Отзыв на датасет "Переписка русских литераторов"

База данных "Переписка русских литераторов" представляет собой интересный материал систематизированных данных как об объемах известной исследователям переписки литературных деятелей, так и о хронологии и статистике ее публикации в академических изданиях. Тем не менее, для успешного введения базы в научный оборот, как мне кажется, необходимо сделать масштабные преобразования исходной таблицы (20221001_muratova ... .csv), необходимые для ее успешной (в первую очередь) машинной обработки.
Суть моего предложения заключается в том, чтобы разделить имеющуюся таблицу на несколько таблиц реляционнной БД, связанных id-ключами. Минимально стоило бы сделать 3 таблицы:

- Таблица источников (акад. издания)
- Таблица персоналий (акторов)
- Таблица писем

Sources (метаданные = акад. изд.)

Разделение на несколько таблиц позволит уйти от смешения, с одной стороны, метаданных об источнике публикации писем (описание выходных данных акад. изданий), а с другой -- данных о самих письмах и их датировках и прочей информации.
Совершенно точно не стоит давать в Readme (неполный) список академических изданий -- лучше сделать отдельную таблицу, где можно будет описывать каждое издание (если угодно -- отдельные тома).
Это нужно, как мне кажется, не только для систематизации, но и для потенциального решения вопроса о том, кто, как и где публиковал письма -- поэтому здесь можно подумать о каких-то дополнительных колонках для описания изданий (например, иметь категорию издания и отмечать, где это ПСС, а где -- исторический журнал типа "Русского архива"); можно, в целом ограничиться и унификацией уже существующей колонки *publication_type*.
За основу здесь должна быть взята колонка "publication" (сейчас это 3799 уникальных значений, а должно быть гораздо меньше), примерные колонки в этой таблице:

- **source_id** -- краткое название источника, например: LN
- **title** -- название: Литературное наследство
- **n_vol** -- количество томов (1, если однотомное)
- **place** -- место публикации
- **year** -- год(ы) публикации и т.д.

Individuals

Отдельная таблица, где стандартизировано описан каждый узел сети (1 актор = 1 строка). В сегодняшнем варианте таблицы все получатели писем даны в формате "А. С. Пушкину. (1). Май 1821 — июнь 1823 / ~ 10 сентября 1826—1827." -- я понимаю, что это понятная с филологической стороны нотация, но она совершенно бесполезна для машинного чтения, а также для соотнесения с отправителем, записанным как "Пушкин, Александр Сергеевич". Мне не совсем ясно, как в таком случае соотнести в одном узле Пушкина и отправляющего, и получающего письма, а сделать это -- вероятно, первостепенная задача для пользователя сети. Смежная Дательному падежу проблема -- если есть разногласия в том, как дается имя адресата ("Редактору [А. Сомову]" не равно для машины просто "А. Сомову"), такие случаи должны быть унифицированы. Для этого и необходима отдельная таблица для авторов/акторов/ узлов, которая будет суммировать такого рода информацию, при этом с помощью ID закрепляя ее за одним человеком.
Как наверное и так известно, для построения любых сетей обычно есть две таблицы: одна описывает направления и содержит две колонки (From --> To: Pushkin A.S. --> Slepushkin F.N.), вторая описывает каждый из узлов (допустим, три колонки: имя, дата рождения и дата смерти: 1 Pushkin A.S.; 1799; 1837; 2 Slepushkin F.N.; 1787; 1848). Предлагаю сделать что-то подобное, таблица авторов в таком случае может быть построена на основании таких колонок:

- **person_id** -- любой уникальный ID для каждого отправителя/получателя
- **person_name** -- Фамилия И.О.
- **person_full_name** -- Фамилия, Имя Отчество
- **person_other_names** -- все значения из скобок с псевдонимами и т.п.
- **sex** -- пол
- **year_birth** -- год рождения
- **year_death** -- год смерти
- **occupation** -- профессия
- **other** -- все остальные сведения
- ! NB сейчас в таблице неверно названа колонка "autho**n**_occupation"

Letters (таблица писем from --> to)

В этой таблице можно использовать *person_id* для решения вопроса об направлении писем. Помимо направления и акторов, сюда же можно добавить все ценные сведения о самих письмах, которые уже есть в таблице. Исходными здесь будут колонки author и personalities (последнюю нужно автоматически разделить по точкам на несколько колонок). Примерный состав этой таблицы:
(данные о письме)

- **from** -- адресант (person_id)
- **to** -- адресат (person_id)
- **n_letters** -- количество писем (автоматически вычленить из круглых скобок)
- датировки самих писем -- поделить на несколько колонок
  - **letter_year_start** -- год начала переписки
  - **letter_year_end** -- год окончания
  - letter*month*start; letter*month*end -- опционально по той же логике

(данные об источнике)

- **source_id** -- ссылка на id издания в таблице *sources*
- **source_pages** -- страницы в издании-источнике (м.б. что-то еще)
- датировка источника публикации писем:
  - за основу -- колонка publication_year : сейчас в ней находятся также страницы ("с. 118-119", обозначения "Б. д." и "Б. г.", а также пустые ячейки -- это нужно унифицировать)
  - **publication_year_start** -- начальный год публикации
  - **publication_year_end** -- конечный год публикации (даже если это период вроде "1941; 1951" -- начальный будет 1941, конечный -- 1951)
- **publication_type** (в этой колонке сейчас 37 уникальных значений, нужно исправить все опечатки и разночтения)

Дополнительно

- **notes** -- из этой колонки таблицы было бы ОЧЕНЬ полезно вытащить в отдельную колонку все упоминания о языках ("на франц. яз.") -- опять же, это можно сделать автоматически.
- **Readme** -- текст описывает скорее не данные, а исследовательские вопросы, которые ставила перед собой автор при работе над базой. Текст должен быть скорректирован с учетом того, как выглядят остальные Readme в репозитории;
- Название файла с базой должно содержать название датасета, ссылка на Муратову вводит в заблуждение.

Подчеркну, что хотя это существенные изменения, большую часть этой работы **можно** сделать автоматически при помощи регулярных выражений и т.п.
Мои предложения, разумеется, опциональны и могут быть скорректированы, однако мне кажется, что разделение на более ясные категории и унификация позволят, во-первых, быстро применить эти данные для построения сетей, во-вторых, расширят потенциальные возможности использования базы для изучения разных (не только ориентированных на сети) вопросов: этим можно перенаправить критику о неполноте и нерепрезентативности базы на решение других потенциально

небезынтересных научных задач.

Антонина Мартыненко
University of Tartu
1.11.2022