

Немного о названиях улиц

Дмитрий Маракасов

OpenStreetMap

amdmi3@amdmi3.ru

Веб и ГИС 2011

<http://amdmi3.ru/files/webgis2011/>

Названия улиц...

Улица Ленина

Ленина, улица

ул. Ленина

Ленина ул.

ул Ленина

Ленина ул

ул.Ленина

ЛЕНИНА УЛИЦА

Ленина,ул.

Ленина

Ленина улица

...

Чем это плохо

- Отсутствие единого стиля
- Потеря информации
 - Алая улица != улица Алая
<http://ru.wikipedia.org/wiki/Алай>
 - Бобров переулок != переулок Бобров
- **Несоответствия названий улиц адресам (проблемы с адресным поиском)**
- Неоднозначности
 - «Майский пр.» — проезд/проспект/просек?
 - «Ленина» — ...что? (особенно когда есть и улица и проезд)
- Война правок (у каждого свой привычный стиль)
- Вторичные ошибки (новички видят «ул.», пишут по аналогии «пр.»)

«Соглашение о именовании дорог»

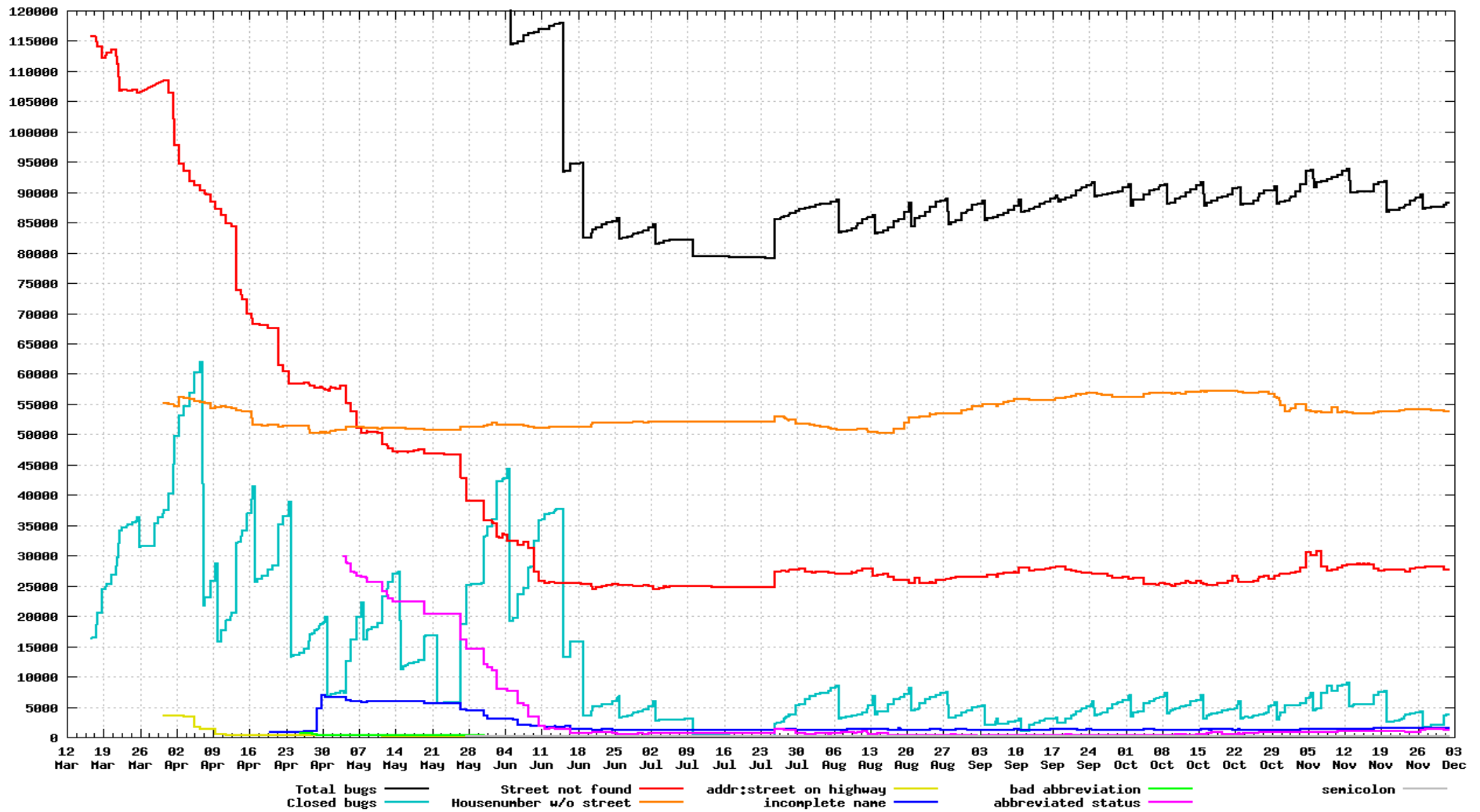
- Естественный порядок слов
 - улица Ленина (существительное, статусная часть в начале)

НО
 - Ленинская улица (прилагательное, статусная часть в конце)
- Статусная часть пишется без сокращений, со строчной буквы

Что решает соглашение

- ✓ Отсутствие единого стиля
- ✓ Потеря информации
 - Алая улица != улица Алая
<http://ru.wikipedia.org/wiki/Алай>
 - Бобров переулок != переулок Бобров
- ✓ **Несоответствия названий улиц адресам (проблемы с адресным поиском)**
- ✓ Неоднозначности
 - «Майский пр.» — проезд/проспект/просек?
 - «Ленина» — ...что? (особенно когда есть и улица и проезд)
- ✓ Война правок (у каждого свой привычный стиль)
- ✓ Вторичные ошибки (новички видят «ул.», пишут по аналогии «пр.»)

Исправление адресных ошибок по ходу нормализации названий



Алгоритм vs. словарь

- Первоначальная алгоритмическая реализация упёрлась в необходимость добавления большого набора костылей и исключений, поскольку без словаря нельзя надёжно исправить значительную часть случаев, например:
 - «улица 3 Текстильщиков» vs. «улица 26 Бакинских Комиссаров» - где номерная часть?
 - «улица Светлая» vs. «улица Тукая» - где прилагательное?
 - «Заозерная улица» vs. «Бульдозерная улица» - где нужна ё?
- Из-за этого — неоправданное усложнение кода и необходимость ручной проверки всех исправлений после работы нормализатора
- В новой реализации во главу угла поставлен словарь

Новая реализация (streetmangler)

- **База-словарь правильных написаний названий *всех* улиц (и не только улиц)**
- Библиотека для обработки названий (как с, так и без использования базы)
- Утилита для извлечения и проверки названий из дампов OSM
- Утилита для исправления названий в OSM (в комплект пока не входит)
- <https://github.com/AMDmi3/streetmangler>

Возможности (1/2)

- Классификация/проверка названий. При наличии в базе «улица Ленина», умеем:
 - Найти точное совпадение с базой (т. е. подтвердить правильность написания):
 - «улица Ленина» = «улица Ленина»
 - Надёжно исправить неканоническую форму (совпадение с точностью до написания, положения статусной части и регистра):
 - «ленина ул.» → «улица Ленина»
 - Найти предположительно правильный вариант с учётом наличия орфографических ошибок:
 - «улица Лениина» → «улица Ленина»
 - «улица Ленеа» → «улица Ленина»
 - Определить отсутствие статусной части
 - «Ленина» — улица с отсутствующей статусной частью, а «грунтовка» — нет
- Постобработка в виде форматирования статусной части
 - например, приведение к «Ленина ул.» для конвертации в форматы навигаторов

Возможности (2/2)

- Быстрый код на C++ (10G XML дамп России из OSM обрабатывается 4 минуты)
- Поддержка других ЯП через swig (есть биндинги к Python, спасибо [PShA](#))
- Поддержка различных (человеческих) языков и разных наборов правил для каждого языка ([Котяра](#) использует для белорусского)
- Библиотеку можно использовать отдельно от базы, например в конверторах
- Базу можно использовать отдельно от библиотеки, например в других картопроектах
- Единственная сложность — нужно сначала заполнить базу

Наполнение базы

- В процессе. На 1 декабря 2011:

| | Всего | Точное | Неканон. | Орфогр. | Без ст.ч. | Ненайд. | Мусор |
|-----------|---------|---------------|----------|---------|-----------|---------------|-------|
| Объектов: | 1012754 | 88.77% | 0.57% | 1.58% | 0.78% | 6.55% | 1.74% |
| Названий: | 44270 | 52.89% | 3.52% | 5.39% | 2.44% | 26.72% | 9.04% |

Из исправленного

- Окрябрьская улица
 - Октбрьская улица
 - Октяборьская улица
 - Октябрская улица
 - **Октябырьская улица**
 - Октябьская улица

 - Комомольская улица
 - КомсКомольская улица
 - **Комсолольская улица**
 - Комсомолькая улица
 - Комсомольскя улица
 - Комсомоьская улица

 - улица Двержинского
 - улица Джержинского
 - улица Дзерджинского
 - улица Держизнского
 - улица Держиского
- Магистрольная улица
 - Лениградский проспект
 - Комунальная улица
 - ул. Буденного
 - ул. Рокосовского
 - улица Зиюля
 - прспект Жукова
 - Фесивальная улица
 - Мухинлй улица
 - улица Алуксандра Невского
 - улица Березовая роща
 - улица Богдана Шмельницкого
 - улица Карлалибкнехта
 - улица Красного Восстания
 - улица Лыткиа
 - улица Энергентиков
 - улица Пискунова (латинская «с»)
 - улица Тимерязева
 - ...

Спасибо.