'est très gros mais il n'est pas nécessaire de garder les photos dans leur taille originale. Nous devrons décider de la taille suffisante en local (256x256 ?)

6.1.2 Intégration de la comestibilité (pas encore introduit)

Découverte d'une base de données, le site www.mycodb.fr. Ce site collaboratif se concentre uniquement sur les champignons. Ce qui se révèle très intéressant, c'est la fiche détaillée sur chaque champignon. Exemple d'une fiche (figure 12)

L'information qui nous intéresse ici est de pouvoir connaître la comestibilité d'un champignon. En récupérant sous forme de liste (création d'un csv) chaque champignon accompagné de sa donnée de comestibilité, il a ensuite été possible de créer une nouvelle colonne à notre tableau principal, la comestibilité avec les modalités : non comestible, comestible, comestible bon, toxique, mortel.

Deux choix à faire : ne garder que les champignons référencés par mycodb ou garder le tableau d'origine et inscrire "non renseigné". Dans ce premier cas, nous garderons 1,2M d'images sur 2,3M.

6.1.3 Récupération des caractéristiques de chaque champignon

Avec le site MycoDB, il devient possible de récupérer sous forme de tableau les caractéristiques de chaque champignon (lamelle, chapeau, anneau, chair etc...) pour trouver d'autres targets (dans le but d'avoir encore plus de photos par modalité)

Réduire le nombre d'images en réduisant le nombre de champignons non comestibles : ne garder que ceux que les gens sont susceptibles de cueillir.