Passerelle - Development #40704

connecteur csv : import xslx impossible

12 mars 2020 19:31 - Thomas Noël

Statut:	Fermé	**Début:**	12 mars 2020
Priorité:	Normal	**Echéance:**	
Assigné à:	Benjamin Dauvergne	**% réalisé:**	0%
Catégorie:		**Temps estimé:**	0:00 heure
Version cible:			
Patch proposed:	Oui	**Planning:**	Non

Description

La création d'un connecteur tableur avec un fichier .xlsx débouche sur une erreur « Fichier CSV invalide : [Errno 2] No such file or directory: u'/var/lib/passerelle/tenants/passerelle-calvados.test.entrouvert.org/media/communes_apcr.xlsx' »

Révisions associées

Révision 14a4cce0 - 23 avril 2020 21:11 - Benjamin Dauvergne

csvdatasource: store data in temporary file when loading excel sheets (#40704)

Historique

#2 - 12 mars 2020 19:40 - Thomas Noël

```
diff --git a/passerelle/apps/csvdatasource/models.py b/passerelle/apps/csvdatasource/models.py
index a8d85467..7c47cdd9 100644
--- a/passerelle/apps/csvdatasource/models.py
+++ b/passerelle/apps/csvdatasource/models.py
@@ -234,7 +234,7 @@ class CsvDataSource(BaseResource):
                 content = get_data_ods(self.csv_file)

             elif file_type == 'xls' or file_type == 'xlsx':
-                content = get_data_xls(self.csv_file.path)
+                content = get_data_xls(self.csv_file)

             if len(content.keys()) == 1:
                 # if there's a single sheet, ignore specified sheet name and
```

Ça n'a jamais du être testé en vrai depuis un changement quelque part, et les tests sont pas bons. J'enquête...

#3 - 12 mars 2020 19:48 - Thomas Noël

(enquête abandonnée pour l'instant, à votre bon coeur)

#4 - 13 mars 2020 14:58 - Thomas Noël

Ca vient de #39971 qui ne peut pas fonctionner en xlsx car le fichier n'est pas encore sur le disque au moment du get_rows()

#5 - 13 mars 2020 15:52 - Thomas Noël

L'API de pyexcel_xls est tellement pas documentée que je doute d'y arriver (tenté plein de truc, fini avec

```
-                content = get_data_xls(self.csv_file.path)
+                self.csv_file.seek(0)
+                afile = six.StringIO(self.csv_file.read())
+                content = get_data_xls(afile)
```

qui se termine en :

```
passerelle/apps/csvdatasource/models.py:239: in get_rows
    content = get_data_xls(afile)
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_xls/__init__.py:32
: in get_data
    return read_data(afile, file_type=file_type, **keywords)
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_io/io.py:73: in ge
t_data
```

```
    afile, file_type=file_type, streaming=False, **keywords
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_io/io.py:91: in _g
et_data
    return load_data(**keywords)
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_io/io.py:216: in l
oad_data
    result = reader.read_all()
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_xls/xlsr.py:175: i
n read_all
    self._native_book = self._get_book()
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/pyexcel_xls/xlsr.py:205: i
n _get_book
    xls_book = xlrd.open_workbook(**xlrd_params)
/tmp/tox-thomas/passerelle/py2-django111-pg-junit/local/lib/python2.7/site-packages/xlrd/__init__.py:138: in o
pen_workbook

    def open_workbook_2007_xml(zf,
                               component_names,
                               logfile=sys.stdout,
                               verbosity=0,
                               use_mmap=0,
                               formatting_info=0,
                               on_demand=0,
                               ragged_rows=0):
        ensure_elementtree_imported(verbosity, logfile)
        bk = Book()
        bk.logfile = logfile
        bk.verbosity = verbosity
        bk.formatting_info = formatting_info
        if formatting_info:
>           raise NotImplementedError("formatting_info=True not yet implemented")
E           NotImplementedError: formatting_info=True not yet implemented
```

Je lâche l'affaire, considérant qu'on devrait abandonner xls et xlsx et ne prendre en charge que ods et csv.

#6 - 13 mars 2020 16:35 - Benjamin Dauvergne

Thomas Noël a écrit :

> L'API de pyexcel_xls est tellement pas documentée que je doute d'y arriver (tenté plein de truc, fini avec
> [...]
> qui se termine en :
> [...]
>
> Je lâche l'affaire, considérant qu'on devrait abandonner xls et xlsx et ne prendre en charge que ods et csv.

L'objet file de Django est caca mais on peut faire plus simple :

```
with open(self.csv_file.path, 'rb') as fd:  # enlever 'b' si pas nécessaire, c'est bizarre que ça marche en St
ringIO, xls c'est pas du binaire ?
     content = get_data_xls(fd)
```

#7 - 13 mars 2020 16:55 - Thomas Noël

Le soucis c'est que self.csv_file.path pointe vers un truc qui n'existe pas (au moment du get_rows() fait par le clean()). C'est un faux chemin (genre media/filename) d'où l'erreur qui s'affiche (No such file or directory: u'/var/lib/passerelle/tenants/xxx/media/file.xlsx)

#8 - 13 mars 2020 20:27 - Benjamin Dauvergne

Thomas Noël a écrit :

> Le soucis c'est que self.csv_file.path pointe vers un truc qui n'existe pas (au moment du get_rows() fait par le clean()). C'est un faux chemin (genre media/filename) d'où l'erreur qui s'affiche (No such file or directory: u'/var/lib/passerelle/tenants/xxx/media/file.xlsx)

Ok, de toute façon c'est moche.

#10 - 23 avril 2020 11:45 - Thomas Noël

- Fichier 0001-csvdatasource-do-not-validate-xls-xlsx-files-40704.patch ajouté

- Statut changé de Nouveau à Solution proposée

- Patch proposed changé de Non à Oui

Je propose ce contournement (ne pas chercher à valider les fichiers xls/xlsx reçus) ; je n'ai vraiment rien trouvé d'autre.

#11 - 23 avril 2020 21:10 - Benjamin Dauvergne

- Assigné à mis à Benjamin Dauvergne

#12 - 23 avril 2020 21:12 - Benjamin Dauvergne

- Fichier 0001-csvdatasource-store-data-in-temporary-file-when-load.patch ajouté

- Tracker changé de Bug à Development

Ça m'a fait de la peine, mais les tests de csvdatasource c'est juste Verdun en '17.

#13 - 27 avril 2020 13:17 - Thomas Noël

- Statut changé de Solution proposée à Solution validée

#14 - 27 avril 2020 17:43 - Benjamin Dauvergne

- Statut changé de Solution validée à Résolu (à déployer)

```
commit 14a4cce0ba826269754a50c9f9836d4bf21b6771
Author: Benjamin Dauvergne <bdauvergne@entrouvert.com>
Date:   Thu Apr 23 21:11:02 2020 +0200

    csvdatasource: store data in temporary file when loading excel sheets (#40704)
```

#15 - 29 avril 2020 02:16 - Frédéric Péters

- Statut changé de Résolu (à déployer) à Solution déployée

Fichiers

0001-csvdatasource-do-not-validate-xls-xlsx-files-40704.patch	1,21 ko	23 avril 2020	Thomas Noël
0001-csvdatasource-store-data-in-temporary-file-when-load.patch	3,56 ko	23 avril 2020	Benjamin Dauvergne