

Применение ИИ для гармонизации данных

Гармонизация метаданных и данных национальных статистических служб в хранилище данных



Владимир Некрасов
ООО «Контур Компонентс»

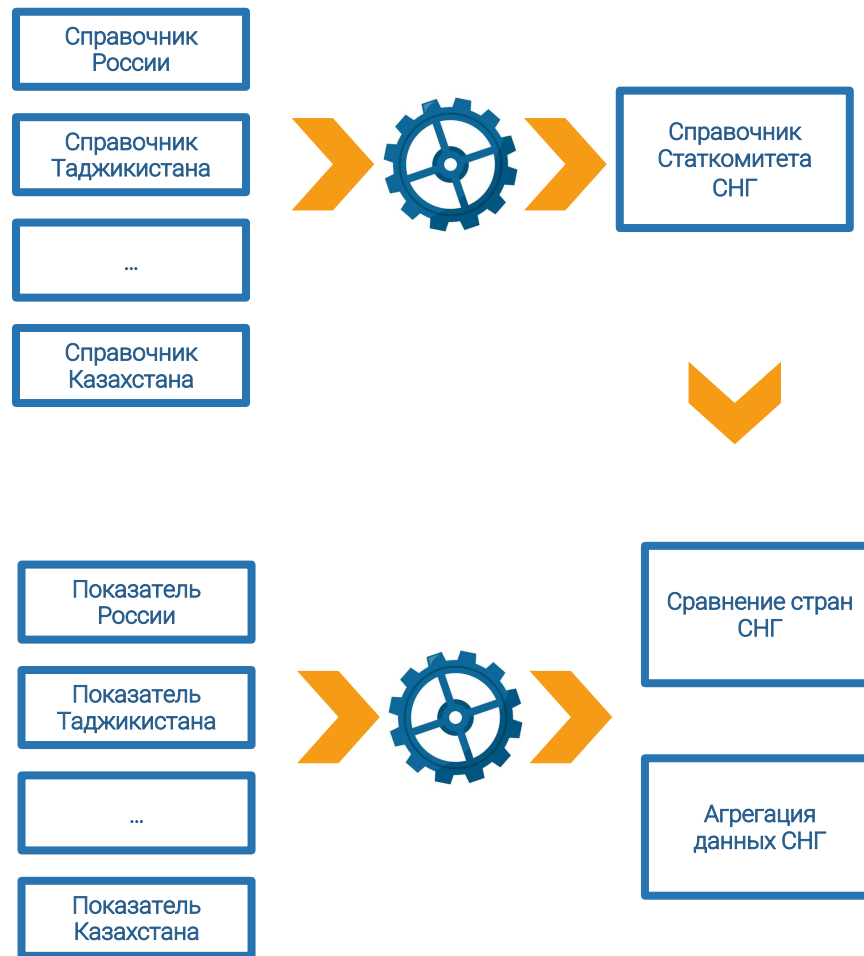
Что такое гармонизация?

- 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

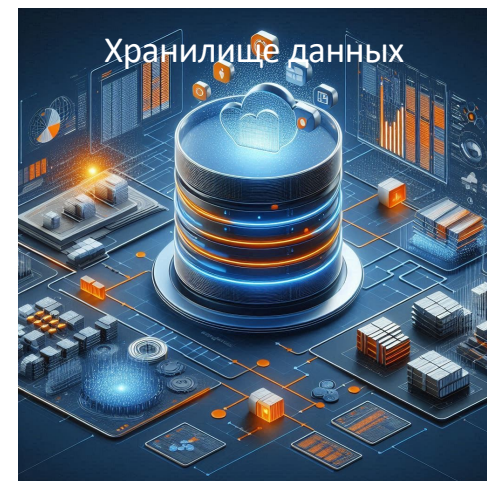
Для чего нужна гармонизация метаданных и данных?

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

- Статистики стран СНГ
- Статистических наборов данных международных организации



Гармонизация при сборе данных



Процесс гармонизации

Гармонизация метаданных

Гармонизация данных

Сбор данных

Национальные
(локальные)
справочники
Национальные
показатели

Сопоставление
справочников

Сопоставление
национальных и
глобальных
справочников

Переходной ключ
справочников

Сопоставление
записей
справочников

Сопоставление
записей
национального и
глобального
справочника

Переходной ключ
записей справочника

Сопоставление
показателей

Сопоставление
национальных и
глобальных
показателей

Переходной ключ
показателей

Перекодировка
данных

Замена локальных
кодов глобальными

Переходные ключи

Место искусственного интеллекта в гармонизации

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



Встроенный ИИ

The screenshot shows a configuration page for AI services. On the left is a navigation sidebar with categories like 'Экземпляр', 'База данных', 'Хранилище', 'Безопасность', 'Внешний вид', 'Поведение', 'Уведомления', 'Мониторинг', 'Журналирование', 'Локализация', and 'ИИ'. The 'ИИ' section is selected. The main area is titled 'ИИ' and contains tabs for 'Ядро', 'Написание', 'Резервная модель', 'Ассистент ИИ', 'Учетные данные', and 'Резервные учетные данные'. The 'Ядро' tab is active, showing a toggle for 'Использовать ИИ' (checked) and a dropdown menu for 'Бренд ИИ' with 'OpenAI' selected. A list of other brands is visible: Azure OpenAI, Anthropic, Google Gemini, Mistral, Cohere, Ollama, Llama, and IBM watsonx. To the right are input fields for 'Модель ИИ' (gpt-4o), 'Таймаут (мс)' (120000), and 'Количество повторов' (1). A blue chat icon is in the bottom right corner. At the bottom left, it says 'Items: 11' and 'Display a menu'.

Конфигурация <

ИИ

Ядро Написание Резервная модель Ассистент ИИ Учетные данные Резервные учетные данные

Поставщик ИИ и настройки запроса.

Использовать ИИ

Да

Бренд ИИ

OpenAI

OpenAI

Azure OpenAI

Anthropic

Google Gemini

Mistral

Cohere

Ollama

Llama

IBM watsonx

Модель ИИ

gpt-4o

Таймаут (мс)

120000

Количество повторов

1

Items: 11

Display a menu

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

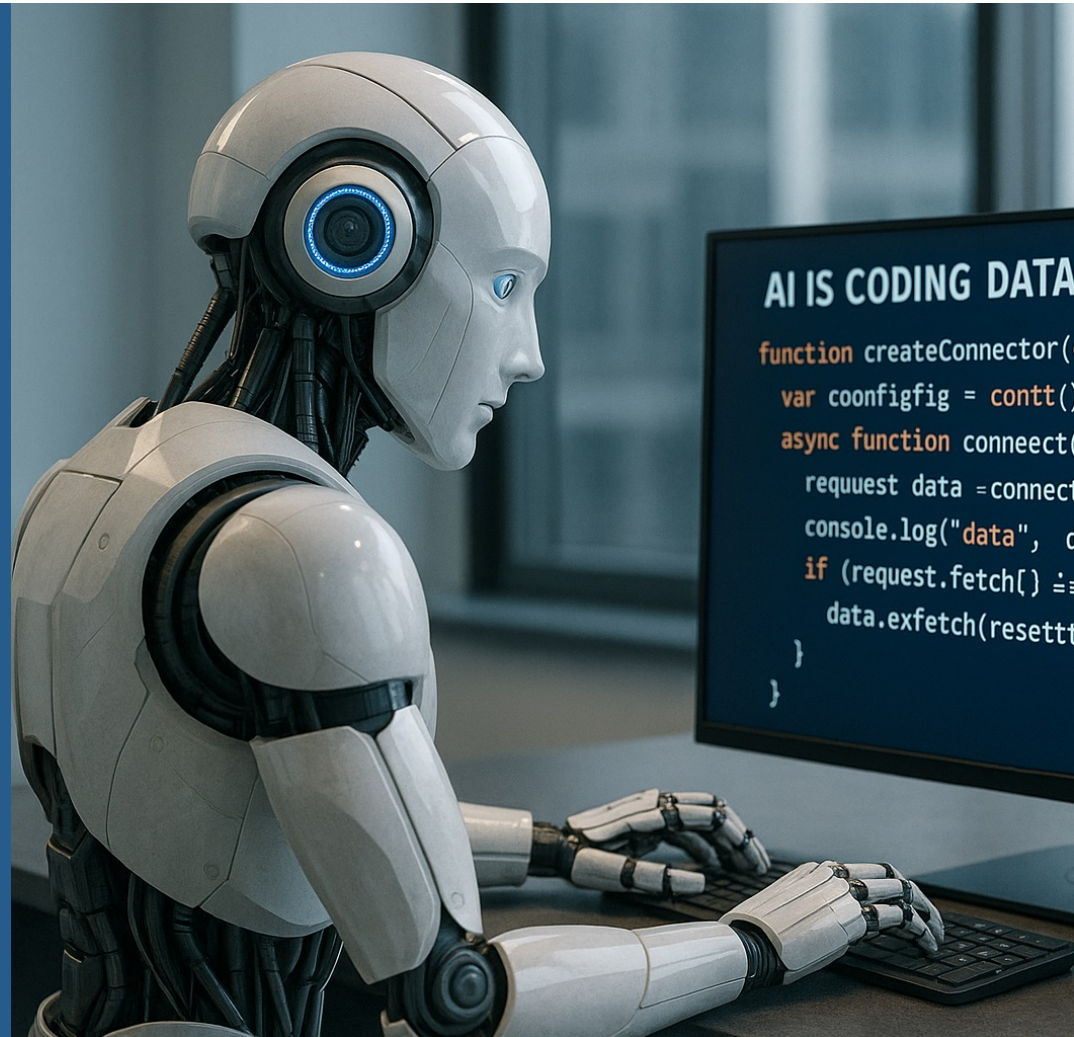


Сопоставления объектов с помощью процедуры, разработанной ИИ

- 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 схожести 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




INDICATOR	SURVEY-BASED (ILO)	ADMINISTRATIVE (REGISTER-BASED)	MODEL-BASED (ESTIMATED)
 Unemployment rate	 Share of unemployed persons in the labor force	 Share of registered unemployed in the labor force	 Estimated share of unemployed in the labor force
Definition	Household population (15+ years)	Registered unemployed (15+ years)	Total population (15+ years)
Population	Labor force survey	Unemployment register	Multiple data sources (surveys, registers, macro data)
Data source	Sample survey	Administrative recording	Statistical modeling
Collection method	National estimates	Registered unemployed only	Complete population
Coverage	Usually 1 week (usual status approach)	Registration date	Monthly / Quarterly
Reference period	Unemployed / Labor force × 100	Registered unemployed / Labor force × 100	Model estimate
Calculation formula	Internationally comparable Captures reason for unemployment	Timely and cost-effective High timeliness	Timely estimates Available at high frequency
Pros	Sampling error More expensive	Not internationally comparable Excludes unregistered unemployed	Model assumptions Less transparent
Cons			

Различия методик производства показателей

Страновые различия методик
Расширение ЕИАС для ведения государственных различий методик производства показателей

Статус: Все

Предпросмотр различий методики - ts_cis_ft720341: Share of population with average per capita money income (expenditure) less than subsistence minimum (poverty level), in %

Страна	Годы	Особенности учета и формирования показателя
 Таджикистан		рассчитано на основе сложившегося распределения населения по величине расходов, использованных на потребление; для оценки уровня бедности применяется национальная черта бедности; применение разных методологических подходов по исчислению показателя уровня бедности делает невозможным проведение сопоставления между странами
 Узбекистан	С 2021	рассчитано на основе сложившегося распределения населения по величине среднедушевых денежных доходов; для оценки уровня бедности применяется национальная черта бедности; изменена методология расчета черты бедности
		для оценки уровня бедности применяется национальная черта бедности; применение разных методологических подходов по исчислению показателя уровня бедности делает невозможным проведение сопоставления между странами;
 Украина		рассчитано на основе сложившегося распределения населения по величине

Rows: 1 Visible rows: 1

Спасибо за внимание!