Lies, Damned Lies, and (OSM) Statistics

Frederik Ramm <frederik@remote.org>

State of the Map Conference
Milan, 2018-07-28

Slide notes:

This is a commented version of the talk given at the State of the Map conference. Slides are not altered; a recording of the talk is also available.
Lies, Damned Lies, and (OSM) Statistics

(or: misunderstandings)

(or generally quantitative statements about mapping)

Slide notes:

The talk title deserves two remarks:

“Lies” is a harsh word, it suggests having an intention to say the wrong thing. Many wrong things are said by mistake though.

Also, “statistics” is a discipline of mathematics and I’m using it here more in the general sense of quantifying things.
What's in OSM?

Slide notes:

Many people new to OSM want to find out what data they can expect from OSM, and the first thing they turn to is often ...
Slide notes:

... the wiki, which contains detailed descriptions of many things we map. Wiki pages explain the tags to be used for mapping things, what other tags to use together with them, and so on.
Here's an example about the tag “natural=wood”, used mainly to map unmaintained woodland.
Here's another example about “power=transformer”, used to map transformers.
Slide notes:

This page is actually very long, with tons of examples of different transformers and so on.
But what's important?

Slide notes:

Suppose you want to find out which is more important for OSM, woodland or transformers, and you were to base your decision on the wiki alone.
power = transformer has the longer wiki article, it has more documented additional tags, and it even is an “approved” feature, meaning a vote has been held and the feature accepted, whereas natural = wood has fewer of everything, and was never accepted in a vote.
we have PROOF:

OpenStreetMap is a bastion of electricity freaks for whom trees are, at best, raw material for power poles!

Slide notes:

It is easy to be misled by these results into thinking that transformers are more important.
...or not?

Slide notes:

but are they?
Let's look at taginfo (taginfo.openstreetmap.org), a web site that counts how many objects of a given tag are in OSM.
Slide notes:

4.8 million objects with natural=wood versus only 62,000 with power=transformer.
Oops ;)

Slide notes:

It seems our initial guess was wrong.
What did taginfo count?

Slide notes:

Let’s clarify what exactly taginfo counts:
- a count (not total area or length)
- of OSM objects (not real-world objects)
- that have a specific tag
- and are in OSM at present

Unclear: how many mappers?

Slide notes:

It counts how many woodland areas there are, not how big they are. Sometimes the same woodland area may be represented by several different objects in OSM.

It doesn’t count things that were in OSM once and have since been removed, and it also doesn’t tell us how many different people have used these tags; for all we know, all transformers could have been added by one single person!
We need to do some work on the command line to research further.
The “osmium” program can be used to filter out objects with a certain tag from the planet file (the world-wide OSM database), and store it in a text file using the “opl” format.
Slide notes:

The text file has 4.8 million lines, as expected.
Slide notes:

This is how the file is formatted: There are space-separated entries on each line, specifying:

- `n262696` – the object type (node) and ID
- `v4` – version 4
- `dV` – object is visible
- `c343748` – last edited in changeset 343748
- `t2008...` – timestamp of last edit
- `i6809` – edited by user ID 6809
- `uTimSc...` – user name
- `T...` – list of tags the object has, comma separated
- `x, y` – coordinates
A simple Unix command tells us how many different values there are in the 7\textsuperscript{th} field (user name): 35114 different users have between themselves last edited the 4.8 million woodland areas.
Slide notes:

We can also show who the most prolific woodland editors are. Most seem to be import accounts.
Until now we have only looked at the person last editing something. But this does not necessarily tell us who actually **introduced** an object or tag; for all we know, one person could have mapped all the woodlands, and then 35,000 different persons could have edited them afterwards, giving us skewed results.
We can also have osmium convert the “history planet” into an OPL file, which then gives us ALL versions of every object, even those meanwhile superseded.
#!/usr/bin/perl

use strict;
my $last;

while(<>)
{
    my @bits = split(/ /, $_);
    my $obj = shift(@bits);
    my %part = map { substr($_,0,1) => substr($_,1) } @bits;
    my %tag = map {/(.*)=(.*)/; $1=>$2 } split(/,/, $part{'T'});
    if (($tag{'natural'} eq 'wood') && ($obj ne $last)) {
        print $part{'u'}."\n"
        $last = $obj;
    }
}

Slide notes:

Since the opl file is a plain text file, it can easily be processed in a scripting language of your choice.

This example in Perl does the following:

- split each line from the opl file into parts
- take the “T” part (tags) and split it into key/value pairs
- if a “natural=wood” tag is present, and we haven’t already seen “natural=wood” on an earlier version of this object, output the user name corresponding to the edit
This has only slightly changed things; we now have 30.412 different users adding natural=wood tags.
Slide notes:

Assuming that people will sometimes “accidentally” create a new natural=wood object by splitting an existing object in two or other geometry modifications, we can filter away the “long tail” of people having less than 5 natural=wood edits, leaving us with 14.546 people who have introduced natural=wood 5 or more times.
Slide notes:

Doing this in a scripting language can be very slow; processing the whole planet like this takes half a day.
#include <iostream>
#include <osmium/io/any_input.hpp>
#include <osmium/handler.hpp>
#include <osmium/visitor.hpp>

class TagHandler : public osmium::handler::Handler {
    osmium::object_id_type lid = 0;

    public:
        void osm_object(const osmium::OSMObject& object) {
            if (object.tags().has_tag("natural", "wood")) {
                if (lid != object.id()) {
                    lid = object.id();
                    std::cout << object.user() << std::endl;
                }
            }
        }

    int main(int argc, char* argv[]) {
        TagHandler handler;
        osmium::io::Reader reader{argv[1]};
        osmium::apply(reader, handler);
    }

Slide notes:

Luckily, osmium also exists as a C++ library, and the C++ program above does exactly the same as the Perl script shown (and can work on the history file directly, without having to convert to opl format). This will only take half an hour.
Wrapping up the “transformer vs. wood” issue, we see that while the transformer (blue) rules on wiki details, the woodland (green) is clearly more important to OSMers.
The wiki and simple statistics are easy to misread.

Slide notes:

This shows how it is easy to come to wrong conclusions if you do superficial research only.
What brought me to give this talk are gender issues in OSM. As everyone knows, we suffer from a gender imbalance in OSM,
an we have vastly more men than women. This can easily be “proven” by visiting a random OSM event, and all of us would welcome a better balance between genders. It has been shown beyond doubt that diverse teams do better work – and we would like to see people from all walks of life, all genders, nationalities and age groups, in OSM.

However, there have been a few recent scientific and journalistic publications that have painfully misrepresented the gender issue in OSM, and I will go over a few of them.
One study made the (in itself relatively sexist) assumption that women were generally more interested in kindergartens, whereas men were more interested in where to spend their nights. The study pointed out that there exist a multitude of tags aimed at depicting night activities (pub, bar, nightclub, even strip clubs and brothels) but only one tag for kindergartens. The study claimed that this was an obvious sign of OSM being designed and dominated by men’s interests.
However, looking at Germany data only, the country has about 49,000 kindergartens, of which about 33,000 are mapped in OSM. Pubs, bars, and nightclubs together make up 28,000 objects in OSM, and a further 1,500 for brothels, strip clubs etc.

Not only is the assumption that women were less interested in pubs or nightclubs flawed – even if they were, apparently we still manage to have many more kindergartens in OSM than any of the night activities taken together.
Tags are not everything.

Slide notes:

People often think that what they read on the wiki about tags is an indication of the reality in OSM.
One publication highlighted the (true) fact that a tagging proposal for “childcare” has been rejected. However, reading the wiki more closely reveals that only a few dozen people participated in the vote, and the rejection was due to a technical issue with the proposal and not due to people disapproving of child care mapping. And as you have seen, the rejection of the proposal hasn’t kept people from mapping kindergartens.
One researcher recently entered the term “brothel” into taginfo and was surprised to see a large variety of tags describing the various services available at brothels.
Slide notes:

In comparison, there seemed much less tags describing the detailed services available at childcare facilities. The researcher took this as a sign of the OSM community being more interested in brothels than in childcare facilities.
<table>
<thead>
<tr>
<th>Category</th>
<th>Wiki</th>
<th>Data</th>
<th>Mappers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>440</td>
<td>730</td>
<td></td>
</tr>
<tr>
<td>Tags</td>
<td>16</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td></td>
<td>230.828</td>
<td></td>
</tr>
<tr>
<td>Mappers</td>
<td></td>
<td>11.699 (1.719)</td>
<td></td>
</tr>
</tbody>
</table>

(of 1.323 detailed brothel:something tags, 1.182 were added by the same person, and only 15 other people have used these tags more than twice. Numbers in parentheses=mappers with 5+ edits)

Slide notes:

Closer inspection shows that while there are indeed many brothel-specific tags (16), the ration of kindergartens to brothels in OSM is 60:1. Only 400 mappers have ever mapped or modified a brothel, only 41 mappers have added 5 or more brothels, and of 1.323 brothel-specific tags in OSM, 1.182 (almost 90%) have been added by the same individual. This proves that OSM leaves room for niche interests – it does not prove that OSM is full of men only interested in what kind of service is available at a brothel.
Lies, or the creative omission of truths

Slide notes:

Other recent publications have quoted some facts from OSM without putting them into the right context. I will highlight only two of them:
Of 87.175 doctors in OSM, only 958 are gynaecologists!

<table>
<thead>
<tr>
<th>Number of doctors in OSM: 87.175</th>
</tr>
</thead>
<tbody>
<tr>
<td>of these, gynaecologists: 958</td>
</tr>
<tr>
<td>general: 3940</td>
</tr>
<tr>
<td>ophthalmology: 896</td>
</tr>
<tr>
<td>internal: 765</td>
</tr>
<tr>
<td>paediatrics: 649</td>
</tr>
<tr>
<td>various others of lesser frequency: 6342</td>
</tr>
<tr>
<td>73.625 without any detail</td>
</tr>
</tbody>
</table>

Slide notes:

It is true that (at the time of giving the presentation), only 958 doctors out of 87.175 in OSM were marked as gynaecologists.

It is equally true that “gynaecologist” is the second most frequently used doctor’s specialisation in OSM, after “general”. The overwhelming portion of doctors does not have any specialisation listed.

This is a matter of general lack of detail, not of anti-women bias.
Of 221.923 toilets in OSM, only 9.389 are for women!

Slide notes:

It is true that (at the time of giving the presentation), only 9.389 toilets out of 221.923 in OSM were marked as being for women.

However, only 9.550 toilets are marked as being for men; 190.861 toilets are not marked with any gender. (Apologies to the audience for the gross simplification of only discussing binary gender here.)

This, too, is a matter of general lack of detail, not of anti-women bias.
Bad science: don't do it!

If you write about OSM, and its undeniable gender imbalance, please try not to misrepresent the efforts of the OSM community. Don’t present the numbers that suit you and ignore the rest.

It can sometimes be difficult to interpret the wealth of information in OSM, and easy to draw the wrong conclusions from a wiki article. If you are unsure, try to talk to the community about it and people will help you.
Thank you

Slide notes: