
Première évaluation de la qualité des données libres d'OpenStreetMap en France

Guillaume Touya, Jean-François Girres

Laboratoire COGIT, Institut Géographique National

73 avenue de Paris 94160 Saint-Mandé

{ prénom.nom}@ign.fr

RÉSUMÉ. Les nouvelles technologies du Web 2.0 ont permis l'émergence des contenus créés collaborativement dont l'exemple le plus intéressant est le projet OpenStreetMap. Ce projet vise à produire une base de données topographiques entièrement libre. Cet article étudie la qualité des données françaises de ce projet en prenant comme comparaison des données de l'Institut Géographique National. La qualité des données est très hétérogène du fait de l'absence de spécifications précises et acceptées.

ABSTRACT. New concepts like free data or Volunteered Geographic Information recently emerged thanks to new Web 2.0 technologies. The OpenStreetMap project is the most significant example. It aims at producing free vector topographic databases. This paper studies the OpenStreetMap french data quality compared to institutional databases of the french NMA. Data Quality is quite heterogeneous due to a lack of precise and well-accepted specifications.

MOTS-CLÉS : OpenStreetMap, qualité, VGI

KEYWORDS: OpenStreetMap, quality, VGI

1. Introduction

Avec l'avènement actuel du Web 2.0, les contributeurs ne se contentent pas de chercher du contenu mais ils en créent eux-mêmes comme le montre le succès des sites FaceBook, MySpace ou YouTube. Le projet OpenStreetMap permet lui de contribuer en construisant une base de données géographique. Ce modèle de production nouveau est appelé le *crowdsourcing* [TAP 07] ou *Volunteered Geographic Information* (VGI) [GOO 07] dans le cas de l'information géographique. Le travail présenté ici constitue une première expérience en terme d'évaluation de la qualité de données géographiques libres. Ce travail reprend l'idée de [HAK 09] qui évalue la qualité d'OpenStreetMap en comparaison des données de l'Ordnance Survey, l'IGN anglais, en l'appliquant aux données françaises et en étendant l'évaluation à d'autres composantes de la qualité.

2. Qu'est-ce qu'OpenStreetMap?

OpenStreetMap est un projet collaboratif tel Wikipedia initié en Angleterre en 2004, qui vise à créer et fournir des informations géographiques libres. Les données sont distribuées sous une licence qui permet d'utiliser les données de manière totalement libre à condition de distribuer librement toute donnée dérivée sous la même licence.

Les données déposées dans OSM par des contributeurs du projet sont modélisées et stockées sous la forme de primitives géométriques étiquetées. Par exemple une route est une polygone avec les étiquettes *highway* = "primary", *oneway* = "no" et *name* = "N10". Pour faciliter le travail du contributeur, il existe également des fiches conseils montrant comment étiqueter une situation donnée. Les données sont saisies à l'aide d'un logiciel type SIG adapté à OSM avec des fonctions d'éditations pour créer des primitives géométriques OSM et les étiqueter. Les données sont généralement issues de sources libres de droit comme des traces GPS personnelles ou des images satellites.

3. Evaluation de la qualité des données OSM

Afin d'évaluer la qualité des données géographiques offertes par OpenStreetMap, nous traitons les différentes composantes de cette qualité : précision géométrique, précision attributaire et sémantique, exhaustivité, cohérence logique, actualité et historique.

Pour estimer la précision géométrique, des comparaisons sont réalisées entre les données OSM et les données de la BD TOPO®, produite par l'IGN, de précision métrique. Les objets dits « homologues » sont sélectionnés et appariés manuellement et des mesures de distance sont calculées (distance de Hausdorff, distance surfacique, etc...). Les résultats, dont on peut voir un exemple en Figure 1 et Tableau 1, peuvent présenter des écarts très variables en fonction des thèmes comparés.

Par exemple, la comparaison des thèmes linéaires routiers montre des écarts maximaux moyens très supérieurs (Distance de Hausdorff moyenne de 13.57 mètres) à la précision de la base de données de référence (Erreur moyenne quadratique de 2 mètres) mais surtout une très forte hétérogénéité dans la distribution des valeurs, du fait du manque de spécifications de saisie précises.

L'étude des carrefours des thèmes routiers a permis d'estimer une erreur de position moyenne trois fois supérieure (distance moyenne de 6.65 mètres) à celle garantie par la BD TOPO®.

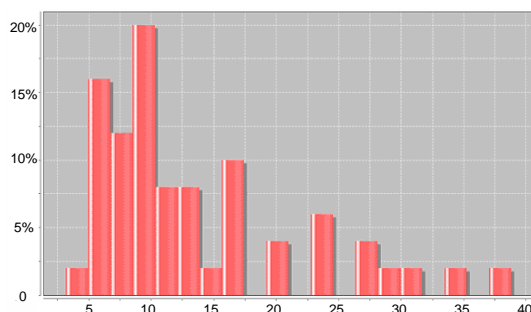


Figure 1. Distribution des distances de Hausdorff à partir de l'échantillon

Statistiques	Dist. de Hausdorff	Dist. moyenne
Maximum	38.8 m.	6.07 m.
Minimum	3.14 m.	0.14 m.
Moyenne	13.57 m.	2.19 m.
Ecart-type	8.32 m.	1.69 m.
Cœf. de variation	61.28 %	76.95%

Tableau 1. Statistiques sur les écarts linéaires

La précision attributive est aussi réduite par l'absence de spécifications précises avec peu de valeurs remplies pour les attributs secondaires (29%) et 40% d'erreurs pour les attributs principaux comme le nom des lacs ou la nature de routes. La précision sémantique dépend elle directement de la clarté des spécifications: les routes de nature "autoroute" ou "principale" sont presque toutes justes sémantiquement (comparaison automatique avec la BD TOPO®) mais les routes "résidentielles" ou tertiaires", plus floues, sont sémantiquement fausses à 50%.

Pour les autres composantes, nous avons particulièrement noté des problèmes de cohérence logique, notamment topologique, intra-thème (5% de routes mal connectées) et inter-thème (68% des communes testées non cohérentes avec les cours d'eau) qui sont handicapants pour des applications SIG classiques. La cohérence est très hétérogène du fait de l'absence de contraintes d'intégrité dans les spécifications.

Concernant l'exhaustivité, OSM en est encore loin avec en moyenne 10% des objets de la BD TOPO® mais la différence est moindre en longueur totale pour les routes (les petites routes sont donc moins saisies). On remarque que les zones exclues (Creuse) sont moins bien couvertes que les zones favorisées (grandes villes).

Concernant l'actualité, nous avons noté un renouvellement de 30 % des objets en trois mois qui montre une forte activité des contributeurs, mais ne garantit pas une mise à jour de la base de données. Enfin, l'historique des données est réduit du fait de métadonnées succinctes.

4. Conclusion et perspectives

En conclusion, nous pensons que le manque de spécifications nuit fortement à la qualité globale des données OSM et limite leurs utilisations. Le COGIT étudie des méthodes permettant de définir des spécifications au sein de systèmes de saisie collaborative.

Bibliographie

- [GOO 07] Goodchild M. F., " Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0 ". International Journal of Spatial Data Infrastructures Research., vol. 2, 2007, p. 24-32.
- [HAK 09] Haklay M., " How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England ". Environment & Planning B, to be published.
- [TAP 07] Tapscott D., Williams A. D., " Wikinomics : Wikipédia, Linux, YouTube... Comment l'intelligence collaborative bouleverse l'économie ". Pearson Education, 2007.