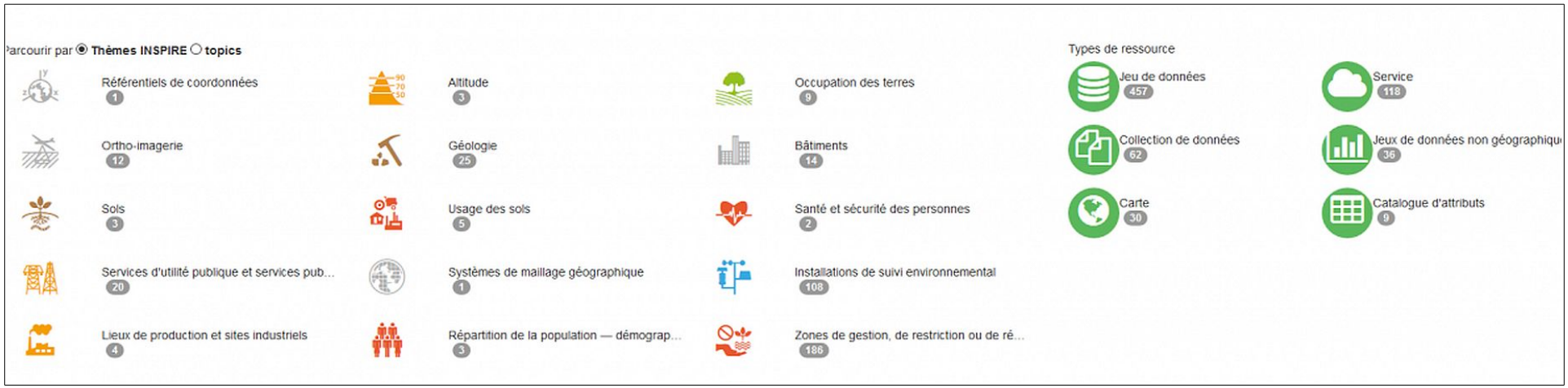
Cerema

SIGN@TURE
Lettre de l'Information Géographique

Avril 2019 68

Les données : leur ouverture, leur qualité, leurs usages



Sommaire

A la une
- CeremaData : l'opendata du Cerema
- Cinq clés pour comprendre le plan d'application satellitaires

Zoom sur
- Du Machine Learning pour estimer le type constructif des bâtiments
- Les nouveautés de QGIS 3,6 « Noosa »

Dossier
- Nouveau : le GT qualification des données géographiques au CNIG !
- La qualité des données géographiques à la Métropole de Lyon
- La qualité des données géographiques en 120 secondes, par le CRIGE PACA

Vu, lu, entendu
- Qualité des données, Machine Learning : quelques propositions de lecture

Éditorial

OPENDATA, LES COMPTES N'Y SONT PAS

Il y a 60 ans naissait l'internet. Ce fut l'occasion de nombreuses célébrations de cette rupture technologique, qui invitèrent parfois à se demander pourquoi on parle encore de transition numérique aujourd'hui tant on ne peut plus revenir en arrière. 60 ans, l'âge d'un essor grandissant pour l'internet qui ne bat pas en retraite. D'un essor, mais aussi de dérives qui le dénaturent et qui inquiètent jusqu'aux Etats et jusqu'à ces tout premiers inventeurs dont fait partie Louis Pouzin. Resté un peu dans l'ombre du choix français d'une commutation de paquets qui a conduit au Minitel, aux transactions plus facilement facturables ("circuits virtuels" normés par Transpac X25), Louis Pouzin défendait, et avait théorisé (Datagramme et "routage adaptatif"), une autre solution qui a conduit les Etats-Unis à développer l'internet que nous connaissons. Quels choix commet-on parfois pour des histoires d'argent. Louis Pouzin milite aujourd'hui activement pour "ouvrir" l'internet, et sa gouvernance, comme on ouvre les sources des logiciels ou comme on ouvre les données.

Plutôt comme on devrait, car c'est plus facile à dire qu'à faire semble-t-il, comme en témoigne un récent référé de la Cour des Comptes. En même temps qu'il se créait, le Cerema se préparait à cette ouverture de données en mettant

en place sa propre infrastructure de données fondée sur la solution Prodige. Il a fallu beaucoup de ténacité aux équipes concernées pour porter ce projet et cela a porté ses fruits comme en témoigne ce nouveau numéro de Sign@ture, permettant techniquement au Cerema d'être en conformité avec la Loi.

Ouvrir des données c'est bien. Merci encore à l'internet de le permettre. Ouvrir des données dont on connait la qualité, voire de qualité, c'est mieux. Et le sujet de la qualité des données, initié par le Cerema et le Crige Paca, est aujourd'hui entre de bonnes mains au CNIG.

Ouvrir les données conduit souvent à des gisements très importants, qu'il peut être difficile de traiter et de visualiser, cela ne fait pas grand débat (la réciproque n'est pas vraie).

Mais la dataviz et le deep learning pourraient venir à la rescousse. Une preuve supplémentaire est illustrée dans ce numéro et qui permet au Cerema de s'engager sereinement dans la voie des données en lien avec son expertise multi-thématique.

Bernard Allouche

SIGN@TURE Lettre de l'Information Géographique
Cerema

CeremaData : l'opendata du Cerema

Entre obligations et recommandations, la plateforme CeremaData, mise en ligne en Février 2019, s'inscrit dans le processus d'ouverture des données.

Lancement de CeremaData

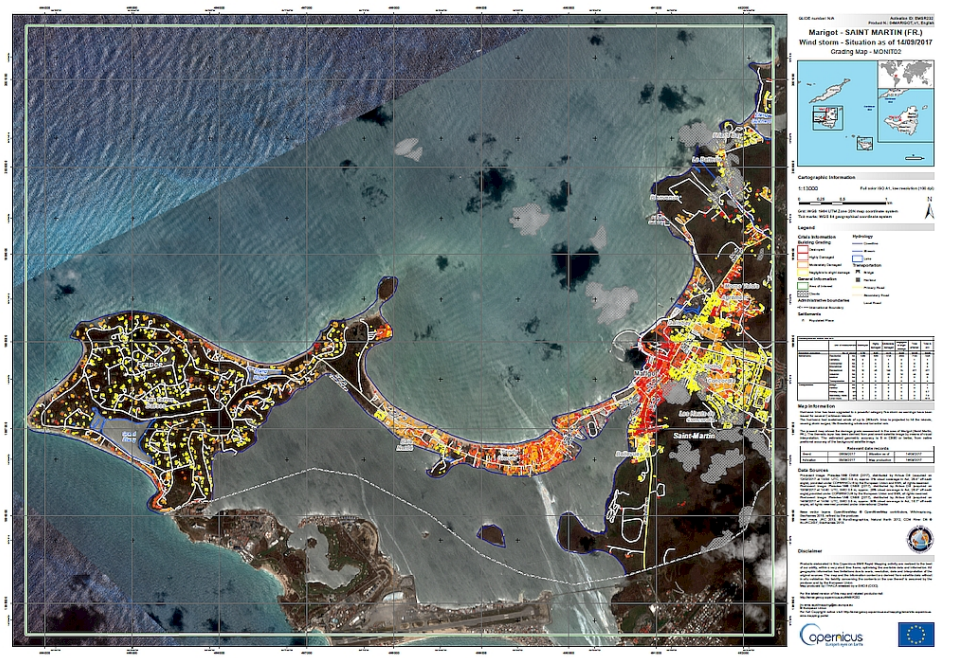
Le Cerema produit et analyse des données depuis de nombreuses années, via les études qu'il produit sur ses domaines d'action qui sont souvent territorialisés. Il gère et diffuse également des données de partenaires, Etat ou collectivités. Et en février 2019, la plateforme CeremaData a été lancée : son objectif est de donner accès aux données et de valoriser les ressources et savoir faire numériques du Cerema.

L'opendata des opérateurs de l'état : un sujet dont s'est saisie la cour des comptes

Le 11 mars 2019, la Cour des comptes a en effet publié un référé sur "La valorisation des données de l'IGN, de Météo-France et du Cerema". En pointant du doigt "des difficultés d'application récurrentes et un pilotage insuffisant" dans l'ouverture des données chez ces 3 opérateurs, la Cour des comptes préconise de clarifier :

- l'application des lois open data pour les établissements publics ;
- le modèle économique des données ouvertes.

Cinq clés pour comprendre le Plan d'applications satellitaires



Le Plan d'applications satellitaires (PAS) 2018 du MTES et du MCTRCT a été publié en septembre 2018.

Un précédent Plan d'applications Satellitaires encourageant

Le PAS 2018 fait suite au PAS 2011, dont le bilan a montré plusieurs avancées :

- d'une part sur les actions thématiques : développement durable des territoires (exemple : produits d'occupation du sol), gestion durable du littoral (exemple : évolution du trait de côte), systèmes d'observation globale de la Terre (exemple : utilisation des images et produits du programme Copernicus), mobilité durable (exemple : apports socio-économiques de Galileo à la gestion du trafic) ;

- d'autre part sur les actions transversales : mise en place du pôle de compétences et d'innovation « Applications satellitaires et télécommunication

Ce bilan favorable a encouragé à la préparation d'un nouveau Plan, qui s'inscrit dans la suite du précédent, et dans le nouveau contexte actuel, tant technique que d'enjeux. Il s'est largement appuyé sur l'expression des besoins des utilisateurs.

Ce nouveau Plan, dont le détail est consultable sur GéoInformation, est organisé autour de 3 parties :

1. l'environnement actuel, favorable à l'utilisation des applications satellitaires
2. les actions inscrites au PAS
3. les suites et la mise en œuvre

Cinq idées structurantes permettent d'avoir une compréhension de ce document.

1 - La phase d'élaboration a été particulièrement riche et dynamique

Cette phase préparatoire a dégagé une réelle envie des services. Dix groupes de travail, qui recouvrent les principales missions des deux ministères, ont été mis en place[1]. Plus de 140 personnes ont participé à ces travaux, venant de services très différents (principalement les Directions Générales des Ministères, les établissements publics, les DREAL). Cette diversité a permis d'exprimer 136 sujets d'intérêt très variés. Après analyse (sujets identiques émis par deux groupes, sujets ne relevant pas du satellitaire, instruments non disponibles, sujets non faisables techniquement), il est resté 85 sujets, reflétant ainsi la richesse des attentes des participants.

2 - Les 20 actions inscrites au PAS donnent l'image de l'évolution qui va se produire

La phase d'élaboration a nécessairement abouti au choix par les maîtres d'ouvrage pressentis des actions qu'ils acceptaient de réaliser. 20 actions ont ainsi été retenues et sont inscrites au PAS 2018. Si ce nombre peut paraître modeste par rapport aux 85 proposées, la réalisation de ces actions pendant la durée du PAS 2018, c'est-à-dire d'ici 2022, constituerait une transformation significative des pratiques satellitaires dans le ministère, au cœur des métiers. Ces actions sont accompagnées d'un véritable engagement des pilotes, leurs perspectives d'aboutir sont donc fortes.

Elles constituent une « carte d'identité » des changements attendus à cet horizon et portent sur les thèmes suivants :

- transport aérien : lutte contre le brouillage des signaux GNSS, suivi des aéronefs, approche satellitaire pour les aéronefs sur les aéroports français ;
- transport ferroviaire : optimisation des coûts de maintenance des voies ;
- lutte conte la pollution atmosphérique : inventaire des émissions de polluants, pollutions générées par les activités terrestres et marines ;
- gestion de crise : emprise des zones inondées, outil d'aide à la gestion des informations OGERIC, estimation des infrastructures dégradées ou détruites ;
- aménagement du territoire : couverture du manteau neigeux, occupation du sol ;
- biodiversité : amélioration du suivi des animaux marins, cartographie des habitats naturels, température des masses d'eau dans les espaces protégés, pression anthropique exercée par le trafic maritime, évolution du bocage et impact sur la biodiversité, caractérisation de l'hydromorphologie des cours d'eau.

L'ensemble de ces actions est décrit dans le PAS 2018, et une fiche descriptive pour chacune d'elle est jointe dans ses annexes.

Exemple de carte de crise réalisée à partir des images Pléiades (Marigot - Saint-Martin)

SIGN@TURE Lettre de l'Information Géographique
Cerema

3 - L'accès aux images reste une priorité

Cette action figurait déjà dans le PAS précédent. La mise en place du dispositif Equipex « Géosud » et le soutien efficace du CNES ont permis de bénéficier aisément d'un accès mutualisé aux images satellitaires à très haute résolution (Pleiades : 50 cm et SPOT 6 et 7 : 1.5m).

La fin du financement de l'Equipex en 2019 nécessite de trouver un nouveau dispositif pour les années à venir : il s'appellera DINAMIS. Les besoins des services sont significatifs pour des images de cette nature, et pas seulement en cas de crise : production dans des délais extrêmement brefs (quelques jours) d'une image neutre et à jour d'un territoire, photo-interprétation et mesure de changements, …

La mise en place du nouveau dispositif constitue un enjeu important pour les services du MTES et MCTRCT.

4 – La communauté des utilisateurs permettra de partager les expériences

Une première analyse des applications développées dans les services montre a priori que leur nombre semble plutôt modeste. Lors d'échanges avec les acteurs territoriaux, cette impression paraît inexacte : les applications développées ne sont effectivement pas très nombreuses, mais un nombre significatif de ces applications ne sont pas connues par les autres utilisateurs potentiels. Ce constat appelle la nécessité de mettre en place une fonction de partage des expériences.

L'intérêt de la création d'une communauté des utilisateurs est renforcé par la participation forte et active à l'élaboration du PAS 2018, qui a traduit une réelle motivation des participants.

La création de cette communauté est en cours de réflexion et pourrait inclure un ensemble de fonctions utiles à un tel groupe : partage d'expérience, forum, agenda collectif, lettre d'information, documentation de référence, … Elle s'appuiera sur une animation dynamique.

Son objectif serait d'encourager l'usage des applications satellitaires (images et GNSS) au service des politiques portées par le MTES et le MCTRCT ; et son périmètre pourrait couvrir l'ensemble des acteurs portant les politiques des 2 ministères : services ministériels, partenaires privilégiés, secteur privé (en tant que pourvoyeur de solutions), collectivités territoriales (pour les politiques dont le ministère est en charge et qu'elles mettent en œuvre).

Elle pourrait être mise en place avant l'été 2019.

5 – Le soutien à l'innovation permettra aux PME et TPE de mieux répondre aux politiques portées par les ministères

Si la France dispose, dans le secteur du satellitaire, d'entreprises de niveau mondial, il semble intéressant d'encourager l'émergence de projets et l'innovation dans les PME et TPE. En ce sens, deux séries d'actions sont en cours :

- le soutien aux Boosters (en charge, pour les Pôles de compétitivité et sur les domaines de l'espace et du numérique, d'encourager le dynamisme territorial du secteur privé),
- l'appui à l'utilisation des financements issus du Programme d'investissement d'avenir (PIA).

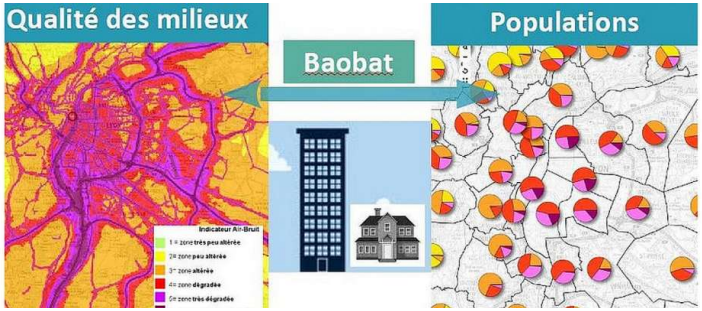
Conclusion

Ces clés de lecture donnent une vision structurante du PAS 2018, même si elles ne remplacent pas une lecture plus détaillée. Elles permettent de décrypter les principaux enjeux liés à la mise en œuvre de ce plan, qui s'achèvera en 2022.

Auteur : José Devers (CGDD/DRI/SDI/MIG)

SIGN@TURE Lettre de l'Information Géographique
Cerema

Du Machine Learning pour estimer le type constructif des bâtiments



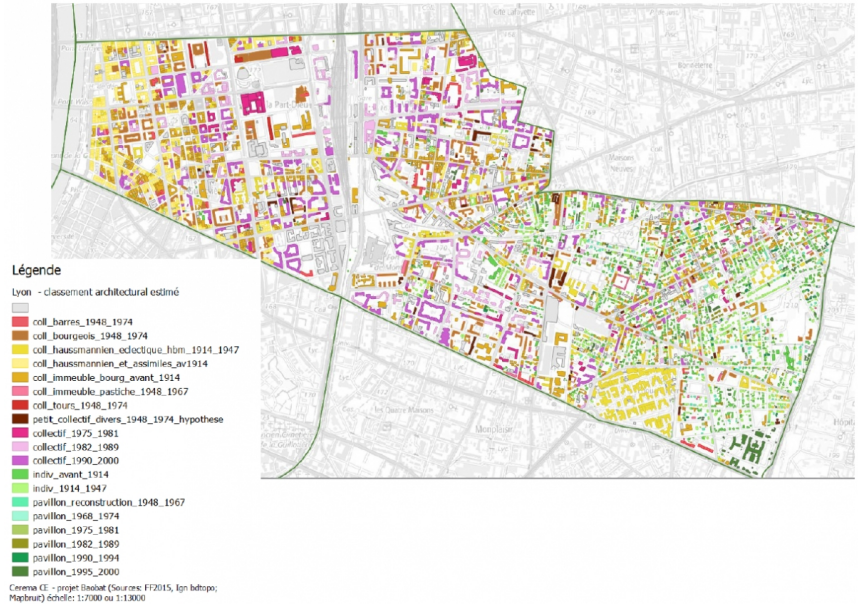
Par le croisement de données géographiques et de connaissances métier (Environnement, Bâtiment, Habitat, Changement Climatique), le Cerema propose de structurer une base de données multithématique relative à la qualité de l'enveloppe des bâtiments (Baobat).

Nuisances environnementales et bâtiments

Nous savons aujourd'hui cartographier la qualité des milieux vis-à-vis de l'air et du bruit, par exemple au travers des indicateurs de co-exposition air et bruit élaborés par le Cerema pour l'observatoire régional orhane. Par ailleurs, nous savons cartographier les populations exposées à ces nuisances. Mais dans le processus visant la connaissance de l'exposition des populations aux nuisances environnementales, demeure un maillon manquant. Il s'agit de la composante « bâtiment », qui joue pourtant un rôle essentiel : isolation acoustique, perméabilité aux polluants, émetteur potentiel.

Classification du bâti suivant le type constructif

Pour envisager cette composante « bâtiment », il a donc été proposé d'intégrer dans la base Baobat, les informations qualifiant le niveau d'isolation acoustique. Pour cette intégration, plusieurs approches sont possibles. L'une d'entre elles repose sur la corrélation existant entre types constructifs et niveau d'isolement acoustique. Pour la mettre en œuvre, il faut donc parvenir à classer les différents bâtiments selon leur type constructif. Nous nous sommes appuyés pour cela sur une classification existante dans la communauté du bâtiment, issue de l' « Analyse détaillée du parc résidentiel existant » rédigée en juillet 2017 par le programme PACTE. Cette typologie répartit les bâtiments à usage de logement en 26 classes (10 pour les maisons individuelles, 16 pour l'habitat collectif).



Si une partie des bâtiments est facile à classer par simples requêtes SQL à partir de quelques variables de la base Baobat (année de construction, hauteur des bâtiments...), cette méthode se révèle trop rigide pour être applicable à l'ensemble du bâti, et ne permet pas de classer les bâtiments en cas de données manquantes. C'est la raison pour laquelle nous avons testé des méthodes d'apprentissage supervisé pour réaliser cette typologie.

Machine learning et typologie

La méthodologie déployée a consisté à :

- construire un échantillon d'apprentissage dans lequel les bâtiments sont déjà associés à un type constructif ;
- faire une analyse exploratoire des données pour relever le nombre de données manquantes par variable ;
- déterminer les variables qui permettent de discriminer les types constructifs ;
- choisir et mettre en œuvre un algorithme de classification (en l'occurrence l'algorithme du Random Forest).

La base de données a été déployée sur le périmètre du département du Rhône et de la Métropole de Lyon.

La librairie Scikit-Learn de Python propose de nombreux algorithmes d'apprentissage adaptés à différents contextes et faciles à mettre en œuvre. L'environnement de travail des Notebook Jupyter fournit un moyen pratique d'écrire et d'exécuter du code, de tester, d'évaluer et d'optimiser un algorithme d'apprentissage.

Les conclusions de ce travail montrent l'efficacité de ces méthodes d'apprentissage automatique, qui ne requièrent que peu de paramètres et sont rapidement déployables sur de grands volumes de données. La base de données Baobat pourrait d'ailleurs servir de base d'apprentissage pour d'autres sujets (performance énergétique des bâtiments, valeur foncière...). Il convient cependant de garder un œil critique quant à la fiabilité des données produites par l'algorithme, en validant les résultats obtenus par de l'expertise sur le terrain.

Auteur : Frédéric Berlioz, chef de l'unité Géomatique (Cerema Centre-Est)

Les nouveautés de QGIS 3.6 « Noosa »

QGIS, le logiciel SIG de l'Open Source Geospatial (OSGeo), outil des ministères de la Transition Écologique et Solidaire et de la Cohésion des territoires, a sorti, le 22 février 2019, sa version 3.6.0 nommée « Noosa ».

QGIS 3, une avancée technologique

Il y a un peu plus d'un an déjà (février 2018) la nouvelle version QGIS 3.0 Girona était disponible en téléchargement. Le passage entre les versions 2.x et 3.x marque une rupture technologique du logiciel SIG :

- parce que le code à été réécrit avec la version 3 du langage de programmation Python (anciennement en Python 2) ;
- parce que l'interface graphique s'appuie maintenant sur la version 5 de la bibliothèque QT (anciennement en QT 4) ;
- parce qu'un grand nombre d'algorithmes ont été ré-écrits en C++ permettant une exécution en multithreading.

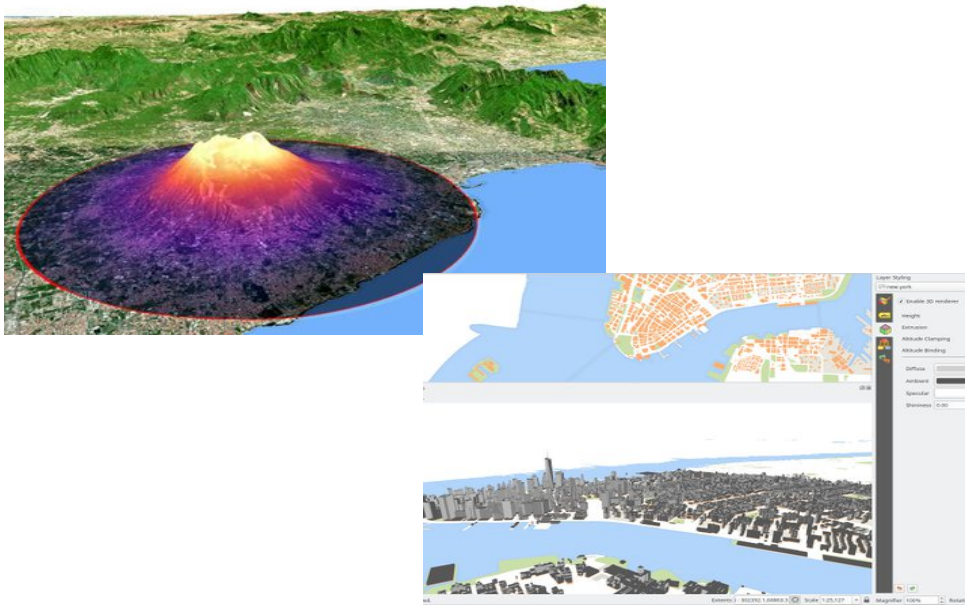
SIGN@TURE Lettre de l'Information Géographique
Cerema

Avantages et inconvénients

Comme à chaque nouvelle version, QGIS propose un nombre important de nouveautés et améliorations qui sont listées et consultables sur le site officiel en suivant le lien https://qgis.org/fr/site/forusers/visualchangelogs.html.

Mais c'est bien le passage à Python 3, QT 5 et au multithreading qui apportent les avantages les plus marquants… et parfois quelques inconvénients.

L'utilisation de QT 5 apporte de nouvelles fonctionnalités graphiques comme l'utilisation de nouveaux widgets et graphiques, et surtout un visualiseur de carte 3D. Cette visualisation est accessible soit en utilisant une couche raster MNT de drapage, soit en utilisant une couche vecteur extrudée grâce à sa 3ème dimension (z), soit en utilisant des symboles 3D (formes géométriques simples comme les arbres).



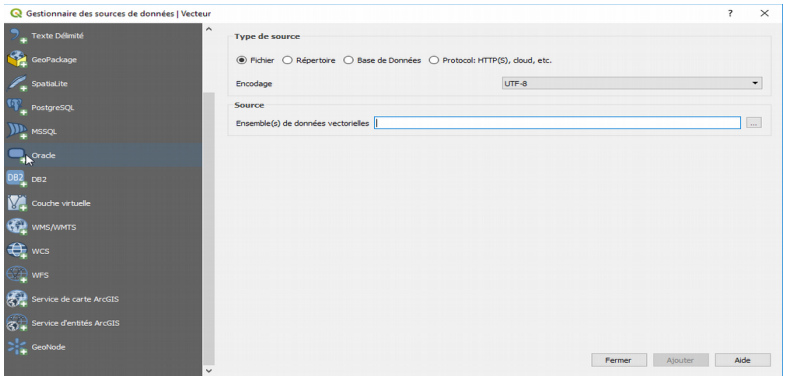
Le passage à Python 3 et à QT 5 a nécessité la réécriture de l'ensemble des extensions QGIS. Les effets sont positifs…. comme négatifs. En effet cette réécriture a entraîné un nettoyage dans la multitude des extensions (avec la suppression des extensions développés en C++ difficiles à maintenir), mais elle a aussi entraîné la disparition d'un certain nombre d'entre elles. Du temps sera donc nécessaire à la réécriture des extensions pour qu'elles soient compatibles avec QGIS 3.x (en Mars 2018, les services techniques du Ministère de la Transition Ecologique et Solidaire ont recensé 827 extensions dans le répertoire officiel QGIS 2.18 contre 103 pour QGIS 3.0).

Enfin le multithreading, exécution en tâche de fond (ou arrière plan), appliqué aux traitements complexes, permet de ne pas bloquer QGIS lors de l'exécution de ces derniers. Pour l'expliquer simplement : le multithreading, permet l'utilisation de QGIS sans attendre que le traitement lancé soit terminé : plusieurs tâches sont donc possibles en même temps.

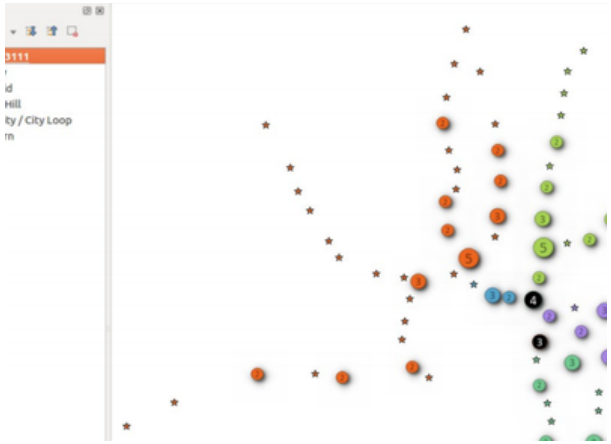
Autres améliorations

A ces 3 évolutions, il peut être ajouté entre autres (la liste est loin d'être exhaustive) :

- Un « Gestionnaire des sources des données » permettant l'ouverture de tous les types de ressources (vecteur, raster, PostgreSQL, WMS, etc.)



- Une barre de recherche (coin bas-gauche) qui permet de tout (ou presque) rechercher : algorithmes, actions, paramètres, composeurs d'impression, couches et entité particulière de la couche active courante, etc.
- L'intégration complète du format GeoPackage dans les outils de QGIS qui est maintenant supporté en format de sortie des traitements ;
- L'intégration des nouveaux formats de données PostgreSQL comme « array » et « hstore » ;
- La détection automatique des relations entre tables PostgreSQL ;
- Le chargement des mises pages « à la demande » permettant d'éliminer le ralentissement important du composeur d'impression des versions 2.x ;
- Une nouvelle symbologie avec notamment la génération d'une nouvelle géométrie (utile pour une représentation de données temporelles) ou encore une représentation ponctuelle par cluster



- La possibilité de créer de nouvelles couches (Shapefile, Geopackage ou Spatialite) avec la dimension Z.

Version packagée 3.4.5 POUR LES MINISTERES

Depuis le début de l'année 2019 le Pôle National d'Expertise Progiciels géomatiques (du ministère de la Transition Écologique et Solidaire) met à disposition un package pour la version QGIS 3.4.5 LTR 64 bits.

Cette version a vocation à devenir la nouvelle version de référence tandis que la version 2.16.3 restera aussi « référence » jusqu'au basculement de l'ensemble des services.

Les détails de ce package et les informations nécessaires à son installation sont disponibles sur le site Géoinformation : http://www.geoinformations.developpement-durable.gouv.fr/qgis-package-3-4-5-ltr-a3662.html.

Auteur : Antoine Lemot (Cerema Centre-Est)

SIGN@TURE Lettre de l'Information Géographique
Cerema

Un tableau de bord pour l'analyse des données de valeurs foncières pour la Métropole Nice Côte d'Azur

Dans le cadre de son observatoire du foncier, la métropole Nice Côte d'Azur s'est intéressée à la dynamique des transactions immobilières sur son territoire sur la période de 2010 à nos jours.

Pour l'étude de cette dynamique, elle a fait appel au Cerema Méditerranée, en particulier au Département Aménagement des Territoires qui a lui-même mis à contribution le service d'Appui Géomatique pour la partie développements informatiques.

Les données : Fichiers Fonciers, Demande de Valeurs Foncières et DV3F

Le Cerema, depuis environ 2011, a acquis une expertise reconnue dans l'utilisation des fichiers fonciers qu'il retraite et distribue aux acteurs publics de l'aménagement (collectivités et agence d'urbanisme en particulier) qui en font la demande. Il est aussi membre du groupe national relatif à DVF (https://www.groupe-dvf.fr/). De la conjugaison de son savoir-faire sur les fichiers fonciers et de celui sur DVF est née la base DV3F qui permet de relier l'information des fichiers fonciers (et parcellaire) à celle des transactions immobilières.

D'un côté, les fichiers DVF fournissent un historique des transactions depuis 2010 à l'échelle de la parcelle et du local ; de l'autre les fichiers fonciers décrivent assez précisément les parcelles. Ainsi, la base DV3F, en associant les deux, permet de retracer les dynamiques du prix de l'immobilier selon le type d'occupation à l'échelle d'un territoire : maison, appartement, local d'activité, dépendance, terrains nus.

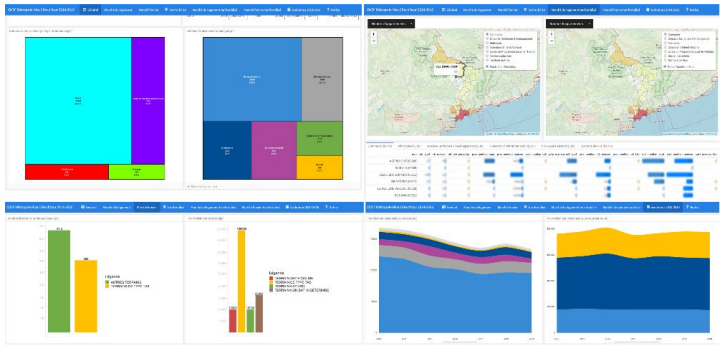
La datavisualisation, des rendus innovants

Le Cerema s'intéresse de plus en plus aux rendus dits innovants.

Après une première étude des marchés immobiliers pour le compte de la DREAL Corse, pour la Métropole Nice Côte d'Azur, il a été décidé d'associer aux livrables classiques (rapport, note technique,...) un rendu novateur des indicateurs, sous forme de tableau de bord. Ce dernier, interactif, a été fourni au format web afin que la Métropole puisse l'intégrer à son site internet.

Sur le contenu, le Cerema a proposé, en concertation avec la Métropole Nice Côte d'Azur, une trentaine d'indicateurs permettant de caractériser le marché immobilier avec des data-visualisations diverses permettant de les illustrer : graphiques en bâtons, tableaux incorporant des histogrammes, treemaps, cartes interactives en rendu "heatmap" ou interrogeables jusqu'à la section cadastrale.

Voici une illustration montrant 4 volets du tableau de bord (pour des raisons de droit de diffusion les données sont floutées) :



Le rendu a été fourni dans deux versions : une version de travail pour la métropole et ses partenaires, et une version diffusable respectant le secret statistique.

Pour chaque indicateur calculé, il est possible de télécharger un fichier excel contenant les résultats relatifs à l'indicateur.

La chaîne de production du tableau de bord est automatisée de bout en bout, car l'objectif était de reproduire par simple clic le rapportage sur un nouveau territoire.

Rmarkdown pour composer facilement des tableaux de bord

Pour composer cette visualisation des données, seules des technologies libres ont été utilisées. Les données DV3F sont stockées nativement dans une base de données PostgreSQL avec une cartouche spatiale PostGIS (pour, notamment, le stockage des géométries des parcelles). Le tableau de bord, lui, s'est appuyé sur R, en particulier le langage Rmarkdown et une librairie appelée flexdashboard.

Le langage Markdown est une syntaxe d'écriture de page créée par John Gruber et Aaron Swartz, dans le courant de l'Open Knowledge. Très simple et légère, elle se rapproche de l'écriture naturelle. Un même fichier markdown (extension .md) permet de générer des pages en beaucoup de formats différents : OpenOffice mais aussi HTML.

Par exemple, quelque chose d'aussi simple que ceci :

```
#Un titre

- Le contenu de ma première puc
- Le contenu de ma seconde puce
- Et encore une autre puce
- [Site du CEREMA](https://www.cerema.fr/fr)
```

Sera traduit en HTML, par quelque chose d'assez complexe :

```
<h1>Un titre</h1>

<ul>

<li>Le contenu de ma première puce</li>

<li>Le contenu de ma seconde puce</li>

<li>Et encore une autre puce</li>

<li><a href=https://www.cerema.fr/fr>Site du
CEREMA</a></li>

</ul>
```

Pour un rendu visuel :

Un titre

- Le contenu de ma première puce
- Le contenu de ma seconde puce
- Et encore une autre puce
- Site du Cerema

Comme on le voit, il est plus immédiat et rapide d'écrire en markdown qu'en HTML.

R est un logiciel OpenSource qui permet de faire des statistiques ainsi que de réaliser des graphiques.

R Markdown est une implémentation de Markdown qui permet d'inclure dans une page Markdown le résultat d'une exécution de code R. Ce résultat peut

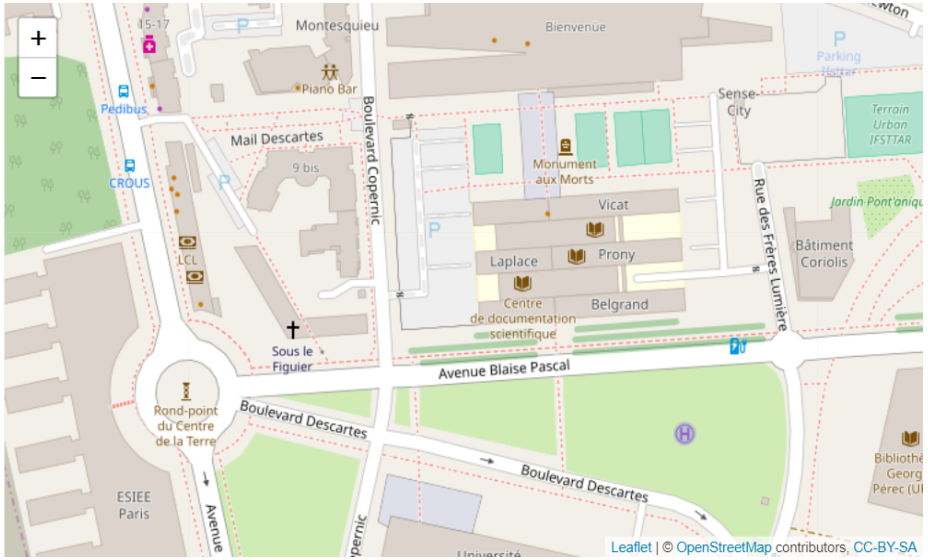
être un tableau de données, un graphique, ou une carte, y compris web et interactif. Il est très apprécié par les chercheurs qui souhaitent rendre leurs résultats reproductibles.

- Par exemple pour l'affichage d'une carte avec la librairie Leaflet :

```
library(leaflet)

leaflet() %>% addTiles() %>% setView(2.58694, 48.8412108,
zoom = 17)
```

Se traduira par du code HTML permettant l'affichage d'une carte leaflet :



Le langage Markdown ne permet pas de structurer une page en onglets, en blocs, et en pages, comme l'est une page web classique. C'est là qu'intervient la librairie flexdashboard. Cette dernière introduit une écriture qui permet de structurer l'information sur une page.

Par exemple :

```
---
title: "Tabset Column"
output: flexdashboard::flex_dashboard
---

Column
-------------------------------------
### Chart 1
```{r}```

Column {.tabset}
-------------------------------------
### Chart 2
```{r}```

## Chart 3
```{r}```
```

Sera traduira visuellement par :


```

R, Markdown et flexdashboard permettent donc, d'une certaine façon, de donner un contexte et d'éditorialiser l'information statistique.

Vers une meilleure visualisation des données foncières

Le tableau de bord DV3F réalisé pour le compte de la Métropole Nice Côte d'Azur rentre dans le cadre d'un nouveau type de prestations à forte valeur ajoutée pouvant être réalisé pour le compte des collectivités.

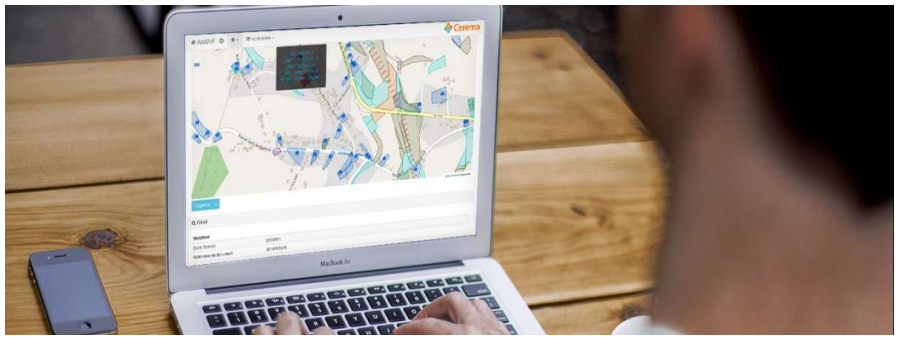
Le tableau de bord, par sa présentation synthétique d'indicateurs pertinents et de graphiques parlants, permet aux décideurs d'asseoir leur analyse et d'échanger autour des dynamiques foncières sur leur territoire.

Fort de cette expérience, le pôle Géomatique du Cerema Med retranscrit maintenant ce mode de rendu à d'autres thématiques, comme la biodiversité.

Auteur : Mathieu Rajerison (Cerema Méditerranée)

SIGN@TURE Lettre de l'Information Géographique
Cerema

Nouveau : le GT qualification des données géographiques au CNIG !



Un nouveau groupe de travail du CNIG s'est réuni le 7 mars dernier : celui sur la qualification des données géographiques.

La mesure de la qualité des données géographiques, et la diffusion des résultats de cette mesure, font l'objet de travaux depuis plusieurs années. Le séminaire QuaDoGeo co-organisé par le Cerema et le CRIGE PACA, en février 2018, a réuni les acteurs intéressés par le sujet pour constituer un plan d'actions.

Dans la continuité de cette dynamique, le CNIG a constitué un groupe de travail "Qualification des données géographiques", rattaché à sa commission Règles de Mise en Oeuvre, et animé par le Cerema. Ce groupe a rassemblé en premier lieu les participants au séminaire de 2018, mais il s'ouvre bien sûr à toute personne intéressée. Sa mission est d'**aider à la réutilisation des données ouvertes grâce à l'information sur la qualité**.

L'une des ambitions de ce groupe est la prise en compte du « retour utilisateur », qui n'est pour l'instant pas encadrée (la norme ISO 19157, qui porte sur la qualification, traite en effet la qualité interne des jeux de données), mais qui joue pourtant un rôle fort dans l'évaluation des données.

Deux actions concrètes réalisables à horizon d'un an sont prévues dans le mandat :

- La confection d'un **« synopsis pratique pour la qualification »** : il s'agirait, face à un type de données, de déterminer une méthode et un ordre des critères à mesurer ;
- La réalisation d'**une première fiche qui validera cette méthode**. Le groupe de travail déterminera le processus à mettre en place : type de fiche, choix éventuel d'un lot de données test, etc.

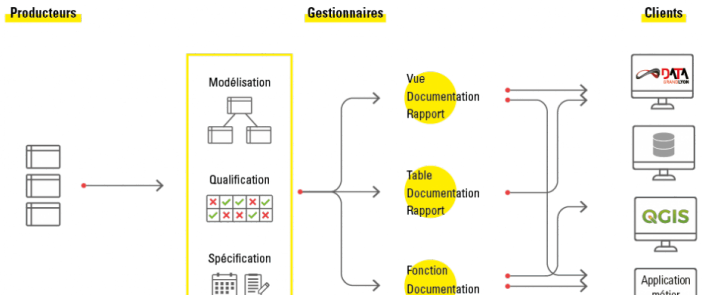
Voir le mandat du Groupe sur le site du CNIG

Retrouvez QuaDoGeo sur Twitter

Auteur : Stéphane Lévèque (Cerema Territoires et Ville)

SIGN@TURE Lettre de l'Information Géographique
Cerema

La qualité des données géographiques à la Métropole de Lyon



Dans le cadre de la politique OpenData de la Métropole de Lyon, la qualité des données est un enjeu important, principalement pour les réutilisations.

Cet article est issu de l'interview de Clément Jamet, Ingénieur données à la Métropole de Lyon. Clément Jamet fait partie de l'Unité Données de références et 3D, au sein du Service Géomatique et Données Métropolitaines, qui se charge de la gestion de la donnée brute et de son parcours jusqu'à sa diffusion (OpenData, intégration à d'autres bases de données…).

D'autres unités de la Métropole travaillent sur les données :

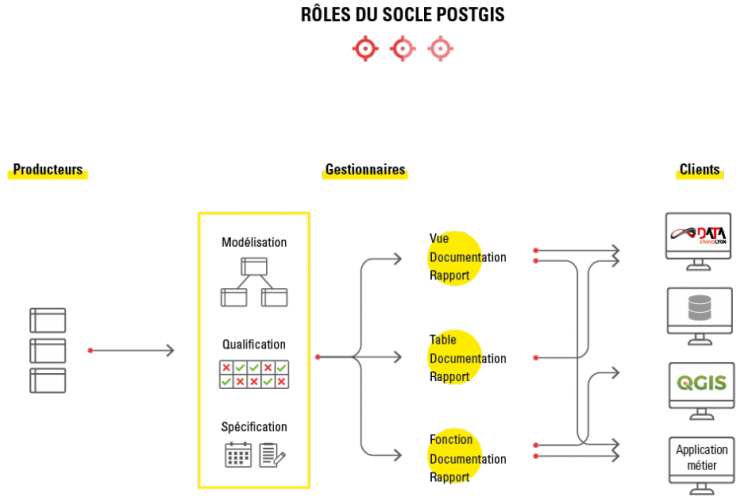
- *l'Unité de Topographie et de délimitation du domaine public ;*
- *l'Unité Diffusion des données (qui s'occupe notamment du portail OpenData).*

Dans le cadre de la politique OpenData de la Métropole de Lyon, la qualité des données est un enjeu important, principalement pour les réutilisations. Clément Jamet revient sur ce qui est mis en place à la Métropole : les enjeux autour de la qualité, et les grands chantiers sur le sujet. Pour illustrer les propos, nous effectuerons ensuite un zoom sur la gestion des données adresse.

Quels sont les enjeux de la qualité des données pour la Métropole de Lyon ?

Il y a deux enjeux majeurs autour de la qualité des données géographiques :

- Le premier est d'assurer un niveau de qualité dans la production des données. Les objectifs sont à la fois de fournir des données utiles, de constituer des socles de références, ainsi que d'être légitime pour alimenter des standards nationaux.
- Le deuxième enjeu est la qualification : pouvoir donner l'information sur la qualité des données produites, notamment pour faciliter les réutilisations, en interne comme par le grand public via les données ouvertes. La qualification est également indispensable pour l'alimentation de bases de données au niveau national.



L'automatisation des process de qualification est un sujet d'actualité. Il est important de pouvoir intégrer à l'exploitation de la donnée des informations permettant de calculer un indice d'incertitude. On peut alors prendre la valeur d'incertitude basse ou haute selon le besoin de tolérance. Dans ce cadre, les fiches du Cerema aident à consolider certaines procédures qui permettent de qualifier la donnée.

SIGN@TURE Lettre de l'Information Géographique
Cerema

Dans le contexte de l'analyse de gros volumes de données, il est important de standardiser cette notion de qualité, pour qu'elle soit intégrée dans des process pour hiérarchiser l'incertitude des données. Au-delà de la méthode proposée par le Cerema, l'un des besoins est donc de standardiser informatiquement les critères qualité.

De même, il est utile de publier les données dans des formats standards (CityGML, Base Adresse Locale – BAL – d'OpenDataFrance, GML, PCRS, INSPIRE), car on trouve des outils qui permettent de valider ou de contrôler la qualité (valid3city, data.gouv.fr ou le BAL). Cela permet de respecter un niveau de qualité et de mutualiser les outils de contrôle.

Quels sont les grands chantiers en cours ?

L'intégration de la mise en qualité des données dès le process de production est un premier chantier. Par ce biais, on peut, dès la production des données, vérifier notamment les critères de qualité géométrique et de complétion attributaire. Cela se fait grâce à la mise en place de contrôles automatiques dans les bases de saisies et dans les environnements de recette et de production.Au-delà de l'action sur la production, la qualité est intégrée à l'ensemble du parcours de la donnée. Dans le process de migration de la donnée jusqu'au portail OpenData, on contrôle le niveau de qualité, ce qui permet d'alimenter les métadonnées et aussi d'avoir des seuils qui peuvent bloquer la propagation de la donnée. Afin de pouvoir élargir à l'ensemble des données cette procédure, aujourd'hui restreinte à quelques jeux de données spécifiques, nous devons définir, pour chaque donnée produite, des critères et des seuils acceptables. Ces seuils d'acceptabilité dépendent de l'usage : ils seront plus bas pour des usages internes de données peu sensibles, et plus haut pour la diffusion en OpenData des données destinées à l'alimentation de base de données nationales.

Si les process de contrôle qualité commencent à être en place, ils ne permettent pas la restitution des métadonnées « qualité ». Aujourd'hui, l'un des objectifs est de générer en sortie une fiche qualité. L'effort est aussi porté, pour les données OpenData, sur le renseignement de la généalogie, en attendant que des critères chiffrés et plus complets soient fournis.

Enfin, dans le cadre de la réception de l'Ortho 2018, le contrôle de la qualité de la livraison servira à fournir des informations qualitatives pour les métadonnées

Produire des données de référence : zoom sur la BD Adresse

L'enjeu : assurer la qualité d'une donnée de référence

En analysant la qualité de sa Base de données Adresse, la Métropole s'est aperçue que le taux d'exhaustivité n'était pas acceptable. Or dans la logique d'en faire la source de données de référence sur le territoire, il a donc fallu mettre en place un contrôle qualité.

La première étape a été de qualifier les manques pour aider les producteurs à mieux produire. Pour cela, le recours à des bases de données externes – elles-mêmes qualifiées – a permis d'analyser les différences. Par exemple, la base des propriétaires de locaux de la DGFip, avec une bonne exhaustivité, permettait de recenser les données manquantes dans la BD Adresse.

D'autre part, dans la Base Adresse de la Métropole, chaque adresse est rattachée à une parcelle. Cette dimension est importante, car la Métropole fournit un service automatique de certificat d'adressage pour les notaires : cet outil applicatif nécessite des données de qualité forte en entrée. Il était donc aussi nécessaire de contrôler ce rattachement.

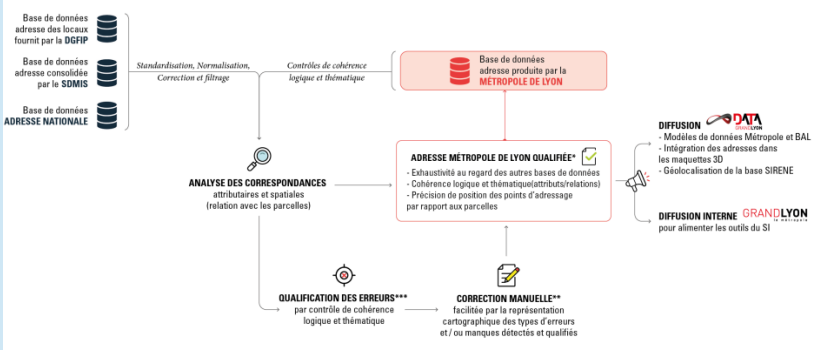
Application de la méthode d'échantillonnage

Sur des données de référence (comme l'adresse), il est difficile de trouver un référentiel pour contrôler la qualité. Suite à une tentative peu concluante de construire un autre référentiel avec différentes sources, le choix s'est alors orienté vers l'échantillonnage et le contrôle terrain. Les fiches Cerema ont été utiles pour aider à structurer le process d'échantillonnage et de contrôle.

Nous sommes en train d'organiser le contrôle sur le terrain par l'unité de Topographie et de délimitation du domaine public de la Métropole. L'échantillon

nage est alors « opportuniste » dans le sens où les lieux de contrôle sont guidés par l'activité des topographes, avec une consolidation pour avoir un échantillon qui réponde aux besoins statistiques. Il est ainsi possible de s'appuyer sur des ressources qui vont sur le terrain pour mutualiser le travail.

Comment définir la donnée de référence ?



Il faut identifier les différences entre une réalité terrain et une réalité officielle, juridique ou de spécification de saisie. Par exemple, à Lyon, tout le monde (même la plaque de rue ou les informations fiscales) parle de la « montée de Choulans », alors que la délibération de création de la voie la nomme « chemin de Choulans ». Le contrôle qualité fait un retour erreur à chaque fois, mais officiellement ce n'est pas une erreur.

Le contrôle qualité pose donc la question : faut-il extraire les erreurs que l'on connaît historiquement (dans une base de données de faux positifs) ou alors diffuser une donnée complète qui contient aussi les noms usuels (c'est-à-dire adapter la base aux habitudes d'usages) ?

En conclusion, le travail pour qualifier une donnée de référence est assez lourd, et nécessite une forte connaissance du territoire.

La qualité des données géographiques en 120 secondes, par le CRIGE PACA

Découvrez une vidéo courte et humoristique pour faire comprendre la qualité des données géographique.

Le CRIGE PACA réalise des actions d'acculturation (journées techniques), de communication, et d'assistance au sujet de la qualité des données géographiques. Dans ce cadre, il a réalisé une vidéo pédagogique et courte, qui vise à rendre accessible ce sujet qui peine à sortir du cercle des spécialistes, et qui pourtant à de nombreux impacts en matière de diffusion et de réutilisation des données.

Lien vers la vidéo (sous YouTube)

En savoir plus sur les travaux du CRIGE PACA sur la qualité des données

SIGN@TURE Lettre de l'Information Géographique
Cerema

Qualité des données, Machine Learning : quelques propositions de lecture

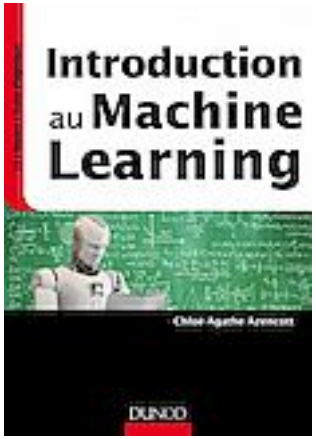
Qualité de l'information géographique

Une parution qui a déjà 4 ans mais qui est toujours d'actualité, sur la qualité de l'information géographique et des données associées.

Sous la direction de Rodolphe Devillers (Department of Geography, Memorial University of Newfoundland) et Robert Jeansoulin (Université Paris-Est)

Chez Hermes / Lavoisier ; collection : Traité IGAT ; octobre 2015 (voir sur le site de l'éditeur)

Introduction au Machine Learning

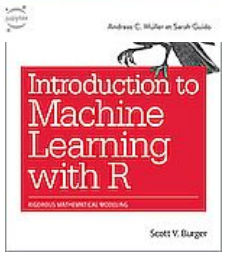


Une introduction au Machine Learning pour savoir quand et comment utiliser le Machine Learning : quels algorithmes utiliser suivant les besoins et comment les mettre en oeuvre.

De Chloé-Agathe Azencott (Maître de conférence à MINES ParisTech et enseignante à CentraleSup)

Chez Dunot ; collection : InfoSup ; septembre 2018 (voir sur le site de l'éditeur)

Machine Learning avec Python ou R



Machine Learning - PyhtonMachine Learning - R L'édition O'Reilly, spécialisée dans l'informatique, propose de se lancer dans le Machine Learning. Faites vos choix : R ou Python ?

"Le machine Learning avec Python" : de Sarah Guido et Andreas C. Muller : février 2018

"Machine learning avec R" : de Scott Burger ; novembre 2018

POUR PLUS D'INFORMATION…

La revue électronique Sign@ture est publiée quadrimestriellement et traite selon son acronyme historique, de la Situation de l'Information Géographique Numérique dans l'Aménagement, les Transports, l'Urbanisme, les Réseaux et l'Environnement mais également d'autres domaines qu'il serait trop long d'énumérer. Elle est destinée à tous les acteurs qui y contribuent (publics, privés et associations). Chaque numéro comprend un dossier technique ou un point de vue qui traite soit des techniques géomatiques soit de l'usage de la géomatique dans l'un des domaines d'études précités ou pas.

https://www.cerema.fr/fr/centre-ressources/newsletters/signture

Directeur de la publication : Pascal Berteaud

Directeur délégué de publication : Christian Curé

Rédacteur en chef : Bernard Allouche

Equipe éditoriale : Antoine Lemot (Cerema Centre-Est), Stéphane Lévêque (Cerema Territoires et ville)

Vous souhaitez participer à la rédaction du prochain numéro de Sign@ture, car votre structure mène une démarche géomatique ou vous avez des événements à promouvoir ? Contactez-nous : revue.signature@cerema.fr