

«ЭПОХА Данных»: Экономический Проект Обработки, Хранения и Анализа данных

Большинство развитых стран (например, Великобритания¹, Германия², Франция³, Канада⁴, Австралия⁵, Швеция⁶, Нидерланды⁷) имеют агрегаторы социально-экономической информации, а в некоторых случаях (к примеру, Германия) — любой статистической информации вне зависимости от области научного знания. Все такие проекты создаются и развиваются либо на базе консорциума, состоящего из университетов, научных центров, органов официальной статистики, министерств и т.д. (как в Германии и Франции), либо на базе центра, который сам по себе является таким консорциумом (например, в Австралии).

До настоящего времени в России отсутствовал подобный единый агрегатор данных, используемых для научных исследований и экспертно-аналитической работы. Научные организации и ВУЗы для своих исследований вынуждены самостоятельно собирать и обрабатывать данные из множества различных источников, отслеживать их обновления и обеспечивать их сопоставимость. Все это вызывало и вызывает большие временные затраты на поиск и предварительную обработку данных.

Институт экономической политики имени Е. Т. Гайдара разрабатывает проект «ЭПОХА Данных»⁸, который является агрегатором открытой социально-экономической, финансовой, ведомственной статистики, административных и больших данных. Проект является открытым. На сегодняшний день платформа объединяет информацию из почти сотни источников и дает доступ к более чем 50 тыс. наборов данных и 600 тыс. временных рядов, покрывающих макроэкономику, финансы, отраслевую статистику, социальную сферу и демографию.

«ЭПОХА Данных» — своего рода «единое окно» открытой статистической информации по экономике и финансам, являющееся одновременно и цифровой инфраструктурой доступа к данным и инструментом для их обработки, и методическим центром, обеспечивающим корректность данных и оперативность их обновления. Важной особенностью проекта является то, что вся его технологическая инфраструктура базируется на решениях с открытым исходным кодом и российских облачных мощностях.

Какие возможности для исследователей и аналитиков предоставляет «ЭПОХА Данных»?

¹ <https://www.ukdataservice.ac.uk/> и <https://www.data-archive.ac.uk/>.

² <https://www.konsortswd.de/en/about/ratswd/>.

³ <https://db.nomics.world/>.

⁴ <https://odesi.ca/en>.

⁵ <https://ardc.edu.au/>.

⁶ <https://snd.se/en>.

⁷ <https://dans.knaw.nl/en/>.

⁸ В настоящий момент все ресурсы проекта расположены на домене Национального центра социально-экономических данных (<https://nsedc.ru/>). До конца 2026 г. предполагается перенос платформы на домен проекта «ЭПОХА Данных» www.dataera.ru.

Во-первых, это доступ к озеру первичных данных (не только статистических, но и текстовых и т.д.) и к платформе открытых статистических данных, структурированных в виде временных рядов и панелей показателей, из российских и зарубежных источников с возможностью их визуализации. Все данные можно скачать в различных форматах. Озеро первичных данных и платформа статистики объединены общей поисковой системой, что позволяет существенно сокращать время поиска. Все данные (как в озере, так и на платформе статистики) автоматически и оперативно обновляются вслед за первоисточником, а все их винтажи хранятся в архиве. Наличие полных метаданных позволяет верифицировать корректность данных и осуществлять быстрый поиск. Ключевая технологическая особенность платформы — разрабатываемый ИИ-ассистент на основе больших языковых моделей (LLM) с использованием RAG-технологий. Система автоматизирует путь пользователя: от формулировки вопроса на естественном языке до автоматического подбора релевантных наборов данных, генерации скриптов для их сборки и выдачи готового датасета.

Во-вторых, доступ к инструментам аналитической обработки, включающим методы статистики, эконометрики и машинного обучения и дающим возможность визуализации, моделирования и прогнозирования данных. Этот блок связан с платформой открытой статистики: встроенная в него поисковая система подберет показатели непосредственно с платформы, а ИИ-эконометрист поможет разобраться в тонкостях предлагаемых процедур и методов. У исследователя есть возможность как сохранять свои персональные корзины данных, обновляющиеся вместе с обновлением данных на платформе статистики, так и сохранять результаты оценки моделей и делать необходимые перерасчеты после обновления данных, а также, при желании, предоставлять доступ к ним другим исследователям через репозиторий моделей.

Третий важный элемент — исследовательская экосистема, которая позволяет, с одной стороны, получать доступ через систему дистанционных рабочих мест к большим данным и к чувствительным ведомственным и корпоративным данным, а с другой, обмениваться с коллегами как результатами своих исследований, так и данными. Последняя возможность осуществляется через репозиторий исследовательских данных и моделей, который уже упоминался выше. Он представляет собой хранилище наборов данных, на основе которых были проведены те или иные исследования, что позволяет верифицировать опубликованные авторами результаты. Мы приглашаем к сотрудничеству научные журналы, предоставляя инфраструктуру и мощности для хранения наборов исследовательских данных из опубликованных в них статей.

Методический блок платформы включает библиотеку официальных материалов статистических органов, методические комментарии исследователей, касающиеся особенностей предоставляемых ими данных.

И, наконец, еще одним блоком проекта является цифровой архив госфинансов и госуправления⁹. Основная цель архива: сбор и сканирование исторических документов, преобразование их в машиночитаемый формат, заполнение карточек метаданных исторических документов и оцифровка исторической статистики. На данный момент архив содержит более 20 тыс. документов из библиотек, государственных и личных архивов, подборки бюджетов и других финансовых документов Российской Империи и СССР с начала 19 века до 1990 г., полные коллекции основных статистических сборников.

*М. Ю. Турунцева
(Институт Гайдара)*

⁹ <https://finlibrary.ru/s/finarchive/page/home>.