Для выполнения курсового проекта я выбрал научный журнал Nature - <https://www.nature.com/>

Для наполнения базы данных я не пользовался фейковыми данными, а написал скрэйпер на Python, который собрал для меня всю необходимую информацию по всем выпускам журнала за 2021 год.

Основной сущностью является научная статься – article, размещаемая в таблице articles.

Пример статьи - https://www.nature.com/articles/d41586-021-01287-2

Основные аттрибуты:

Id – идентификационный номер

Issue\_id – номер выпуска

category\_id – ссылка на категорию

title - название

body – текст (сократил его до 200 символов, чтобы не раздувать размер файлов)

is paid – BOOLEAN значение, обозначающее, находится статья в свободном доступе (0) или платная (1)

created\_at – дата публикации статьи.

Помимо таблицы articles, база данных включает ещё 12 таблиц:

Volumes – тома, в которые собираются журналы

Issues – журналы

Authors – авторы статей

Organizations – организации, в которых состоят авторы

Categories – содержит названия категорий (раздел журнала), к которым относится статья (каждая статья имеет только одну категорию)

Subjects – Темы, затрагиваемые в статье (каждая статья может затрагивать несколько тем)

Media – хранит url ссылки на файлы, используемые в статьях и на обложках журналов

Recommendations – таблица отношений, указывающая какие статьи рекомендованы на страницах статьи.

Author\_organization – указывает в каких организациях состоит автор.

Article\_media – ссылки от статей на media.

Article\_author – таблица, указывающая на авторов статей (обычно одна статья имеет большое количество авторов).

База данных связывает научные статьи, авторов, организации, позволяет выполнять широкий спектр запросов для исследования трендов в мире научных публикаций, поиску самых продуктивных авторов и институтов.

В связи с тем, что я использовал реальные данные в том виде, как они представлены на сайте, некоторые принятые мной решения могут показаться не самыми оптимальными, а именно:

1. Поле name в таблице authors является уникальным, потому что сайт не представлял никакой возможности как-то по-другому проидентифицировать автора (у них нет своих личных страниц), поэтому в рамках данной учебной базы данных поле name у автора считается уникальным.
2. Address в таблице организаций не разбит на доп. столбцы как город, страна и т.д., а указан так, как он виден на сайте, без дальнейшего анализа. В рамках данного проекта, как я посчитал, это значения не имеет. Можно было написать парсер, вычленяющий страну, город из адреса, или сфейковать данные, но я посчитал это нецелесообразным.

Файлы проекта:

project\_tables – файл создания таблиц

insert\_nature – файл для вставки данных

nature.mwb – диаграма, сгенерированная в mysql workbench. В случае, если будут проблемы с открытием, приложил еще и изображение диаграмы – diagram.png.

Остальные файлы – представления, процедура, триггеры и характерные выборки.