LA QUALITÉ DES DONNÉES GÉOGRAPHIQUES

par Henri Pornon

1) Quelques lieux communs sur la qualité

"Des données de qualité sont indispensables à la réussite d'un projet de SIG".

"Des données précises sont des données de qualité".

"La qualité ça coûte cher".

"Des données de qualité, ça se paye".

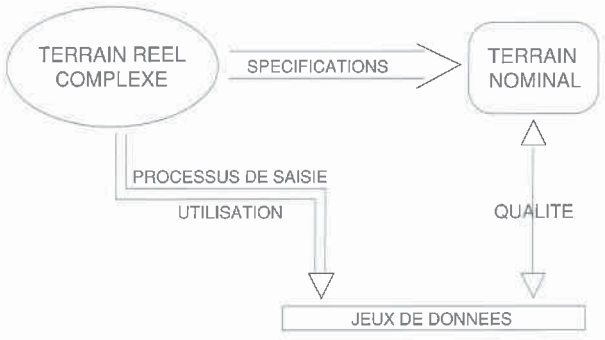
Ces quelques lieux communs montrent que la notion de qualité est un concept utilisé couramment dans le domaine des données géographiques, même si elle est difficile à cerner. Ils mettent également en évidence le fait qu'une clarification est nécessaire.

2) Définition de la qualité

La norme internationale ISO 8402 qui précise le vocabulaire de la qualité définit la qualité de la façon suivante : "ensemble des propriétés et caractéristiques d'un produit ou service qui lui confèrent l'aptitude à satisfaire des besoins exprimés ou implicites".

Cette définition se transpose facilement au domaine des données géographiques. Les besoins des utilisateurs permettent de définir des spécifications qui serviront de fondement à la définition de la qualité d'un jeu de données.

Le graphique suivant (extrait d'un schéma de la norme française EDIGEO pour l'échange de données géographiques) illustre cette notion :



La notion de qualité ne s'applique pas à la comparaison du terrain réel (trop complexe et trop dense) avec un jeu (ou une base) de données, mais à la comparaison du jeu de données avec le terrain nominal, qui est la vision qu'a l'utilisateur du terrain réel à travers ses besoins, traduits en spécifications.

Ainsi, les rues d'une ville seront traduites en des terrains nominaux différents par des utilisateurs de SIG différents. La société de livraison à domicile souhaitera une représentation filaire sous forme de graphe topologique, de précision ordinaire (métrique) mais incorporant toutes les rues. Le service municipal chargé de la délimitation domaine public/domaine privé souhaitera une représentation surfacique des limites d'îlots avec une précision décimétrique. Le transporteur national souhaitera une représentation filaire des principaux axes de trafic et des zones industrielles de la ville.

Un plan topographique ou photogrammétrique satisfera les exigences de qualité du deuxième cas, mais pas du premier ni du troisième.

On remarquera enfin que dans la définition de la norme, les besoins peuvent être implicites ou explicites. on peut effectivement considérer qu'un utilisateur qui commande un plan topographique s'attend implicitement à un certain résultat : précision de lever, symbolisme de représentation, choix des objets à représenter.

Pour une carte routière numérique, il n'a pas non plus besoin d'expliciter le fait que celle-ci doit comporter des villes reliées par des routes, dans une organisation permettant la recherche d'itinéraires.

Cependant, dans ces deux cas, certains aspects des besoins ne sont pas forcément implicites : pour le plan topographique, quelle organisation en couches est requise ? Pour la carte routière, combien de routes et de villes sont nécessaires. Toutes les communes, tous les hameaux ? Toutes les petites routes ou seulement les principales ?

3) Du terrain réel au terrain nominal ?

Que définit l'utilisateur quand il spécifie son terrain nominal ?

• Le système de coordonnées de référence (Lambert II et NGF)

• La précision géométrique de la planimétrie et de l'altimétrie (planimétrie en précision décamétrique, altimétrique en précision métrique)

• Le territoire concerné (le bassin versant de la Seine)

• Le choix des objets à saisir (les cours d'eau, les limites communales, les installations polluantes, les stations d'épuration, les points de mesure de pollution et de débit).

• Pour chaque objet :

- Le mode de sélection : les cours d'eau de longueur supérieure à 1 Km, toutes les stations d'épuration, les points de mesure gérés par l'Agence...),

- L'organisation graphique : les limites communales en polygones fermés, les cours d'eau sous forme d'un graphe arborescent...,

- Les attributs attendus (nom et population de la commune...),

- Le type de primitive graphique (station d'épuration par des points, zones d'épandage par des polygones...) et le choix de modèles mathématiques (courbes lissées polynomiales ou non...).

• Les relations entre objets (exemple : relation station d'épuration rejette son eau dans cours d'eau).

• etc...

4) Du terrain réel à la base de données

Le producteur de données va s'efforcer de mettre en œuvre un processus de production capable de satisfaire les exigences de l'utilisateur pour fournir un produit répondant aux spécifications.

Les lots de données qu'il va livrer vont être comparés aux critères énoncés dans le paragraphe précédent, mais aussi à d'autres critères de qualité relatifs aux données et à la prestation.

4.1) Critères de qualité relatifs à la prestation

• planning d'exécution et de livraison,

• type d'édition graphique à fournir,

• seuils de conformité des données avec les spécifications.

L'utilisateur définira par exemple comme acceptable les cartes numérisées dans lesquelles 95% des cours d'eau de plus de 1 Km auront été saisis.

• format des fichiers à livrer,

• mode de réalisation de la prestation : l'utilisateur exigera par exemple certains matériels, logiciels ou périphériques ou certaines technologies ou méthodes (vectorisation semi-automatique plutôt que digitalisation). Le calage entre elles pour assurer la continuité de la base de données devra être réalisé d'une certaine façon. Les choix à opérer quand il existe des sources multiples pourront aussi être prédéfinis.

4.2) Critères de qualité sur les données concernées par la prestation

• Validité dans le temps des informations. L'utilisateur peut souhaiter savoir, par exemple dans le cas d'un lever topographique ou photogrammétrique, la date de réalisation du lever.

• Origine des informations. Dans le cas où le producteur mélange plusieurs sources, l'origine de chaque information peut devoir être fournie avec le lot de données.

La question se pose alors de savoir comment comparer le lot de données au terrain nominal (rappelons que la comparaison avec le terrain réel n'a pas de significations, même si le contrôle peut s'exercer en visitant le terrain réel).

5) Le contrôle du lot de données

5.1) Pas de contrôle

L'utilisateur court tout simplement le risque de disposer d'une base incohérente et inutilisable. C'est par exemple le cas de collectivités qui font réaliser une digitalisation du fonds de plan cadastral sans contrôle ni spécifications et découvrent, à la première tentative de réaliser une carte thématique à la parcelle, qu'il s'agit d'un document non structuré contenant tous les traits et symboles du document papier initial, mais ne disposant ni de parcelles, ni des îlots, ni des bâtiments...

5.2) Démarche traditionnelle : contrôle du produit

La production de données géographiques est encore très souvent un processus artisanal. Le producteur réalise la prestation demandée par son client et réalise parfois mais pas toujours un contrôle avant livraison. Le client contrôle le lot de données, identifie des erreurs, renvoie le lot des données pour corrections. Le producteur corrige et renvoie les données corrigées. Le client contrôle à nouveau...

Dans un grand nombre de cas, ce type de contrôle coûte très cher aux deux acteurs. Les corrections vont réduire la marge du producteur et dans certains cas rendre la prestation non rentable. Chacun sait que la modification des données géographiques coûte beaucoup plus cher que leur création ou numérisation. Le contrôle coûte également cher à l'utilisateur, qui va être obligé de consacrer beaucoup de temps à dépister les erreurs contenues dans les lots de données.

La qualité du résultat final est par ailleurs difficile à garantir car elle dépend de la quantité de contrôles exercés par l'utilisateur et du sérieux des corrections ou de la prestation initiale du producteur.

5.3) Démarche qualité de premier niveau : contrôle de processus

De plus en plus de producteurs de données géographiques, notamment ceux dont l'approche est plus industrielle qu'artisanale, incorporent au processus de production les contrôles qui leur permettront de connaître avant livraison la valeur du produit livré et de corriger les erreurs au fur et à mesure.

Ceci a pour conséquence que les erreurs détectées par l'utilisateur au contrôle sont moins nombreuses et que le taux de retour de lots de données pour correction est beaucoup plus faible. Le producteur en tire donc un double avantage, économique (rentabilité de la production) et stratégique (satisfaction et confiance du client).

Le client, s'il a confiance dans le processus de production du producteur, peut également en tirer un double avantage : il réduit le nombre de contrôles et disposera de données d'une qualité mieux maîtrisée.

5.4) Démarche qualité de deuxième niveau : management de la qualité

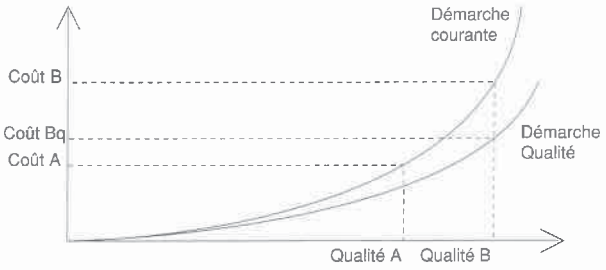
Le producteur, qui ne se contente pas d'intégrer des contrôles à son processus de production, mais met en place une organisation du personnel et des méthodes de travail orientées vers la recherche de la qualité et peut faire la preuve à son client de cette prise en compte de la qualité, va offrir des garanties à son client. Celui-ci aura l'assurance que le producteur contrôle sa prestation et pourra donc limiter son propre contrôle.

C'est l'objectif des démarches de certification dans le cadre des normes ISO 9000.

Le producteur de données, après avoir mis en œuvre une démarche qualité interne, fera certifier sa démarche par un organisme certificateur (en France, l'AFAQ).

5.5) La qualité, une démarche économique

Le graphique suivant illustre que le débat sur la qualité est avant tout économique. A niveaux de qualité différents (A<B) correspondent des coûts variant dans le même sens (coût A>coût B). En revanche, pour un niveau de qualité donné (Qualité B), la production de données dans le cadre d'une démarche qualité coûte moins cher (coût Bq<coût B). Le débat sur la qualité contient souvent la confusion entre ces deux aspects, entretenue par la difficulté qu'ont les utilisateurs à définir explicitement leurs besoins. Il ne faut donc pas confondre le premier aspect ("Il y a mieux, mais c'est plus cher") avec le second ("Si vous savez ce que vous voulez, la démarche qualité coûte moins cher à résultat identique et garantit mieux le résultat").



6) La mesure de la qualité sur le lot des données

Si le producteur est certifié ou peut justifier d'un plan d'assurance qualité, le client va limiter les contrôles à quelques sondage ou va faire totalement confiance au producteur. Dans le cas inverse, le client va chercher à mesurer la qualité du lot de données.

6.1) Mesure par la manipulation des données

Le contrôle du lot de données peut s'avérer être une opération longue et délicate, si le client souhaite vérifier un grand nombre de critères. La vérification de la concordance du lot de données avec le terrain nominal peut nécessiter de nombreuses manipulations sur les données : superposition avec d'autres documents graphiques, mesures de contrôle sur le terrain, exécution de requêtes dans la base de données, calculs particuliers ...

6.2) Fourniture de l'information sur la qualité avec le lot de données

C'est le principe de la boîte de petits pois. Il ne viendrait à personne d'ouvrir toutes les boîtes de petits pois extra-fins avant l'achat pour vérifier que les petits pois sont réellement extra-fins. C'est écrit sur la boîte.

On peut espérer que dans une avenir proche, il sera possible de connaître la qualité d'un lot de données par simple lecture d'une série de caractéristiques dans l'entête du fichier. Ceci suppose deux conditions : que les valeurs fournies par le producteur inspirent confiance (intérêt d'une certification qualité) et que les critères et leurs valeurs soient normalisés.

La norme d'échange de données géographiques EDIGEO, utilisée en France et proposée au niveau européen, permet de fournir neuf critères de qualité dans un échange : généalogie, actualité, précision planimétrique, précision altimétrique, précision métrique, exhaustivité, précision sémantique, cohérence logique, qualité spécifique.

7) La qualité, une évidence ?

Au-delà de l'évidence pour chaque producteur de souhaiter fournir des lots de données de qualité et pour chaque utilisateur d'exploiter une base de données de qualité, on peut constater que le débat sur la qualité comporte un certain nombre d'exigences qui représentent un bouleversement culturel pour chacun des interlocuteurs.

Pour l'utilisateur, s'engager dans la recherche de données de qualité suppose de pouvoir clairement exprimer ses besoins au producteur et donc de savoir ce qu'il veut et comment il le veut. Cela suppose également d'établir une relation différente avec le producteur, non plus basées sur la recherche du plus bas prix, mais sur la confiance.

Pour le producteur, il s'agit de réorganiser complètement le processus de production pour définir correctement les phases automatisées, les phases manuelles et les contrôles internes. Il s'agit de remplacer aussi une perception du client ("j'ai eu le marché à très bas prix, il en aura pour son argent") par une autre ("je cherche le meilleur moyen de satisfaire ses exigences et de lui faire la preuve de mon aptitude").

On voit bien qu'il faut des deux côtés une évolution à la fois technique et culturelle. Cette évolution est cependant nécessaire pour améliorer la qualité des bases de données.