



# Lies, Damned Lies, and (OSM) Statistics

Frederik Ramm <[frederik@remote.org](mailto: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, *(or: misunderstandings)* Damned Lies, and (OSM) Statistics *(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 ...

English Create account Log in

Main page Discussion Read View source View history Search OpenStreetMap Wiki

**Available languages — Main Page** Help


• Afrikaans • asturianu • azərbaycanca • Bahasa Indonesia • Bahasa Melayu • bosanski • brezhoneg • català • čeština • dansk • Deutsch • eesti • **English** • español • Esperanto • euskara • français • Frysk • galego • hrvatski • interlingua • isienska • italiano • kréyòl gwadloupeyen • kurdî • latviešu • Lëtzebuergesch • lietuvių • magyar • Nederlands • norsk • occitan • polski • portugues • română • shqip • slovenčina • slovenscina • suomi • svenska • Tiếng Việt • Türkçe • Zazaki • срpski / srpski • Български • македонски • русский • українська • Ελληνικά • Հայերեն • नेपाली • ལྷོ་ཡུལ་སྐད་ • ལྷོ་ཡུལ་སྐད་ • ភាសាខ្មែរ • 한국어 • 日本語 • 中文 (简体) • 中文 (繁體) • עברית • العربية • بھٹیو • فارسی

**Other languages — Help us translate this wiki**

**Welcome to OpenStreetMap**, the project that creates and distributes *free* geographic data for the world. We started it because most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways.

[More about OpenStreetMap](#) | [How to contribute](#) | [Where to get help](#)


**Use OpenStreetMap**



Using OpenStreetMap

- Browse our world map
- Check the ready-to-use products for your mobile device, your desktop computer or the web services
- [...more on using OpenStreetMap](#)


**Contribute free map data**



Beginners' Guide

- Browse the map feature documentation
- Browse the Mapping projects
- [...more on contributing map data](#)

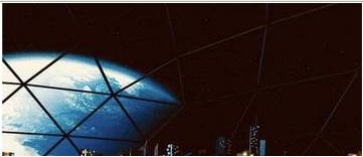
**Software Development**



Develop and use the Platform

- Use OpenStreetMap for your software
- Contribute to the OpenStreetMap software

**Image of the week**




**Event Calendar** — See also Past Events

21 Jul.	Mappertreffen, Essen, Germany	
21 Jul.	東京1街歩き!マッピングパーティー:第21回 増上寺, Tokyo, Japan	
21 Jul.	Je cartographie Blois - Cartopartie sur le thème des espaces verts, Blois, France	
21 Jul.	More Joy Diversion, Manchester, United Kingdom	
24 Jul.	Vélocité, Bordeaux, France	
24 Jul.	Pub Meetup, Nottingham, United Kingdom	

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.



[English](#) [Create account](#) [Log in](#)

---

[Page](#) [Discussion](#)

[Read](#) [View source](#) [View history](#)

---

## Tag:natural=wood

**Available languages — Tag:natural=wood**

• čeština • Deutsch • **English** • español • polski • português • русский • українська • 日本語

**Other languages — Help us translate this wiki**

[purge](#) • [Help](#)

---

**Forest**, by some used to tag woodland with no forestry.

There are major differences in the way this tag and [Landuse=forest](#) are used by some Openstreetmap users. Some use this tag to show an area is covered in trees, others use it for woodland not impacted by human maintenance. This problem is explained in the page **Forest**.

See the page **Forest** to understand the usage of this tag and [Landuse=forest](#).

**Contents** [\[hide\]](#)

- 1 How to map
  - 1.1 Additional tags
- 2 Rendering
- 3 Tagging mistakes
- 4 See also

### How to map

Create an area and tag it [natural=wood](#).


If you are not sure of its border you can place a single [node](#) in the middle and tag it [natural=wood](#) but the area is preferable.

### Additional tags

- [name=\\*](#) - name of woodland
- [leaf\\_type=broadleaved/needleleaved/mixed](#) - describes the type of leaves / needles.
- [leaf\\_cycle=deciduous/evergreen/mixed](#) - describes the phenology of leaves / needles.

### Rendering


**natural = wood** v · d · e



**Description**





Forest. Sometimes considered to have restricted meaning "Woodland with no forestry".

**Rendering in [openstreetmap-carto](#)**



**Group:** Forest

**Used on these elements**







**Useful combination**

- [name=\\*](#)
- [leaf tvne=\\*](#)

Slide notes:

Here's an example about the tag "natural=wood", used mainly to map unmaintained woodland.



[English](#) [Create account](#) [Log in](#)

---

Page [Discussion](#)
Read [View source](#) [View history](#)

---

## Tag:power=transformer

**Available languages — Tag:power=transformer**  
 · Deutsch · **English** · français · italiano · polski · русский · 日本語  
**Other languages — Help us translate this wiki**

[purge](#) · [Help](#)

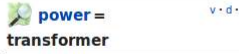
---


A power **Power Transformer** (421-01-01<sup>ⓘ</sup>) static device which converts a given power voltage to another power voltage. A transformer is usually located within a **power=substation**.

In more technical terms, a power transformer is composed of two or more windings which, by electromagnetic induction, transforms a system of alternating voltage and current into another system of voltage and current for the purpose of transmitting electrical power. The delivery is done at the same frequency than the input.

**Contents** [hide]

- 1 How to map
- 2 Advanced mapping
  - 2.1 Where do I find such data ?
  - 2.2 Tagging
  - 2.3 Transformer values
  - 2.4 Location values
  - 2.5 Transformers interfaces
    - 2.5.1 Voltage tagging
    - 2.5.2 Windings configuration
  - 2.6 Transformer sets
- 3 Examples
  - 3.1 Transmission transformers
  - 3.2 Distribution transformers
  - 3.3 Traction transformers
  - 3.4 Auxiliary transformers









**Description**

A static device for stepping up or down electric voltage by inductive coupling between its windings. Large power transformers are typically located inside substations

**Used on these elements**

**Useful combination**

- [transformer=\\*](#)
- [operator=\\*](#)
- [frequency=\\*](#)

Slide notes:

Here's another example about "power=transformer", used to map transformers.

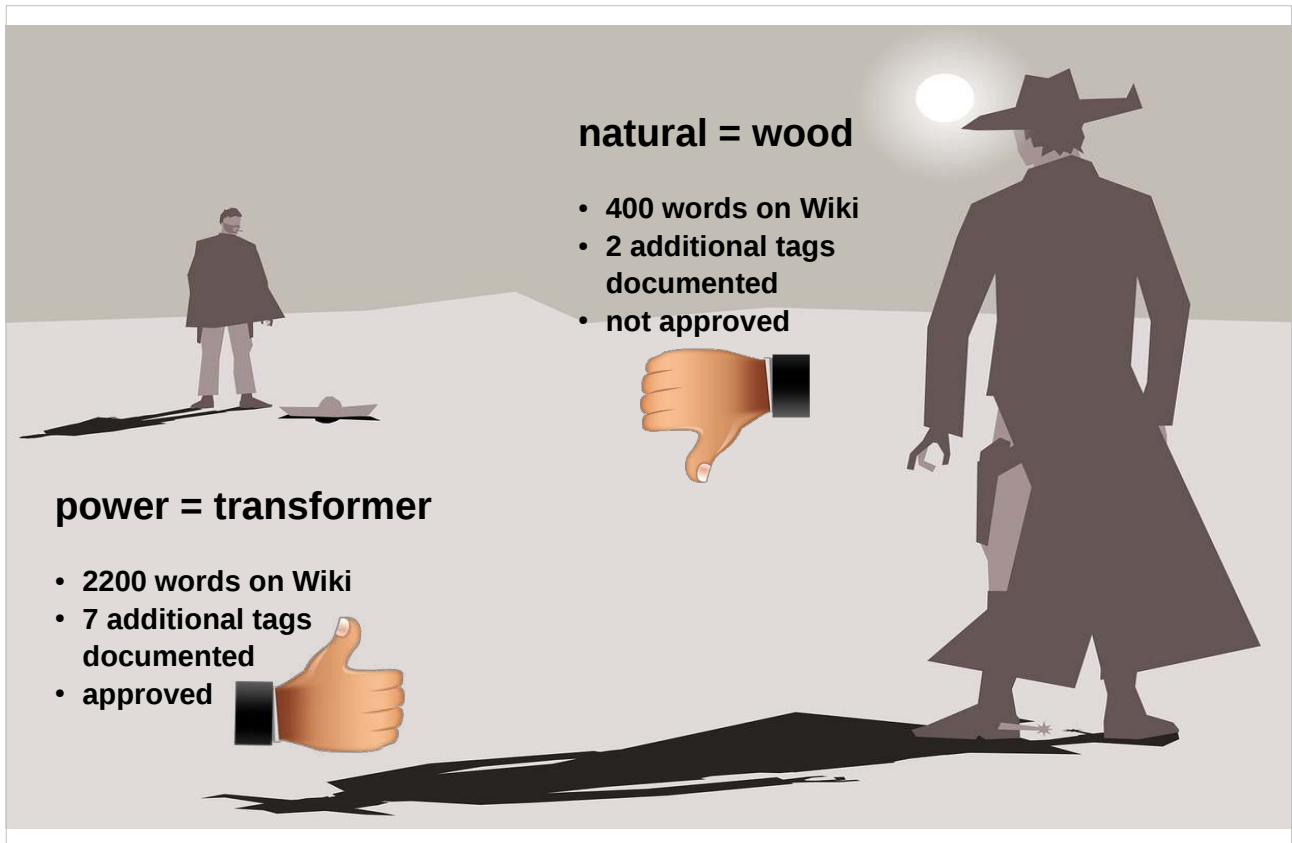


# **But what's important?**

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.





Slide notes:

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?



## KEY/TAG COMPARISON

power=transformer ✕

A static device for stepping up or down electric voltage by inductive coupling between its windings. Large power transformers are typically located inside substations

<input checked="" type="checkbox"/> All	62 816
<input type="checkbox"/> Nodes	51 440
<input type="checkbox"/> Ways	11 364
<input type="checkbox"/> Relations	12

Wiki pages about this tag:

[de](#) [en](#) [fr](#) [it](#) [ja](#) [pl](#) [ru](#)natural=wood ✕

Forest. Sometimes considered to have restricted meaning "Woodland with no forestry".

<input checked="" type="checkbox"/> All	4 874 615
<input type="checkbox"/> Nodes	6 970
<input type="checkbox"/> Ways	4 495 062
<input type="checkbox"/> Relations	372 583

Wiki pages about this tag:

[cs](#) [de](#) [en](#) [es](#) [ja](#) [pl](#) [pt](#) [pt-br](#) [ru](#) [uk](#)

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!





\$

Slide notes:

We need to do some work on the command line to research further.

```
$ osmium tags-filter -R planet.osm.pbf -o wood.opl natural=wood  
[=====] 100%
```

Slide notes:

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.

```
$ osmium tags-filter -R planet.osm.pbf -o wood.opl natural=wood  
[=====] 100%  
  
$ wc -l wood.opl  
4874615
```

Slide notes:

The text file has 4.8 million lines, as expected.

```
$ osmium tags-filter -R planet.osm.pbf -o wood.opl natural=wood
[=====] 100%

$ wc -l wood.opl
4874615

$ head -1 wood.opl
n262696 v4 dV c343748 t2008-06-30T12:00:55Z i6809 uTimSC_Data_
CC0_To_Andy_Allan Tname=Craigs%20%Wood,natural=wood,created_by
=Potlatch%20%0.5d x-0.7375861 y51.1050004
```

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

```
$ osmium tags-filter -R planet.osm.pbf -o wood.opl natural=wood
[=====] 100%

$ wc -l wood.opl
4874615

$ head -1 wood.opl
n262696 v4 dV c343748 t2008-06-30T12:00:55Z i6809 uTimSC_Data_
CC0_To_Andy_Allan Tname=Craigs%20%Wood,natural=wood,created_by
=Potlatch%20%0.5d x-0.7375861 y51.1050004

$ cut -d\ -f7 wood.opl | sort -u | wc -l
35114
```

Slide notes:

A simple Unix command tells us how many different values there are in the 7<sup>th</sup> field (user name): 35114 different users have between themselves last edited the 4.8 million woodland areas.

```
$ head -1 wood.opl
n262696 v4 dV c343748 t2008-06-30T12:00:55Z i6809 uTimSC_Data_
CC0_To_Andy_Allan Tname=Craigs%20%Wood,natural=wood,created_by
=Potlatch%20%0.5d x-0.7375861 y51.1050004

$ cut -d\ -f7 wood.opl | sort -u | wc -l
35114

$ cut -d\ -f7 wood.opl | sort | uniq -c | sort -rn | head -5
70058 uCanvecImports
67422 uGISHulyak
56915 uAmateurCartographer_import
52904 uMilos%20%Cekovic
50887 umrsid_linz
```

Slide notes:

We can also show who the most prolific woodland editors are. Most seem to be import accounts.

**last editor**  
**!=**  
**first mapper**

Slide notes:

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.

```
$ osmium cat history-latest.osh.pbf -o history.opl
[=====] 100%

$ head -5 history.opl
n1 v1 dD c9257 t2006-05-10T18:27:47Z i1298 ut12 T x y
n1 v3 dV c524633 t2009-04-14T15:42:57Z i5164 uwoodpeck T x2 y2
...
n262696 v4 dV c343748 t2008-06-30T12:00:55Z i6809 uTimSC_Data_
CC0_To_Andy_Allan Tname=Craigs%20%Wood,natural=wood,created_by
=Potlatch%20%0.5d x-0.7375861 y51.1050004

$
```

Slide notes:

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

```
$ perl filter.pl < history.opl | sort -u | wc -l
30412 (before: 35114)

$ perl filter.pl < history.opl | sort | uniq -c | sort -rn |
head -5
 74181 GISHulyak
 73377 CanvecImports
 63290 mrsid_linz
 58918 AmateurCartographer_import
 55137 Milos%20%Cekovic
```

Slide notes:

This has only slightly changed things; we now have 30.412 different users adding natural=wood tags.

```
$ perl filter.pl < history.opl | sort -u | wc -l
30412 (before: 35114)

$ perl filter.pl < history.opl | sort | uniq -c | sort -rn |
head -5
 74181 GISHulyak
 73377 CanvecImports
 63290 mrsid_linz
 58918 AmateurCartographer_import
 55137 Milos%20%Cekovic

$ perl filter.pl < history.opl | sort -u |
grep -v "^ [1-4]" | wc -l
14546
```

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;
            }
        }
    }
};

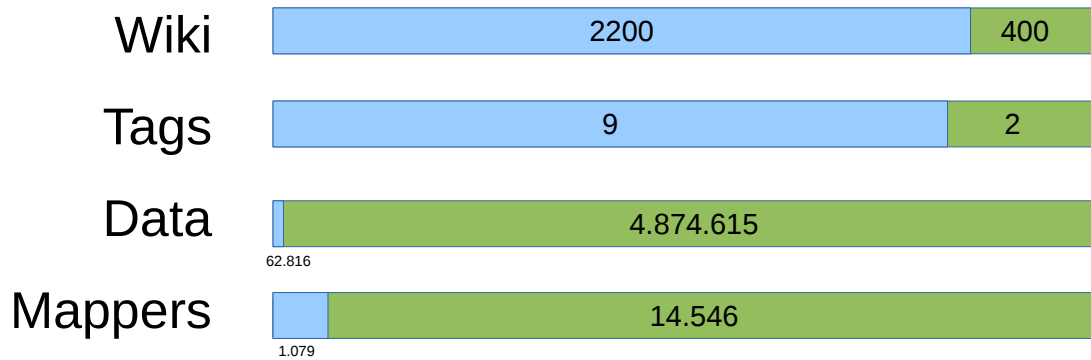
int main(int argc, char* argv[]) {
    TagHandler handler;
    osmium::io::Reader reader{argv[1]};
    osmium::apply(reader, handler);
}

```

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.

# Transformer vs. Wood



Slide notes:

Wrapping up the “transforme 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:

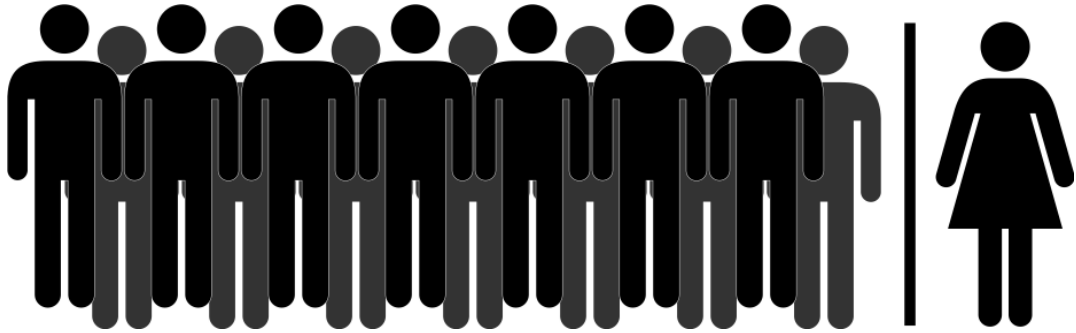
This shows how it is easy to come to wrong conclusions if you do superficial research only.



Slide notes:

What brought me to give this talk are gender issues in OSM. As everyone knows, we suffer from a gender imbalance in OSM,





Slide notes:

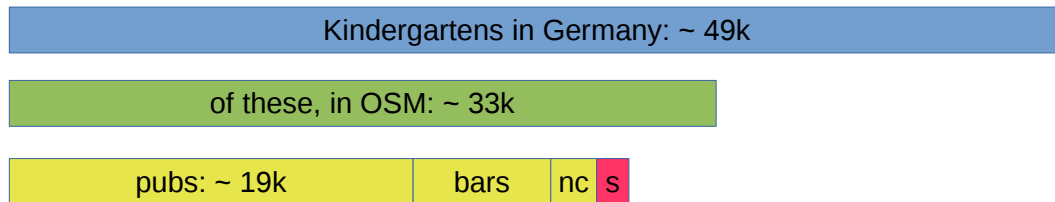
an we have vastly more men than women. This can easily be “proven” by visiting a random OSM event, and all of us would welcome a better balance between genders. It has been shown beyond doubt that diverse teams do better work – and we would like to see people from all walks of life, all genders, nationalities 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.



Slide notes:

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.



(bars: 6306, nightclub: 1605, brothel/stripclub etc:1488)

### Slide notes:

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 night clubs 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.

Page [Discussion](#) [Read](#) [View source](#) [View history](#)

## Proposed features/childcare

[< Proposed features](#)

**childcare**

**Status:** Rejected (inactive)

**Proposed by:** flaimo

**Tagging:** `amenity=childcare`

**Applies to:**

**Definition:** A place for children to do homework, play and spend time outdoors after school or kindergarten.

**Rendered as:** Like amenity=school

**Drafted on:** 2011-04-21

**RFC start:** 2011-04-21

**Vote start:** 2011-04-21

**Vote end:** 2011-04-25

**REJECTED**

- 1 Key
- 2 Description
- 3 Additional tags
- 4 Rendering
- 5 amenity=childcare vs amenity=social\_facility
- 6 Changes based on comments from the RFC phase

Slide notes:

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.

## SEARCH RESULTS

You were searching for: brothel

Keys Values Relation types Roles Full text

### Keys

Page 1 of 2 JSON Displaying 1 to 12 of 21 items

Count	Key
181	<a href="#">brothel:saunaclub</a>
174	<a href="#">brothel:club</a>
141	<a href="#">brothel:apartment</a>
133	<a href="#">brothel:eros_center</a>
59	<a href="#">brothel:flat_rate</a>
54	<a href="#">brothel:escort_services</a>
35	<a href="#">brothel:contact_bar</a>
30	<a href="#">brothel:gangbang</a>
16	<a href="#">brothel:street_prostitution</a>

Slide notes:

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.

## SEARCH RESULTS

You were searching for: childcare

Keys Values Relation types Roles Full text

### Keys

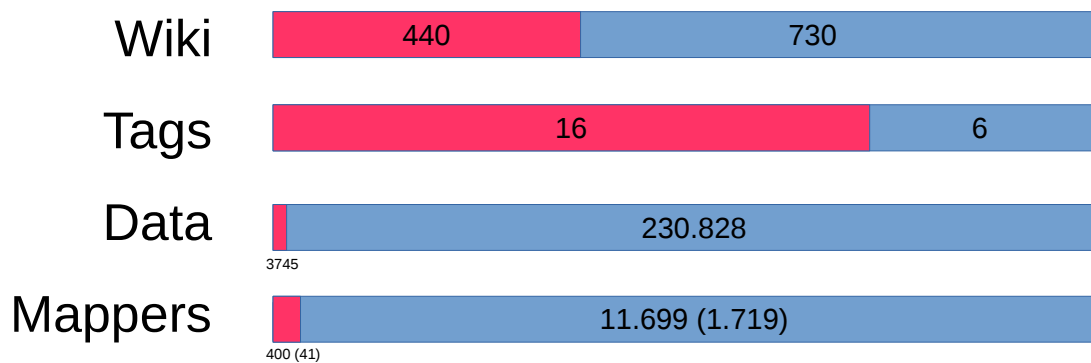
Page 1 of 1 JSON Displaying 1 to 10 of 10 items

Count	Key
46	<a href="#">childcare</a>
15	<a href="#">childcare:capacity</a>
2	<a href="#">childcare:creche</a>
1	<a href="#">childcare:afterschool</a>
1	<a href="#">service_times:childcare</a>
1	<a href="#">childcare:montessori</a>
1	<a href="#">amenity:childcare</a>
1	<a href="#">childcare:type</a>
1	<a href="#">childcare 1</a>

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.

# Brothels vs. Kindergartens



(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.

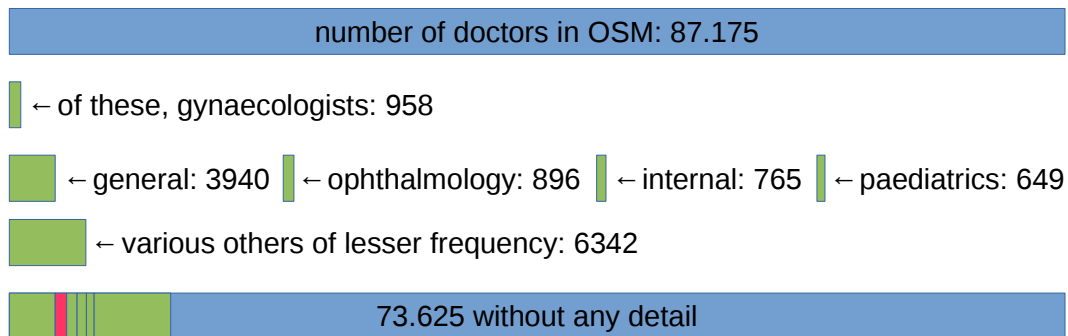


# Lies, or the creative omission of truths

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!



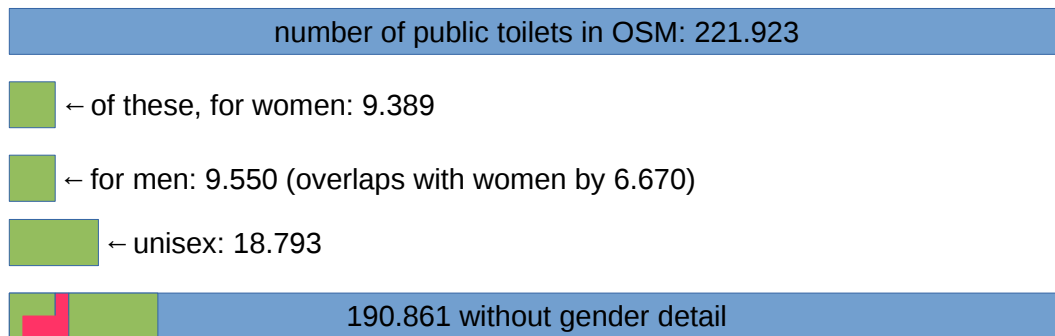
## Slide notes:

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!



Slide notes:

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: