

Neogeographic data quality – is it an issue?

Andrew Coote and Les Rackham, ConsultingWhere Ltd

An updated version of the paper given at the AGI Conference in September 2008

© 2008 by ConsultingWhere Ltd

1. Background

The Information Technology world, of which geographical information systems (GIS) are but a small component, is going through another seismic upheaval. The emergence of mass online collaboration, through what is loosely described as Web 2.0, is changing the way we think about knowledge and introducing challenging new business models. The growth of Wikipedia into a vast encyclopaedia of online knowledge is just one manifestation of the changes to the way we create, manage and disseminate knowledge. Google, with its business model based on advertising, has already transformed the landscape of the media industry. The advent of transactional charging for data and software usage will have an equally fundamental affect upon the systems and services sector.

Closer to home these “paradigm shifts” are manifested in the arrival of consumer mapping sites such as Google Maps and Virtual Earth, innovative applications through “mash ups” and “open source” datasets such as OpenStreetMap and new communities of data authors and users. In recent times, these inter-related themes have come to be referred to as “neogeography”.

At the AGI conference in 2007, a debate entitled, “Will neogeography kill GIS?” was well attended and enthusiastically contested. For those of us who have been around the industry for a while and have lived through various “paradigm shifts” observe that there are some underpinning principles that have been important throughout. Two of these principles are to i) understand the users' requirements and ii) be able to assess the “fitness for purpose” of data and systems in that context.

This paper looks at these principles in respect to neogeographic data and contrasts them with some examples from more “conventional” geographic datasets.

2. Objectives

The objectives of this paper, ordered to reflect its structure, are to:

- Establish a common language for discourse by examining the terminology used in this emerging area;
- Examine the characteristics of neogeographic and more conventional datasets;
- Classify potential users and outline their generic requirements;
- Make the case for assessing data quality prior to use of any dataset;

- Describe the criteria that characterise data quality;
- Apply these criteria to exemplar neogeographic datasets and some “conventional” commercial equivalents;
- Draw some tentative conclusions to stimulate debate

3. Terminology

It is always important to agree your frame of reference for any discussion. So, let us start with the term “neogeography” which has variously been defined as:

“Neogeography is about people using and creating their own maps, on their own terms and by combining elements of an existing toolset.”

Turner, Introduction to Neogeography (2006)¹

“Literally means ‘new geography’, and is commonly applied to the usage of geographical techniques and tools used for personal and community activities or for utilization by a non-expert group of users. Application domains of neogeography are typically not formal or analytical.”

Wikipedia (August 2008)

This in itself illustrates one of the challenges of the subject area. The definition of what neogeography covers is in a state of flux. However, this is not surprising in a subject area which is still in the early stages of development.

For the purposes of this discussion we are focusing on that aspect of neogeography whereby datasets are created and maintained by local communities and special interest groups. Goodchild² refers to this as Volunteered Geographic Information (VGI) and in the Web 2.0 world it is called “crowd sourcing”. The products of such activity we refer to as neogeographic datasets.

The other relevant term in vogue currently is “cloud computing”, the idea that software can be delivered over the web as a service using a shared technology infrastructure. Cloud computing and Software as a Service (SaaS) are similar concepts if not synonymous.

¹ Introduction to Neogeography, Andrew Turner, O’Reilly Media Short Cuts Series (2006)

² Michael F Goodchild, Citizens as Voluntary Sensors: Spatial data Infrastructure in the World of Web 2.0, (2008)

4. Characteristics of Geographic Datasets

Here are some of the characteristics that can be used to distinguish between neogeographic and conventional datasets.

4.1 Neogeographic Datasets

The following can be recognised in most of these products:

- Creation has been stimulated by lack of available data or by frustrations with costs, restrictions and limitations of existing conventional data sources;
- Collaborative: involving the capture, processing and dissemination of geographic information provided voluntarily by individuals;
- Approaches to creation and management are intuitive and not necessarily tied to accepted standards or methodologies;
- Data is licensed using some open source approach, which allows for users to consume the data without charge provided the original creator is acknowledged and any other user can do the same with anything you produce.³

Examples used in this discussion are:

i) OpenStreetMap – to quote from their website:

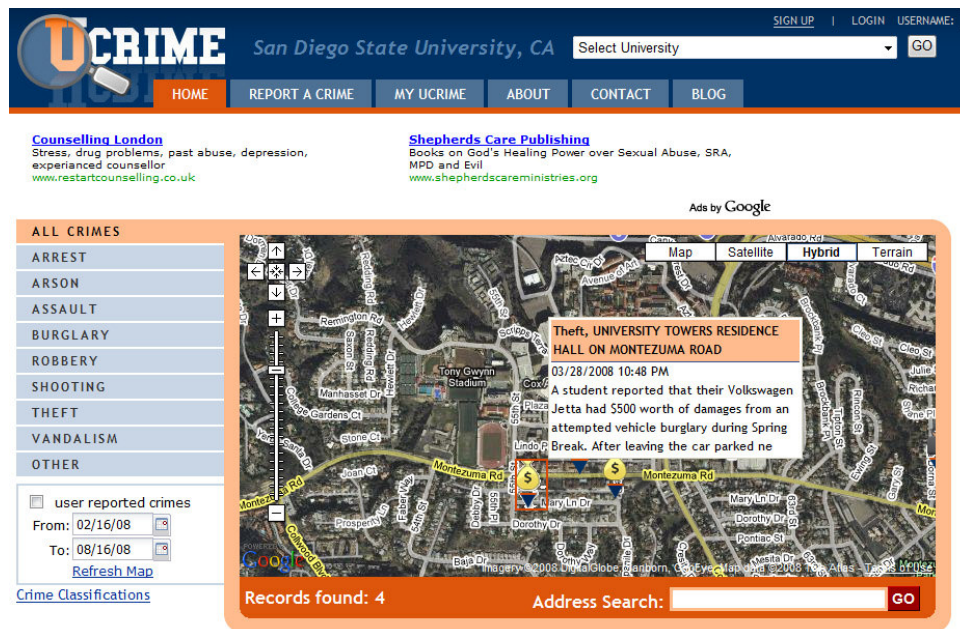
"OpenStreetMap is a project aimed squarely at creating and providing free geographic data such as street maps to anyone who wants them. Contributors to OpenStreetMap take handheld GPS devices with them on journeys, or go out especially to record GPS tracks. They record street names, village names and other features using notebooks, digital cameras, and voice-recorders. Back at the computer, contributors upload those GPS logs showing where they travelled, and trace-out the roads on OpenStreetMap's collaborative database.

"That data is then processed to produce detailed street-level maps, which can be published freely on sites such as Wikipedia, used to create handheld or in-car navigation devices, or printed and copied without restriction."

³ One open source approach to licensing is creative commons, see <http://creativecommons.org/licenses/by-sa/2.0/> for an example legal definition.

ii) UCrime – again quoting from the website:

“UCrime.com is an online service to provide easy to read crime maps and automated alerts to those who do or plan to attend, work for, have children in and live around colleges and universities. The service is free to members of the public. Users can sign up to receive alerts automatically via email if a crime occurs near a selected school or schools. Information on UCrime.com is obtained through police departments, daily newspapers and other publicly available sources. The site incorporates "Web 2.0" technology to enable users to comment on particular crimes. In addition to knowing what is happening where, users are empowered to provide tips and ideas to help solve crimes and improve public safety."



(© 2008 Google - Imagery © 2008 Digital Globe, Geoeeye; Map data © 2008 Tele-Atlas)

Figure 1: UCrime – University of San Diego

UCrime uses Google maps as its spatial base but is community focused, in that it encourages data creation and enhancement by local people for a shared purpose, as illustrated in Figure 1 above.

4.2 Conventional Datasets

What we might describe as "conventional datasets" are characterised by being:

- Created for a specific and defined set of requirements whether for legal, administrative or commercial purposes. Depending on the context these may or may be freely available but usually there is, at least, some dissemination charge and most likely restrictions on access and use;
- Managed by organisations established for the purpose whether as public or commercial bodies. There may be collaboration between organizations but on the basis of legal agreements including commercial contracts;
- Collected by professional staff who are paid to do so;
- Based on established methods, standards, specifications and practices;
- Quality assured in varying degrees during the production of the data and some information, however basic, supplied on the quality of the data;
- Protected by some form of copyright and usage is governed by formal agreements or licences;
- Access limited, in many cases, to only certain organisations or individuals for reasons of security, data protection or commercial advantage.

The examples of such datasets cited here, are Teleatlas⁴ Street Centrelines as used in Google Maps and Ordnance Survey's MasterMap® topographic layer⁵. As these datasets are in relatively wide usage, readers are referred to the producers' website or literature if they require further information.

5. User Communities and their Requirements

At a generic level, we can think of users being of four basic types according to their requirements:

5.1 Consumers

A consumer is a person who purchases any product or service for personal use. Typically when business people and economists talk of *consumers* they are talking about *person as consumer*, a commodity item with little individuality other than that expressed in the buy/not-buy decision.⁶ When we make a decision to buy a SatNav to plan journeys we are acting as consumers.

⁴ <http://www.teleatlas.com/index.htm>

⁵ <http://www.ordnancesurvey.co.uk/mastermap/>

⁶ Wikipedia – August 2008

Consumers do not want to have to think about the product particularly. They want it to work – that is, provide reliable routes or clear maps. They have some other relatively simple requirements which they may not be able to articulate coherently, such as attribute accuracy. However, they have little loyalty to the product; they will desert it very quickly if it goes wrong and tell their friends!

5.2 Special Interest Groups (SIG)

Special Interest groups are individuals who come together to collaboratively achieve some shared goal. In this context, these individuals require datasets with regional or wider geographical extent. An example might be Sustrans⁷ an organisation which designs, creates and manages routes for cyclists, walkers and people with disabilities, covering the whole of the UK.

Special Interest Groups usually have focused requirements, based around a hobby or interest, whether that is Green issues or cycle routes. They are interested in completeness – they want to see all the cycle routes in the area they are visiting next weekend; rich attribution – where are pubs along the route? What are their opening times? And currency – has the track been affected by recent rainfall, so extra care is necessary in certain sections. Less concern is attached to positional accuracy and errors are tolerated, to some extent, – after all it is free and crowd sourced.

5.3 Local Communities

This covers local people who have a common desire to protect and improve their local area. Their demands and interests usually concern a relatively small, but often ill-defined, geography but have a much wider set of interests including history, conservation, local planning and crime prevention.

Community users are also tolerant of errors but expect them to be corrected relatively quickly using the “wisdom of the crowd”. Positional accuracy is important – do you want a burglary to be placed right on your house if in fact it just happens to be the coordinates of the centre of a large postcode? Again rich attribution is where the real value is added for the community user.

Completeness would be nice but if the information is not there the community user is probably only a Google search away from another source.

As a slight aside Professor Peter Dale's⁸ concept of the local land manager might be thought of as the neogeographer in this type of community.

⁷ See www.sustrans.org for further details

⁸ Prof. Peter Dale, National Mapping Agencies – a new model for the 21st Century, Scotland. FIG XXII International Congress, Washington, D.C. USA, April 19-26 2002

5.4 Professionals

In this context the term professional is used to describe users who are employed by organisations that use geographic data to perform their business activities, whether to analyse, report, navigate or otherwise maintain systems.

Here the user requirements are much more complex and diverse, currency is often paramount. Questions of completeness – my application cannot tolerate finding streets missing in one area but captured in others; positional accuracy – is that mine shaft really under the property I'm planning to buy?; and thematic attribution – is that electricity main really disused?; can also be matters of commercial or actual life and death.

6. Data Quality

Let us now turn to considering what we mean by data quality.

6.1 Quality Characteristics

Quality can be described and measured in a number of ways. The most common definition is "fitness for purpose". In other words, how suitable is this data for my requirement in solving some problem or other, whether it is finding a route from A to B or analysing the distribution of deprivation in England and Wales.

Put another way, what is the ability of this data to satisfy my stated or implied needs. Not all my requirements may be stated but I will have reasonable expectations that certain quality levels will be met. Back to route planning data for example, my expectations will be that the route network is complete and up-to-date even if I do not explicitly state this.

A more formal and limited way of looking at quality is "performance against specification". Assuming my dataset has a specification of some sort, has it actually been captured or maintained to that specified level?

Quality is a relative, there are no absolutes. Saying something is of high or low quality is meaningless unless it is expressed as a measure against a production specification or a user requirement. Data producers and users may have very different views of quality according to the purposes for which the data was captured and then used subsequently. This is illustrated in Figure 2.

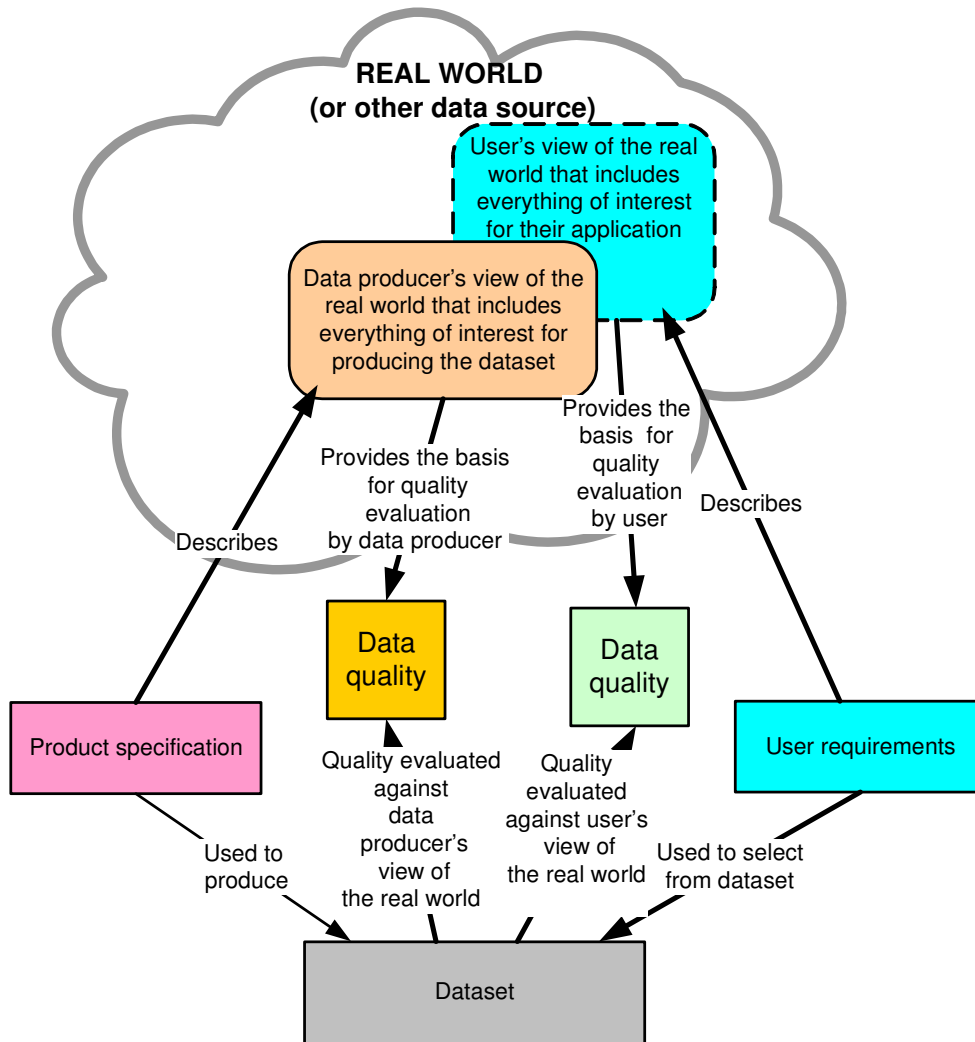


Figure 2: contrasting producer views of quality (derived in part from ISO 19113: 2003 Geographic information - Quality principles)

6.2 Assessing Data Quality⁹

Quality, whether for geographic or neogeographic data, can be assessed and reported against a number of quality elements. These can be either subjective or quantitative. Subjective or non-quantitative quality elements provide an overview of the data and explain why the data was captured, how it was captured, how it is maintained and how it has been used or subsequently modified. Such elements usually go under three headings:

- purpose – the rationale for creating the dataset;
- usage – the application to which the dataset has been put;
- lineage – the history of the dataset.

They provide a very useful initial indication of whether or not a dataset is likely to be useful for a particular purpose.

Other data quality elements can provide quantitative information about aspects of the quality of a dataset and imply a quality evaluation process involving measurement and an objective result. The commonest elements and sub-elements are:

Positional accuracy – the accuracy of the position of features or geographic objects whether in two or three dimensions. This can be expressed, for example as:

- absolute accuracy – closeness of coordinate values to values accepted as true;
- relative accuracy – closeness of relative positions of objects in a dataset to those relative positions accepted as true e.g. distance along a road centre line to a culvert given in a dataset compared to that measured in the real world;
- gridded data position accuracy – closeness of gridded data position values to those accepted as being true.

Temporal accuracy – accuracy of the temporal attributes (dates and times) and temporal relationships (e.g. later or earlier than) of features. This can be expressed for example as:

- accuracy of time measurement – for example if the data states that the object was captured or created on this date was that date correctly recorded;

⁹ The following is based largely upon the International Organization for Standardization standard: ISO 19113: 2003 Geographic information - Quality principles

- temporal consistency – the correctness of ordered events or sequences e.g. is the creation date earlier than the deletion date;
- accuracy of time measurement – for example if the data states that the object was captured or created on this date was that date correctly recorded;
- temporal validity – the validity of data with respect to time e.g. is the data a valid date or time.

Thematic accuracy – accuracy of quantitative attributes and the correctness of non-quantitative attributes and of classifications. For example:

- classification correctness – comparison of the classes assigned to objects or their attributes to ground truth or dataset accepted as true e.g. classes of thoroughfares or structures;
- non-quantitative attribute correctness – correctness of non-quantitative attributes e.g. geographic or other names;
- quantitative attribute accuracy – accuracy of quantitative attributes e.g. population, area.

Completeness – presence and absence of objects in a dataset (by inference at a particular point in time) their attributes and relationships. These can be either errors of:

- omission – data is absent from the dataset which according to the specification should have been included at the time of capture or update e.g. missing streets, street names or classifications;
- commission – data is present in the dataset which should be omitted according to the specification e.g. buildings now demolished (assuming the dataset is being maintained).

Logical consistency – the degree of adherence to logical rules of data structure (whether conceptual, logical or physical), attribution and relationships. For example, these can be characterised as:

- conceptual consistency – degree of conformity to the conceptual schema;
- domain consistency – constraining of values to the value domains;
- format consistency or validity – storage of the data in conformance to the physical structure of the dataset;
- topological consistency – degree of topological fidelity.

7. Analysis: some examples

We can illustrate some of these aspects of data quality in relation to user requirements by reference to the example datasets referred to in section 4.

7.1 Positional Accuracy

In Figure 3, if we assume that the imagery in this Google screenshot represents the “true” position (which it may or may not do) then the road network’s absolute accuracy is not what we might expect – it runs down the sides of roads rather than along the centreline. However, the accuracy of the road centre-line dataset relative to itself appears consistent with that of the imagery - the road segments are of a similar length and the angles comparable.

This “mash up” may be fit for road navigation purposes but not for the professional utilities user locating services relative to the street centreline. This very obvious example does illustrate the difference between specialised professional use and other uses in respect quality. We, as consumers, are tolerant of this kind “inaccuracy”, Google is free for personal use, so can be expected to have limitations and most people's spatial awareness allows them to mentally correct it anyway.



(© 2008 Google –Imagery: ©2008 Infoterra Ltd & Bluesky, Map data; © 2008 Tele-Atlas)

Figure 3: Screenshot of imagery and roads data superimposed

So in comparison, how positionally accurate are neogeographic datasets? Objective and rigorous analysis is still relatively sparse. However, we are indebted to Dr. Muki Haklay, for early access to his, shortly to be published, research results comparing OpenStreetMap and Ordnance Survey (OS) data in both London and the rest of England.¹⁰ He finds OpenStreetMap to be “fairly accurate”. For example, when comparing motorway representation, OpenStreetMap data was, on average, within about 6 metres of the position recorded by OS. Further, there was an 80% intersection, approximately, of “buffered” motorway objects.

7.2 Temporal Accuracy

The first problem with many datasets, whether conventional or neogeographic is to find any information about time of data capture or update let alone have any idea of its accuracy. For example on the Google Maps website¹¹ we are told:

“Google Maps uses the same satellite data as Google Earth. Google Earth acquires the best imagery available, most of which is approximately one to three years old.

“We strive to update our data regularly. Unfortunately, we’re not able to provide detailed information about when a specific area will be updated. Also, we don’t offer high resolution data by order, as this information is added as it becomes available from our data providers.”

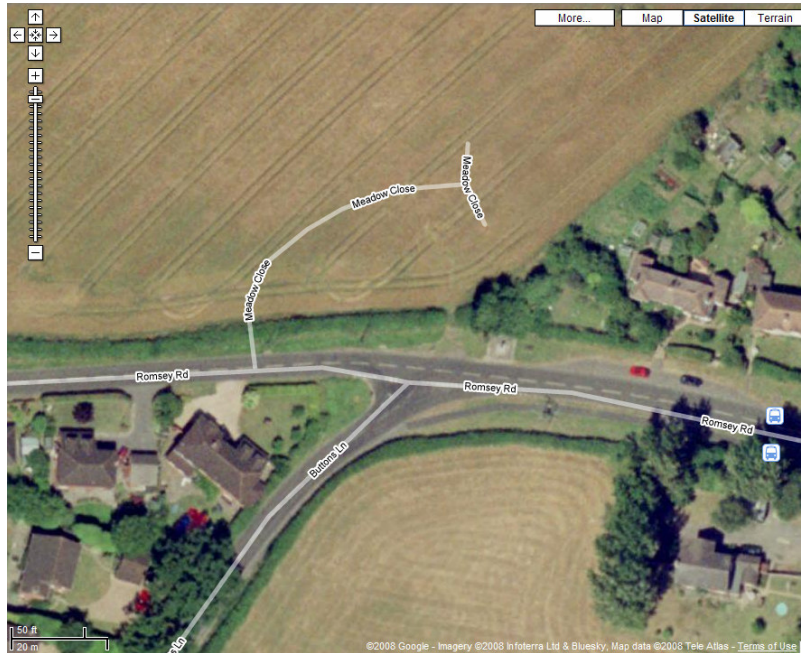
What about the Teleatlas data? Other than a copyright date there is no indication of when it was captured and to what consistency but it is unlikely to have all been captured at the same time as the imagery – this is a problem true of any “mash up”. This is evident from the screenshot at Figure 4.

It is easy to criticise Google for its lack of temporal information and the inability of the user to assess temporal accuracy but in most established areas where there has been little change, it is more than adequate for most consumer purposes.

A neogeographic dataset like UCrime has potentially very good temporal accuracy. The users are also the maintainers - it is in their interest to update the data as soon as an event, in this case a crime, occurs. They are also “on the ground”, so are able to observe change, such as new roads being built, as it occurs. Furthermore, they are more likely to record the date on which events occur (real world time) rather than the time they are updated on the database.

¹⁰ How good is OpenStreetMap Information? Dr. Muki Haklay, under review in Environment and Planning B

¹¹ <http://maps.google.com/support/bin/answer.py?answer=22040&topic=10778>



(© 2008 Google -Imagery © 2008 Infoterra Ltd & Bluesky; Map data © 2008 Tele-Atlas)

Figure 4: Screenshot of imagery and roads data superimposed illustrating the difference in dates of capture of the respective datasets

7.3 Thematic Accuracy

Conventional datasets, by and large, are created to a defined specification in respect to attribution. The attributes themselves have a range of valid values and established methods for indicating where attribution is missing. This enables users of all types to make informed decisions on their usefulness for their purpose. The MasterMap TopoLayer specification¹² for instance is comprehensive in this respect.

In contrast, OpenStreetMap has a very comprehensive list of “map features”¹³ but some freedom in how these are used as indicated by statements such as:

“OpenStreetMap does not have any content restrictions on tags that can be assigned to Nodes, Ways or Areas. You can use any tags you like. However, there is benefit in agreeing a recommended set of features and corresponding tags in order to create, interpret and display a common

¹²

<http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/userguides/docs/OSMMTopoLayerUserGuide.pdf>

¹³ See: http://wiki.openstreetmap.org/index.php/Map_Features

basemap. This page contains a core recommended feature set and corresponding tags."

This would not appear to be approach conducive to achieving a high level of thematic accuracy or very helpful to users.

7.4 Completeness

A common expression of completeness is in terms of "up-to-dateness", the date may be temporally accurate but not be up-to-date or current, at least in the eyes of the user. This is an aspect of quality that matters to all users.

At first glance one might think that conventional products are more "up to date" than neogeographic ones, after all the suppliers employ staff to maintain its currency. However, we are grateful to Steven Feldman¹⁴ for the example of OpenStreetMap in the new town of Cambourne near Cambridge. Since the area is in the process of rapid change, the data captured by the local community, is far more up to date than Teletlas or OS Mastermap® could ever be. The main contributor for the area works in Cambourne and captures change on a weekly basis.

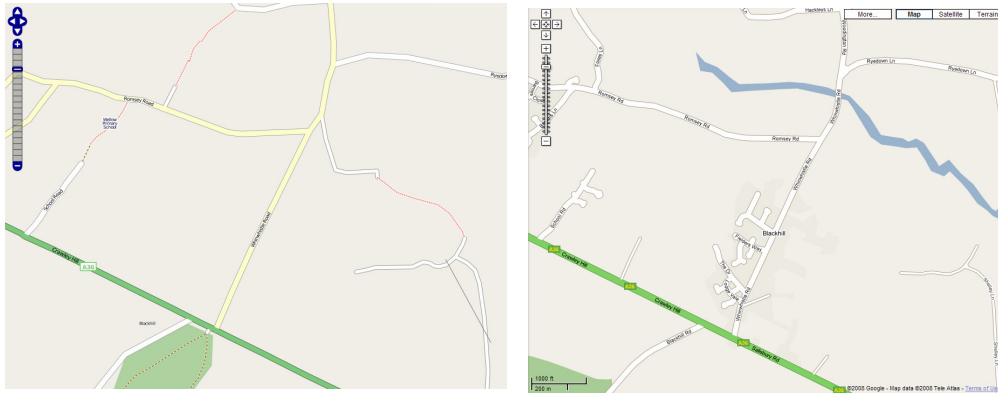
Although, it may not be just up-to-dateness for a particular area that matters to users but also its consistency as Figure 5 illustrates.

Again, Muki Haklay's study is valuable. It comments on OpenStreetMap that "the centres of the major cities of the UK are well mapped and covered. However, many areas that are not covered so well – only 25% of the land area is assessed as having "good" coverage in England, rural areas which require more time to capture, are a particular issue. There are also quality differences related to social deprivation, 8% more of existing roads had been captured in wealthy areas than the poorest areas, based on the Government's index of social deprivation."

OpenStreetMap make no claim that their map is complete as an investigation of their website will show. Indeed there is a table indicating the estimated degree of completeness of various parts of the UK – not always easy to interpret but indicative nevertheless.¹⁵ However, it is the perception of completeness that is perhaps the real problem here. Non-expert users can be easily misled into thinking that the completeness of areas used to advertise the product is achieved universally.

¹⁴ Private communication

¹⁵ See for example: http://wiki.openstreetmap.org/index.php/WikiProject_United_Kingdom#Non-county_areas



(© 2008 OpenStreet Map & © 2008 Google -Imagery © 2008 Infoterra Ltd & Bluesky; Map data ©2008 Tele Atlas)

Figure 5: Screenshots of OpenStreetMap and Google maps of the same area

8. Quality Assurance

There is a very marked difference in conceptual approach between neogeographic and conventionally created datasets in respect to measuring adherence to data specification. Ordnance Survey, for instance, invests much time, money and prestige in its commitment to a fixed data specification, consistent capture methodology and quality assurance applied consistently across its datasets. Jenny Harding 's¹⁶ paper on OS vector data is a very good example of a professional approach to describing data quality and their approach. For instance, OS has a regular programme in which it conducts comprehensive independent sample checks on all its datasets. The quality of its up to dateness is one of the top-level performance monitors for the organisation, expressed as a target percentage of real-world change to be represented in the national database within six months of construction.

In contrast, neogeographic datasets make little mention of quality assurance. Haklay notes that some areas of OpenStreetMap which he studied were obviously covered by a single contributor which implies that no quality assurance was undertaken. Furthermore, there seems to be little control over the ability of one contributor to “undo” the good work of another. One contributor on the OpenStreetMap¹⁷ blog recently commented “I'd like to know when nodes/ways I've created are modified. Since I obviously know these places (I've mapped/contribute to it) I can check that there is no unintentional vandalism. I

¹⁶ Vector Data Quality: A Data Provider's Perspective, Jenny Harding in Fundamentals of Spatial Data Quality edited by Devilliers and Jensoulin, Wiley 2006

¹⁷ <http://wiki.openstreetmap.org/index.php/Amsterdam>

think it would add an important Quality Assurance to OSM if such feeds are possible."

9. Fit for Purpose?

So, to return to the question posed in the title – does data quality matter in the context of neogeographic data? We have seen that this rather depends on:

- the user requirement and their expectations;
- the benefits the user wants to derive;
- what we mean by data quality.

However, the simple answer is that *before* using a dataset *all users* need to decide whether it is "fit for purpose" however simple and undemanding their requirements might be.

For professional users, there is no excuse for avoiding a structured analysis of data quality. This paper sets out the criteria to consider and there are plenty of consultants out there ready to provide external advice, if required. However, this should start with a review of the producers' quality statements and assurance methodology.

There is always going to be a trade off between data quality, user requirements and the specification. Cost will inevitably increase as a higher data quality is required. Similarly, the greater the extent and complex the dataset the greater the elapsed time required to capture and maintain it. You may also be prepared to sacrifice user requirements to reduce cost.

Consumers, special interest groups and community users need a quality statement that needs to be expressed with clarity but brevity. We are not sure that this OpenStreetMap statement really meets these requirements:

"There are no real standards in OSM, the only thing that is defined is how to get data from OSM. That data can be created in many different ways, in a hope that entropy will fix things with time."

However, the Google statement¹⁸ is somewhat better but by no means perfect:

"Information provided through Google is intended for planning purposes only. You may find that weather, construction projects, traffic conditions or other events may cause road conditions to differ from the map results."

¹⁸ http://maps.google.com/help/terms_maps.html

It seems that rather more thought needs to be put into these statements in order to provide adequate guidance to potential users and this applies to both conventional and neogeographic data producers.

10. Conclusions

Neogeographic data is clearly an increasingly significant resource. Its potential to provide more "up to date" data by virtue of being locally sourced is exciting. Positional accuracy is perhaps less of an issue for many purposes than might be imagined.

The main concerns focus around:

- Completeness – the perception that data is universally complete for the area of coverage;
- Consistency – the lack of fixed specification for capture and encoding, whilst being superficially attractive by giving freedom of expression, will constrain use for analysis;
- Quality control - how to assess repeatedly the "wisdom of the crowd" and ensure that "vandalism" unintentional or otherwise does not undermine users' confidence in the product.
- Quality Assurance – the lack of clear and verifiable statements relating to quality, and the methods used to assess it, will be serious impediments to widespread adoption.