Cerema

DGALN

Direction territoriale Nord-Picardie

Import des données DVF

dans une base PostgreSQL

selon le modèle de données « ADEF-Expertise »

Crédit photo : © Arnaud Bouissou/MEIL-MEDDE

Mars 2015

Bordereau Documentaire

Informations du document

Nature du rapport :
- ☐ Intermédiaire
- ☑ Définitif

Diffusion :
- ☐ Confidentiel (diffusion réservée au Cerema)
- ☐ Diffusion restreinte
- ☑ Diffusion libre

Maître d'ouvrage

DGALN

Historique des versions du document

Version	Date	Commentaire
1.0	10/10/2014	Version initiale
1.1	30/10/2014	Prise en compte des remarques lors de la présentation au groupe national DVF
1.2	13/03/2015	Amélioration des scripts et corrections mineures

Auteurs

Antoine Herman - Cerema – Direction territoriale Nord Picardie / RDT / IGS
Tél. 03 20 49 62 34 - Courriel : antoine.herman@cerema.fr

Jérôme Douché - Cerema – Direction territoriale Nord Picardie / RDT / IGS
Tél. 03 20 49 62 59 - Courriel : jerome.douche@cerema.fr

Contributeur

Annabelle Berger - Cerema – Direction territoriale Nord Picardie / Responsable du PCI Foncier Stratégies Foncières
Tél. 03 20 49 63 59 - Courriel : annabelle.berger@cerema.fr

Résumé

Les travaux autour de la création d'un modèle permettant de faciliter le passage des données DGFiP, « demande de valeur foncière » (DVF), en base de données aisément exploitable s'inscrivent dans un processus qui a débuté en 2011 à l'initiative d'un groupe technique lancé par l'ADEF. Ce groupe s'est réuni plusieurs fois et a associé en 2013 le CETE Nord-Picardie (aujourd'hui Direction territoriale Nord-Picardie du Cerema). Un modèle de données pour l'import des données natives de DVF a été produit et partagé collégialement par le groupe.

Dans son étude relative à DVF pour l'EPF Nord-Pas de Calais et la DGALN, le CETE Nord-Picardie a mis en œuvre, développé et testé des programmes d'import rapide des données sources DVF dans une base de données PostgreSQL selon le modèle défini.

Ce document présente les scripts SQL permettant l'import des données sources DVF de la DGFiP dans un modèle de données proche de celui défini initialement par l'ADEF.

Contexte

Une réflexion participative

Dès 2010, les demandes relatives à l'exploitation des données DVF ont conduit l'ADEF à lancer un groupe de réflexion participatif pour structurer les données issues du service de téléchargement « Demande de valeurs foncières », mis en place par la DGFiP. Ce service permet aux collectivités territoriales, à leurs groupements dotés d'une fiscalité propre et à leurs établissements publics d'obtenir, sur leur demande, les informations relatives aux valeurs foncières déclarées à l'occasion des mutations intervenues dans les cinq dernières années et qui sont nécessaires à l'exercice de leurs compétences en matière de politique foncière et d'aménagement. La disponibilité de cette donnée permet de reconstituer une capacité d'observation des marchés fonciers en région, telle qu'elle avait existé de 1991 à 2003 dans le Nord-Pas de Calais grâce à l'Observatoire Régional de l'Habitat et de l'Aménagement (ORHA).

Ce groupe a initialement réuni les bénéficiaires directs des données, principalement les établissements publics fonciers. Le groupe technique, composé de géomaticiens et d'informaticiens a été piloté par Pascal Barillé (EPF des Yvelines – aujourd'hui Pays d'Aurais), et a réuni 7 autres représentants d'organismes bénéficiaires de DVF : Arnaud Buisson Delandre (EPF Lorraine), Nicolas Debey (EPF Normandie), Julien Déniel (EPF Bretagne), Julien Place (EPF Ile de France), Edouard Hyvernat (EPF Yvelines), Julien San José (Agam), Thomas Brosset (Quelleville?). Le CETE Nord-Picardie (Jérôme Douché - CETE Nord-Picardie - Pôle de Compétences et d'Innovations « Foncier et Stratégies Foncières ») a également participé à ce groupe notamment pour faire bénéficier de son expérience sur le traitement et l'enrichissement des Fichiers fonciers (issus de MAJIC de la DGFiP) au niveau national depuis 2009 et faciliter dans un second temps le rapprochement des données DVF avec les Fichiers fonciers.

Un modèle de données partagé

Les travaux de ce groupe ont permis d'élaborer une première structuration du modèle des données DVF pour en faciliter l'usage et ainsi la connaissance des marchés régionaux.

Le premier modèle de données, élaboré par le groupe a été diffusé le 12 novembre 2012 sous licence Creative Communs en paternité et partage dans les mêmes conditions (CC-BY-SA-3.0). Cela implique une liberté de partage, de remixer et d'adapter le modèle, sous condition d'attribuer le modèle de la manière indiquée par l'auteur de l'œuvre ou le titulaire des droits (en l'occurrence, « ADEF-expertise »). Dans le cas où le modèle est modifié, transformé ou adapté, la distribution du nouveau modèle doit se faire sous une licence identique ou similaire à celle-ci.

Une expérimentation du traitement des données

Dans son étude relative à DVF pour l'EPF Nord-Pas de Calais et la DGALN, le CETE Nord-Picardie a mis en œuvre, développé et testé des programmes d'import rapide des données sources DVF dans une base de données PostgreSQL selon un modèle proche de celui défini par le groupe « ADEF-expertise ».

Description des éléments produits

Éléments produits et droits d'utilisation

Pour la mise en œuvre de l'import des données, trois scripts SQL ont été créés et ont nécessités des modifications mineures du modèle de données DVF du groupe « ADEF-expertise ».

Création de trois scripts SQL

Afin de passer des fichiers livrés par la DGFiP au modèle de données, trois scripts SQL pour PostgreSQL ont été réalisés sous licence Creative Commons 3.0 CC-BY-SA (cf. mentions ci-après). Ces scripts permettent d'importer à la fois les données principales de DVF mais également les données annexes descriptives de champs spécifiques.

Au final, 12 tables principales et 5 tables annexes sont créées (cf. modèle de données de DVF).

Mentions légales des scripts

Traitement particulier pour la table « suf »[1]

Dans les données DVF, il n'y a pas d'identifiant de suf de la parcelle. De ce fait, il est complexe de repérer :

- d'une part les informations redondantes ne représentant qu'une seule suf au sein de la parcelle ;
- d'autre part les informations liées à deux suf distinctes mais disposant des mêmes caractéristiques au sein de cette parcelle.

Pourtant, il est important d'effectuer cette distinction afin de s'approcher au mieux de la surface de la suf et ainsi déterminer la surface de la parcelle mutée (en sommant les surfaces des suf différentes). Aussi, un simple regroupement de lignes identiques n'est pas suffisant.

La méthode mise en œuvre dans les scripts livrés est la suivante :

- regroupement des suf similaires au sein d'une parcelle en les comptant et en sommant le champ « surface_terrain » ;
- puis pour chacun des regroupements distincts de suf d'une même parcelle, division de ces deux valeurs (nombre de sufs identiques et somme des surfaces identiques) calculées par le plus grand commun diviseur du nombre de sufs identiques regroupées.

Adaptations légères du modèle de données

Dans sa phase de mise en œuvre, le modèle de données a dû être légèrement adapté :

- suppression du champ « insee_commune » de la table « mutation » car une mutation peut concerner plusieurs communes ;
- déplacement du champ « nombre_de_lots » de la table « disposition_parcelle » à la table « disposition » comme indiqué dans la documentation fournie par la DGFiP. Néanmoins, certaines dispositions contiennent des nombres de lots différents. Le champ a donc été mis en type tableau d'entier (noté « entier[] ») ;
- ajout des champs « idsuf_tmp » et « idadr_tmp » respectivement dans les tables « suf » et « adresse » afin d'accélérer l'import des données ;
- ajout d'un champ « nb_suf_idt » dans la table « suf » indiquant le nombre de sufs a priori identiques d'une même parcelle (« nature_culture », « nature_culture_speciale » et « surface_terrain » identiques) mais distinctes. Le champ « surface_terrain » correspondant à la somme des surfaces réelles pour une même « nature de culture/nature de culture spéciale ».

Le modèle, fourni en annexe, a été publié de nouveau selon la licence initiale (également identique à celle des scripts fournis).

[1]Une « suf » est l'acronyme de subdivision fiscale et représente une partie homogène de parcelle quant à son affectation.

Présentation des trois scripts

Les trois scripts SQL livrés fonctionnent sous une base de données PostgreSQL 9.x et réalisent les opérations suivantes :

	Description
00_creation_base_vide_v1.2.sql	Création du schéma « dvf » et des tables dans le schéma (principales et annexes). **Important : lors de l'utilisation de ce script, toutes les tables du schéma « dvf » et toutes les vues utilisant les tables du schéma « dvf » sont supprimées.**
01_import_donnees_annexes_v1.2.sql	Import des données annexes livrées lors de la récupération des données du millésime DVF. Ce script nécessite de paramétrer l'emplacement et le nom des trois fichiers annexes à importer.
02_import_donnees_dvf_v1.2.sql	Import des données principales de DVF dans le schéma « dvf » en utilisant un schéma temporaire « import » (supprimé automatiquement en fin de script). Ce script nécessite de paramétrer l'emplacement et le nom du fichier des données principales à importer. **Les fichiers doivent être importés un par un. Afin de prendre en compte les dernières mises à jour des données, il est a priori plus pertinent d'importer les millésimes du plus récent vers le plus ancien.**

Le script « 00_creation_base_vide_v1.2.sql » n'est à exécuter que la première fois pour créer la base vide. **En cas de ré-exécution, la base DVF serait supprimée sans message d'avertissement**.

Utilisation des scripts

Fonctionnement général

Trois étapes différentes permettent de mettre en œuvre l'import des données.

En préalable à tout import, il est nécessaire de préparer la base de données (étape 0). Cette étape n'est à réaliser qu'**une seule fois** avant le premier import de données.

Ensuite, **lors de chaque nouvelle mise à disposition** des données DVF par la DGFiP, il est nécessaire d'importer **dans cet ordre** :

- les données annexes (étape 1) ;
- les données principales (étape 2).

Étape 0 : préparer la base de données

La préparation des données nécessite de créer une base PostgreSQL 9.x (nommée par exemple « dvf ») via une commande SQL de type « CREATE DATABASE dvf » par exemple.

Il faut ensuite exécuter le script « 00_creation_base_vide_v1.2.sql » qui créera un schéma « dvf » et les différentes tables dans lesquelles seront importées les données (principales et annexes).

Important : lors de l'utilisation de ce script, toutes les tables du schéma « dvf » et toutes les vues utilisant les tables du schéma « dvf » sont supprimées sans message d'avertissement.

Étape 1 : importer les données annexes de DVF

L'import des données annexes (groupes de natures de cultures, groupes de natures de cultures spéciales et description des articles cgi) s'effectue en deux temps :

- import dans des tables temporaires ;
- ajout des nouvelles données dans les tables annexes situées dans le schéma « dvf » (respectivement « ann_nature_culture », « ann_nature_culture_spec » et « ann_cgi »).

Cet import s'effectue via le script « 01_import_donnees_annexes_v1.2.sql ».

Trois lignes (bien indiquées dans le script) sont à modifier dans ce fichier afin de spécifier les noms et chemins des fichiers annexes.

Étape 2 : importer les données principales de DVF

L'import des données principales s'effectue en deux temps :

- import du fichier et prétraitement de données dans un schéma « import » qui est supprimé en fin de script ;
- ajout des nouvelles données dans les différentes tables du schéma « dvf ». La table « ann_nature_mutation » sera également complétée dans cette étape.

Cet import s'effectue via le script « 02_import_donnees_dvf_v1.2.sql ».

Une ligne (requête '\COPY … FROM …', bien indiquée dans le script) est à modifier dans ce fichier afin de spécifier le nom et chemin du fichier principal.

Il ne faut importer qu'un seul fichier à la fois.

Sur une machine récente, il faut compter moins d'une minute pour l'import d'un fichier départemental de DVF.

Nota :Si le script est exécuté depuis l'interface pgAdmin pour un import en local, il faut, au préalable, supprimer l'antislash (\ ) devant la requête '\COPY … FROM ...'.

Automatisation possible

Pour effectuer plusieurs requêtes SQL successivement, il est possible de créer un fichier batch contenant un ensemble de commandes « psql » exécutant chacune un fichier SQL.

Cette automatisation est particulièrement intéressante dans le cas où vous voulez importer l'ensemble des données DVF en partant de la dernière livraison.

Préparer les fichiers SQL spécifiques à chaque livraison

Pour chaque livraison de DVF, il faut préparer les deux fichiers « 01_import_donnees_annexes.sql » et « 02_import_donnees_dvf.sql » :

- en indiquant les noms et chemins des fichiers à importer (cf. étape 1 et 2 décrites ci-dessus) ;
- puis en les renommant de manière explicite.

Créer un fichier exécutant l'ensemble des scripts

Il faut ensuite écrire un fichier permettant d'exécuter l'ensemble des scripts SQL comme ci-après.

Dans l'exemple ci-dessous :

- <adresse_ip> est à remplacer par l'adresse IP du serveur PostgreSQL (exemple : *127.0.0.1*),
- <port> est à remplacer le numéro du port (a priori *5432* pour une installation standard),
- <nombdd> est à remplacer par le nom de la base de données (par exemple *dvf*),
- <nom_user> est à remplacer le nom de l'utilisateur de PostgreSQL (par exemple *postgres*).

Si pour chaque ligne un mot de passe est demandé, il est alors nécessaire de compléter le fichier de mots de passe de PostgreSQL[2].

Exemple de fichier exécutant l'ensemble des scripts pour intégrer 4 livraisons de DVF (2 livraisons par an)

```
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 00_creation_base_vide.sql

psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 01_import_donnees_annexes_2014_10.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 01_import_donnees_annexes_2014_04.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 01_import_donnees_annexes_2013_10.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 01_import_donnees_annexes_2013_04.sql

psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 02_import_donnees_dvf_2014_10.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 02_import_donnees_dvf_2014_04.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 02_import_donnees_dvf_2013_10.sql
psql -h <adresse ip> -p <port> -d <nombdd> -U <nom_user> -f 02_import_donnees_dvf_2013_04.sql
```

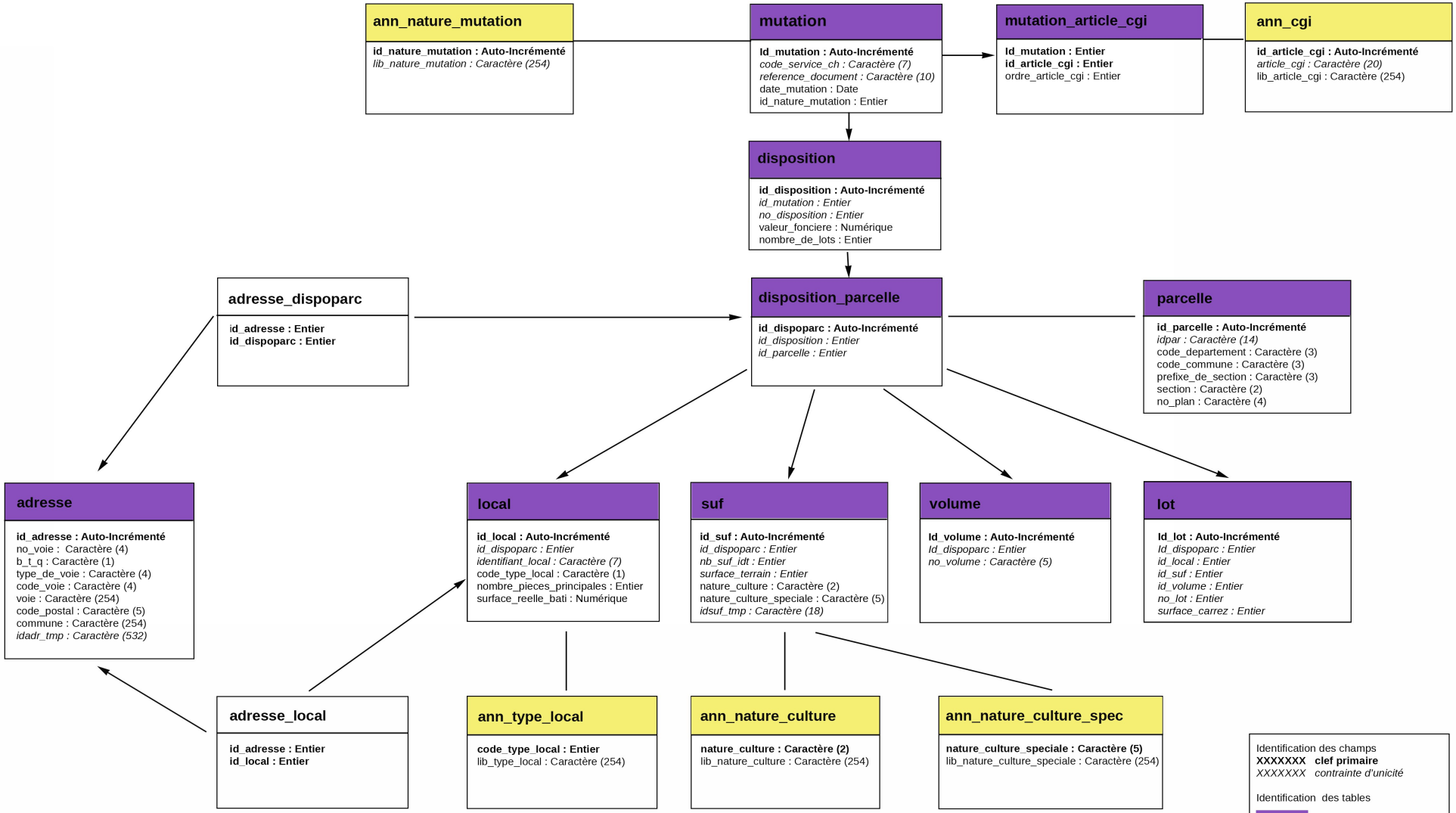
[2] cf. documentation http://docs.postgresql.fr/9.1/libpq-pgpass.html

Cerema

Direction territoriale Nord-Picardie

Modèle de données de DVF

CEREMA Direction territoriale Nord-Picardie - Mars 2015



CEREMA Direction territoriale Nord-Picardie
PCI Foncier Stratégies Foncières / Jérôme Douché