1. Перед началом работы необходимо переформатировать в OpenRefine формат JSON в CSV.
2. Далее необходимо проверить насколько удобно проименованы колонки и если нет – исправить это. Удобнее работать в Excel.
3. Работу делим на этапы, в которых занимаемся только одной задачей (например только чисткой файла, или только перенесением атрибутов в онтологию). Делать это нужно для последующей организации времени. Если какой-то из этапов отнимает слишком много времени, возможно от этого этапа стоит отказаться.
4. Затем осматриваем файл в OSM, объединяем multi-valued cells если есть необходимость, просматриваем файлы в формате строк, а не записей. Чистим файл от пустых строк, если есть необходимость.
5. В первую очередь проверяем ключевые атрибуты (имя, адрес, координаты, некоторые спецэфические атрибуты для разных классов), если на этом этапе присутствуют какие-то проблемы (например отсутствует в значительном числе полей) необходимо об этом сразу сообщать. Возможно не имеет смысла вообще тратить время на этот датасет.
6. Если есть хоть какие-то сомнения по поводу того, стоит ли включать какие-то атрибуты в онтологию обязательно следует спрашивать об этом ведущего проектом (с приложением описания этого атрибута). Вообще стоит обсуждать все атрибуты, которые не являются явно бесполезными или пустыми.
7. Необходимо соблюдать правило: если что-то проще сделать по-другому – делать это по-другому (например удалять колонки намного проще в экселе, следовательно, если ненужных колонок много, то лучше сделать это в экселе, однако если удалить нужно одну-две колонки, то открывать для этого эксель не имеет смысла).
8. В документацию необходимо вносить только те атрибуты, которые либо а) Относятся к этому классу, либо б) Пока присутствуют только у индивидов новодобавленных классов и их судьба в будущем неизвестна.
9. Описание атрибутов должно быть грамотным, полным, но при этом не излишним.