Применение ИИ для гармонизации данных

Гармонизация метаданных и данных национальных статистических служб в хранилище данных

Владимир Некрасов
ООО «Контур Компонентс»

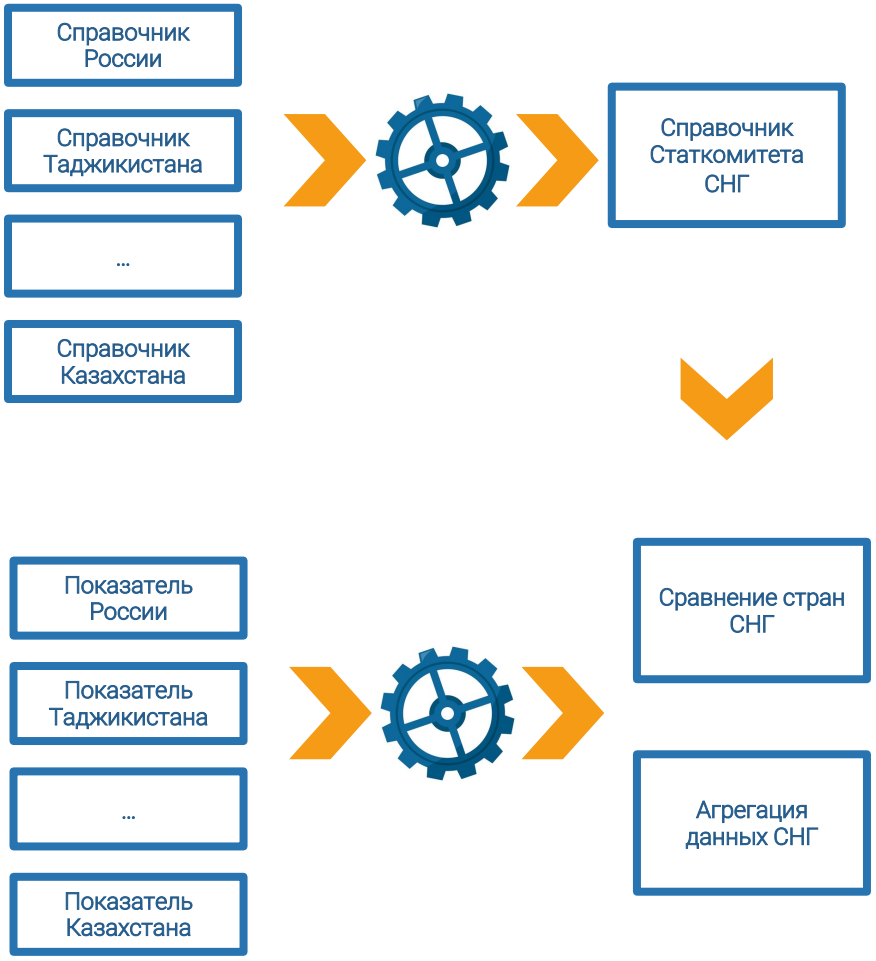
Что такое гармонизация?

- Гармонизация – это преобразование метаданных и данных, собранных из разных источников, в единый формат, в единую систему классификации, в однородную периодичность, в общие единицы измерения и масштаб
- Гармонизация метаданных
  - Сопоставление показателей
  - Сопоставление справочников и классификаторов
  - Сопоставление записей справочников и классификаторов (переходные ключи)
- Гармонизация данных
  - Приведение к единым единицам измерения и масштабу
  - Приведение к единой периодичности
  - Приведение к общим уровням агрегации
  - Приведение значений данных, рассчитанных по разными методикам через коэффициенты, формулы, к сопоставимым величинам

Для чего нужна гармонизация метаданных и данных?

Гармонизация нужна для сопоставления статистики, собранной из разных источников, например:

- Статистики стран СНГ
- Статистических наборов данных международных организации



Гармонизация при сборе данных

НСС страны

НСС страны

Нацбанк страны

Минфин страны

Международная организация

API

API

Коннектор

Коннектор

Коннектор

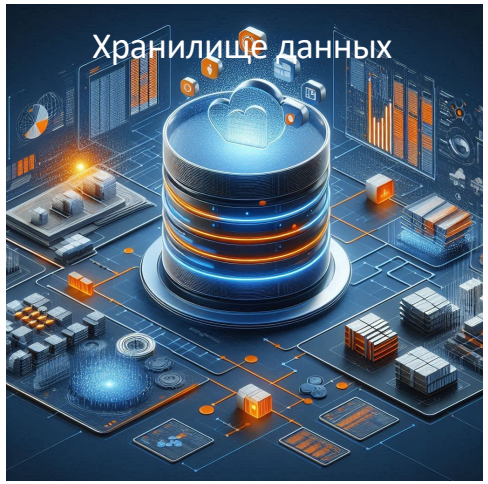
Коннектор

Локальные Показатели (Наборы данных)

Локальные справочники и классификаторы

Данные

Гармонизация



Глобальные показатели

Глобальные справочники и классификаторы

Сопоставимые данные в разрезе стран

Процесс гармонизации

Гармонизация метаданных

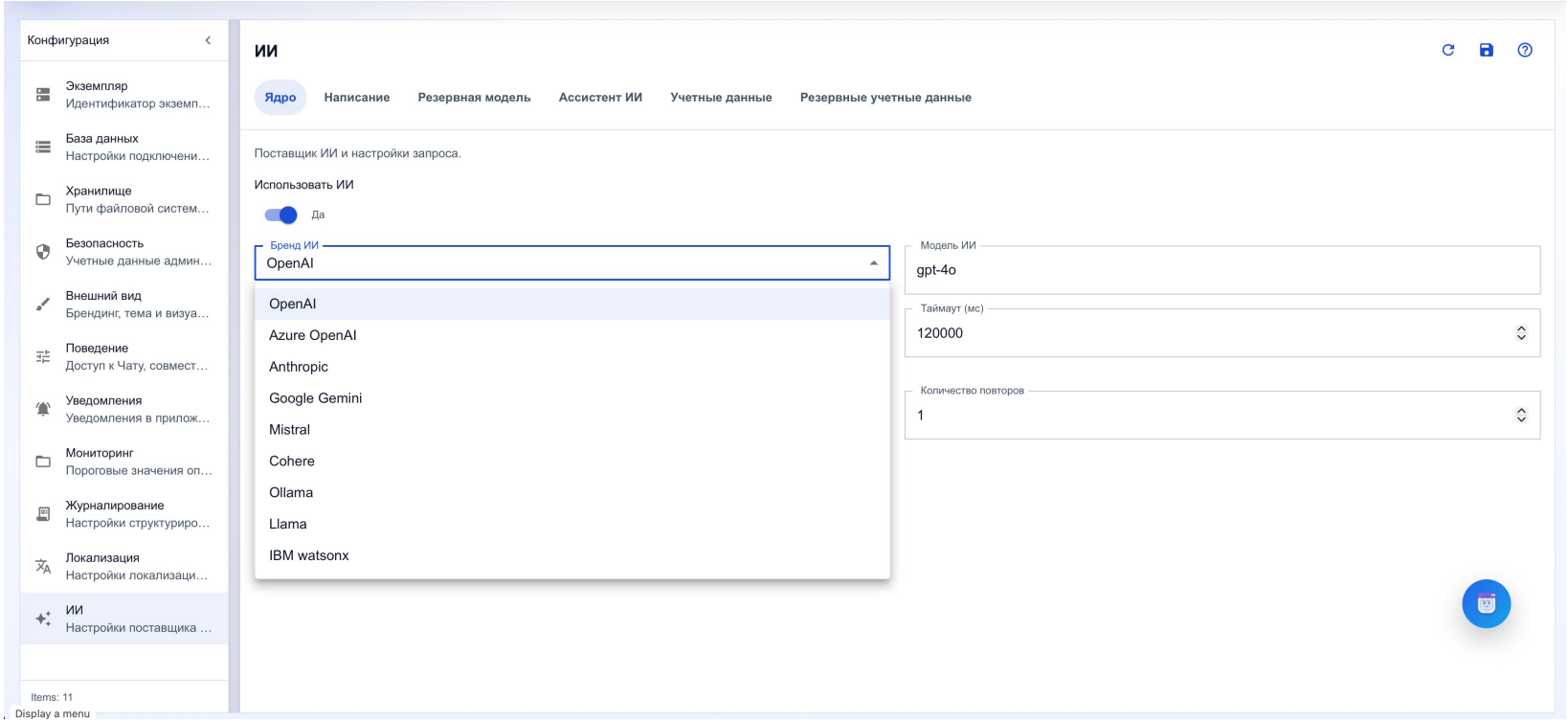
Гармонизация данных

Сбор данных	Сопоставление справочников	Сопоставление записей справочников	Сопоставление показателей	Перекодировка данных
Национальные (локальные) справочники Национальные показатели	Сопоставление национальных и глобальных справочников	Сопоставление записей национального и глобального справочника	Сопоставление национальных и глобальных показателей	Замена локальных кодов глобальными
	Переходной ключ справочников	Переходной ключ записей справочника	Переходной ключ показателей	Переходные ключи

Место искусственного интеллекта в гармонизации

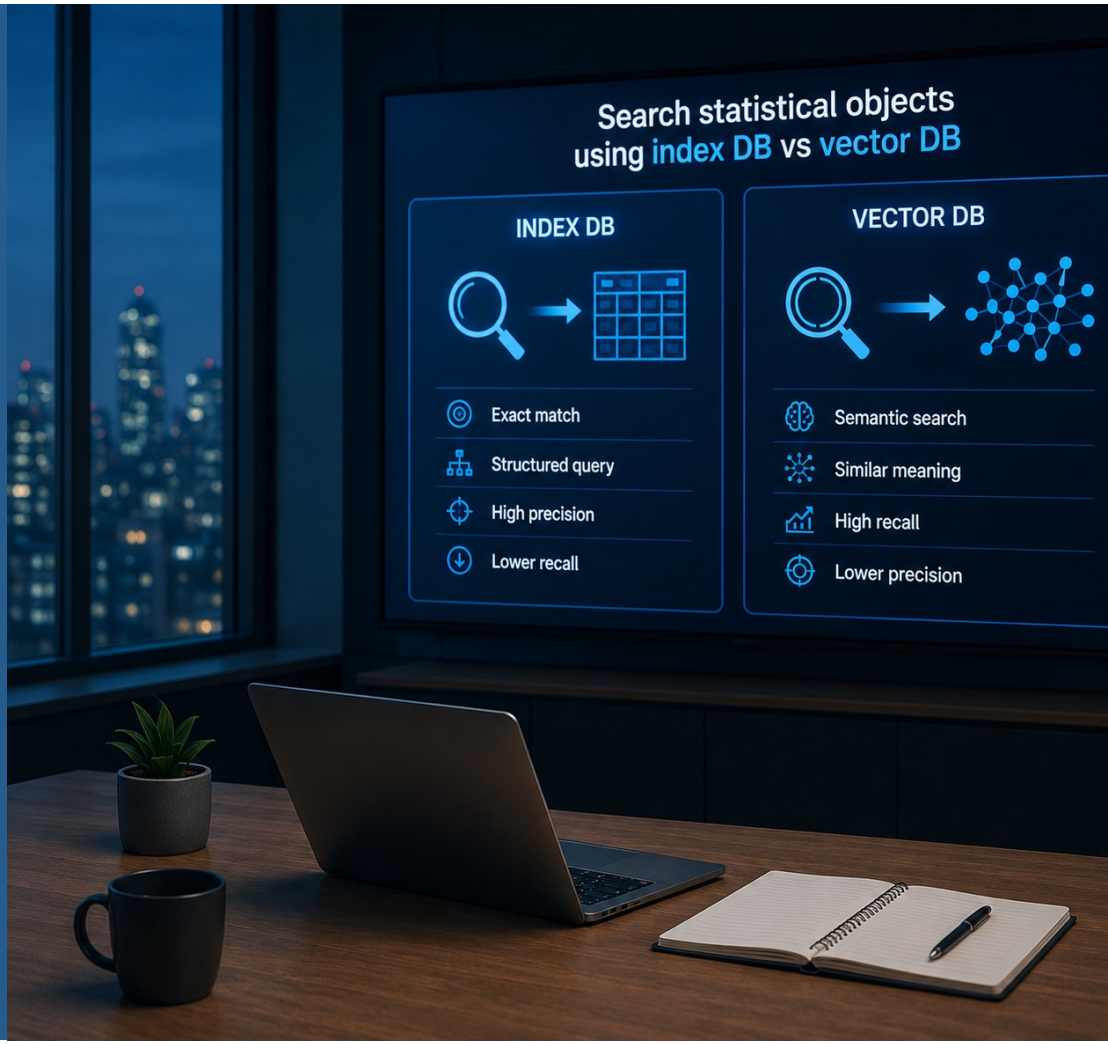
- Сопоставление справочников и классификаторов
- Сопоставление записей справочников и классификаторов
- Сопоставление показателей

Встроенный ИИ



Методы сопоставления объектов с помощью ИИ

- Оффлайн – индексная база данных, генерация искусственным интеллектом кода – процедуры поиска, работа процедуры без ИИ
- Онлайн – векторная база данных, онлайн работа агента ИИ при сопоставлении:
  - Облачный ИИ
  - Локальный ИИ



Сопоставления объектов с помощью агента ИИ

- Разработка агента ИИ для доступа к хранилищу данных и взаимодействия с пользователем
- Генерация векторной базы данных, содержащей до миллиардов атрибутов показателя, справочника, записи справочника
- Генерация векторной записи по объекту и поиск ее в векторной базе

Преимущества – семантическое сопоставление неочевидных пар

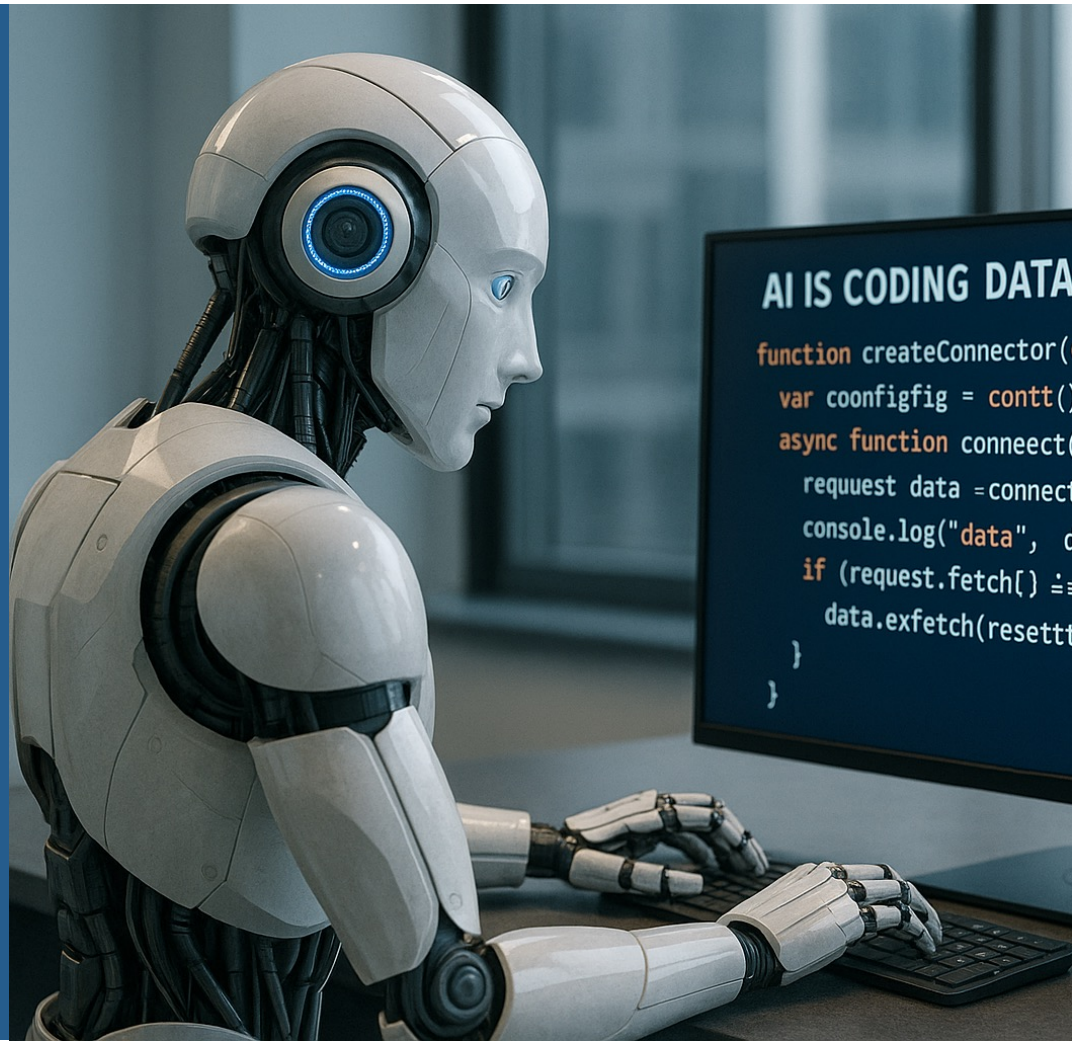
Минусы – дорого, в случае с облачным ИИ, не конфиденциально, могут быть ошибки

Сопоставления объектов с помощью процедуры, разработанной ИИ

- Разработка процедуры сопоставления с помощью ИИ
- Создание индексной базы данных, содержащей до сотен атрибутов показателя, справочника, записи справочника
- Поиск объекта по атрибутам процедурой без ИИ

Преимущества – бесплатно, конфиденциально, точное сопоставление

Минусы – процент сопоставления ниже



Обсуждение с агентом ИИ

- Агент ИИ, встроенный в BI платформу, имеет возможности:
  - Выполнение запросов к метаданным и данным Хранилища данных
  - Выполнение агрегации и сравнения данных в разных разрезах в OLAP-кубе
  - Ведение текстового и голосового диалога с экспертом
- Эксперт может:
  - Задавать вопросы по структуре данных
  - Давать задания по анализу разрезности национальных и глобальных показателей, сравнению по сходности значений агрегатов

Взаимодействие ИИ и экспертов

Порядок работы:

- Автоматическая настройка переходных ключей с помощью ИИ
- Проверка и корректировка экспертами
- Обучение ИИ
- В следующем цикле – более высокий процент точного сопоставления

Автоматическая перекодировка

- Процедура загрузки данных выполняет автоматическую перекодировку данных страны с локальных на глобальные справочники
- Исходная кодировка сохраняется в хранилище данных для выверки

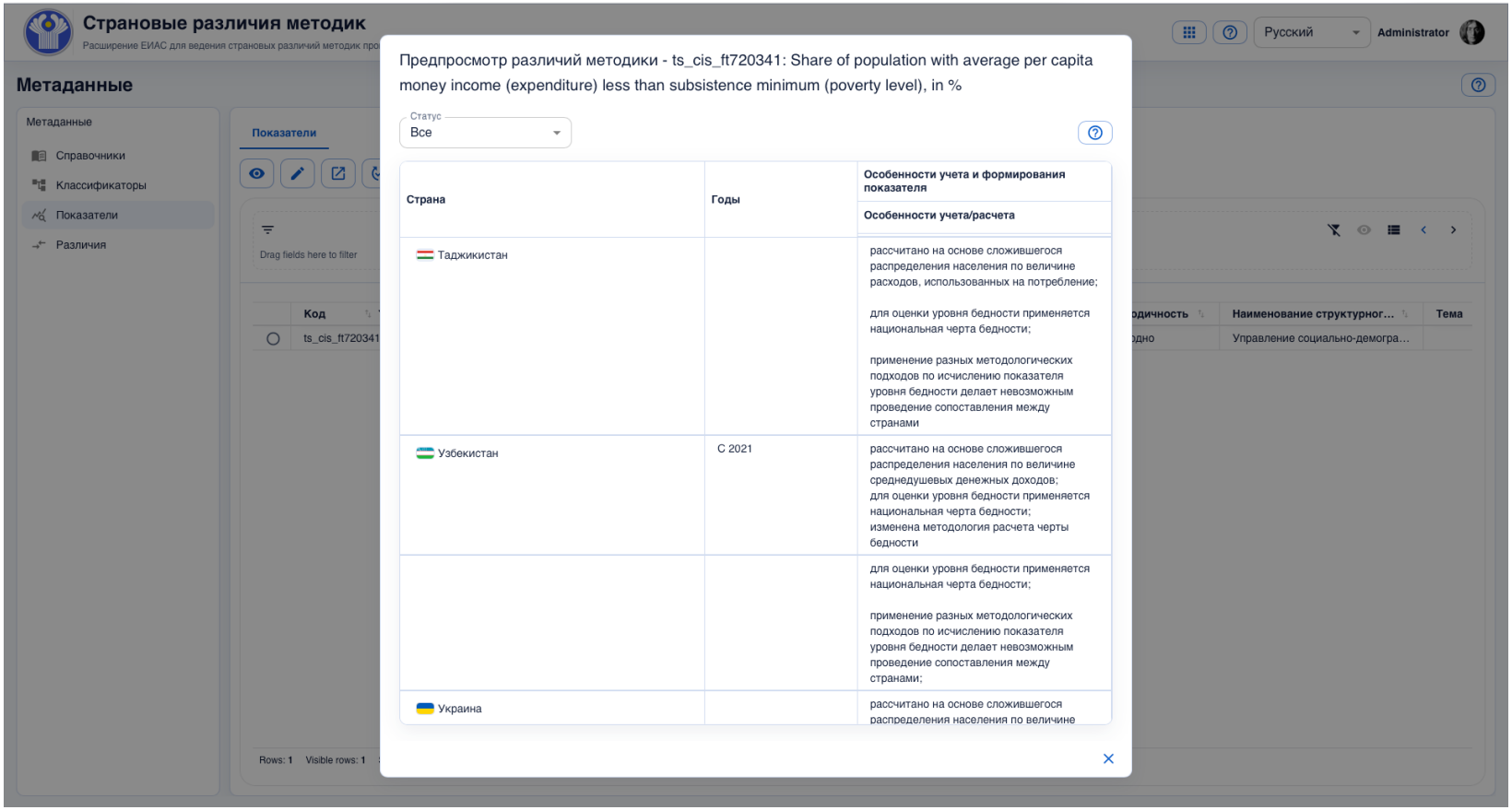
Финальная проверка – многомерный анализ

- Финальная проверка результата выполняется с помощью многомерного анализа данных – OLAP кубов
- Проверка выполняется экспертом, сравниваются значения показателей в разрезах исходной национальной кодировке и глобальной единой кодировке

Различия методик производства показателей

- Самая сложная часть сопоставления – методические различия
- Данные могут быть совершенно несопоставимыми по невидимым в метаданных причинам – различиям методик, применяемых странами
- Мы оставим эту тему на будущее

Различия методик производства показателей



Спасибо за внимание!