Import de médias et textes dans TaxHub - Format v1.0

Processus d'importation

Ce document est en cours de travail. Les informations qu'il contient peuvent donc être amenées à changer à tout moment !

Transmission des fichiers

Les fichiers seront transmis dans un fichier d'archive au format ZIP. Le nom du fichier devra être en minuscule et contenir plusieurs parties séparées par des underscores ("_"). Les parties du fichier seront les suivantes :

1. date au format ISO 8601 : 2020-08-26
2. sujet : sinp
3. abréviation de la région concernée : paca / aura
4. abréviation de l'organisme fournisseur : cbna / cbnmed / cbnmc / cenpaca ...
5. type de données : media / text
6. extension : .zip

Exemple : *2020-08-26_sinp_paca_cbna_media.zip*

L'archive devra contenir le fichier suivant :

- ***meta_archive.ini*** : fichier contenant les métadonnées sur le fournisseur de l'archive et la version du format d'échange utilisée.

L'archive pour les données de type *media* devra contenir le fichiers suivant :

- ***media.csv*** : ficher contenant les informations sur les médias (images, pdf...) à lier aux taxons.

L'archive pour les données de type *text* devra contenir le fichiers suivant :

- ***text.csv*** : fichier contenant les informations sur les textes à lier aux taxons et concernant un *attribut*.

L'archive pour les données de type *text* pourra contenir le fichiers suivant :

- *attribut.csv* : fichier contenant les informations des *attributs* d'un *thème*.
- *theme.csv* : fichier contenant les informations d'un *thème*.

Format du fichier des métadonnées de l'archive "meta_archive.ini"

Ce fichier au format INI a pour objectif de fournir les informations sur l'origine des autres fichiers fournis dans l'archive.
Il concerne le créateur de l'archive.

Ce fichier devra:

- être encodé en **UTF-8**
- être nommé (en minuscule) : ***meta_archive.ini***

Les règles à respecter pour ce format INI sont le suivantes :

- une ligne peut contenir soit un commentaire débutant par le caractère **#** ou une entrée ***clé / valeur***
- la clé doit être séparé de sa valeur par un **=**
- les clés doivent être en minuscule et utilisé l'underscore (_) comme séparateur de mots
- des espaces peuvent encadrer les clés et valeurs (ils seront supprimés).
- Si la valeur contient plusieurs lignes, encadrée là par des guillemets doubles (") et indenter les ligne supplémentaires.

Format (en gras les champs obligatoires) :

- **format_version** [VARCHAR(8)] : version du format d'échange utilisé pour les fichiers à importer.
- export_date [DATE(YYYY-MM-DD HH:MM)] : date et heure de l'export des observations de la synthèse hors de la base d'origine.
- taxref_version [VARCHAR(8)] : version de TaxRef utilisée lors de la génération de l'archive.
- habref_version [VARCHAR(8)] : version de HabRef utilisée lors de la génération de l'archive.
- **editor** [VARCHAR(100)] : nom de l'organisme créateur de l'archive.
- **contact** [VARCHAR(100)] : infos sur la personne ayant créé l'archive. Format : `NOM Prénom <email>`.
- notes [TEXT] : remarques divers sur les fichiers de l'archive.

Exemple :

```
format_version = 1.0
export_date = 2020-08-27 10:15
taxref_version = 13
editor = Conservatoire Botanique National Alpin
contact = jp.milcent@cbn-alpin.fr
notes = "Données de test.
    À utiliser seulement lors de la phase de conception."
```

Format des fichiers d'import

Pour importer les données, nous utiliserons des fichiers CSV associé à la commande *COPY*. Ces fichiers CSV devront :

- être encodée en **UTF-8**
- avoir un nom au singulier, en minuscules et avec des underscores comme séparateur de mots.

- avoir l'extension *.csv*
- avoir un des noms suivant : *media.csv*, *text.csv*, *attribut.csv*, *theme.csv*

Le format CSV (en réalité plutôt TSV) qu'ils contiendront devra respecter les règles suivantes :

- utiliser une **tabulation** comme caractère de séparation des champs
- posséder une **première ligne d'entête** indiquant les noms des champs
- utiliser les caractères **\N** pour indiquer une valeur nulle (*NULL*) pour un champ
- si nécessaire utiliser le caractère **guillemet** (") pour préfixer et suffixer une valeur de champ
- si nécessaire utiliser **deux guillemets** successifs ("") pour échapper le caractère guillemet dans une valeur de champ préfixé et suffixé par des guillemets.

Il faut vous assurer d'avoir supprimé, remplacé ou protégé les caractères suivant dans les valeurs des champs :

- les caractères **anti-slash** (\) doivent être supprimé
- les caractères **tabulation** (Tab, ASCII 9) doivent être absolument supprimé du contenu des champs ou remplacé par `\t`
- les caractères **fin de ligne** (LF, Newline, ASCII 10) sont à supprimer ou à remplacer par `\n`
- les caractères **retour chariot** (CR, Carriage return, ASCII 13) sont à supprimer ou à remplacer par `\r`
- les caractères **tabulation verticale** (Vertical tab, ASCII 11) sont à supprimer ou à remplacer par `\v`

Format MEDIA d'import

- But : Permet de transmettre les informations associé à un média (images, pdf...).
- Table GeoNature : "*taxonomie.t_medias*".

Description du format MEDIA

Pour chaque ligne : `nom_du_champ [format du champ] (=`*`nom_champ_table_geonature`*`) : description du champ.`. Les champs **en gras** sont obligatoires.

- **cd_ref** [INT(4)] (=*cd_ref*) : correspond au champ "*cd_ref*" de TaxRef.
- **title** [VARCHAR(255)] (=*titre*) : titre court du média.
- **url** [VARCHAR(255)] (=*url*): URL qui servira à récupérer le document.
- author [VARCHAR(1000)] (=*auteur*) : liste des auteurs du média. Voir le détail du format à plat des infos sur une personne.
- description [TEXT] (=*desc_media*): description détaillé du média.
- date [TEXT] (=*date_media*) : date de création du média.
- source [VARCHAR(25)] (=*source*) : acronyme/abréviation de la source du média.
- licence [VARCHAR(100)] (=*licence*) : acronyme/abréviation standard de la licence du média. Privilégié les licences Creative Commons.
- **meta_change_date** [DATE YYYY-MM-DD HH:MM:SS] : date et heure du dernier changement effectué sur l'enregistrement du media.
- **meta_last_action** [CHAR(1)] : permet d'identifier les lignes ajoutées depuis le dernier import ("*I*"), modifiées ("*U*") ou supprimées ("*D*").

Notes

Au 2023-05-16 l'intégration des données se base sur le champ **url** pour réaliser un UPSERT (ajout/modification) des enregistrements transmis. Il n'y a donc pas de possibilité de supprimer les médias précédemment ajouté. À l'avenir, nous nous baserons surement sur les champs **meta_change_date** et **meta_last_action** pour réaliser les suppressions.