Модели данных

А4. Модели справочников и реестров

Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

сен 2024, окт 2025 года

Москва

Артемьев Валерий Иванович © 2024-2025

A4. Модели справочников и реестров

- Определение справочника
- Виды справочников
- Способы классификации и кодирования
- Примеры классификаторов и калькуляционных справочников
- Структуры данных справочников (линейные, иерархические и рекурсивные)
- Определение реестра (основных данных)
- Способы формирования суррогатных ключей (счётчик, последовательность, датчик случайных чисел, универсальный уникальный идентификатор UUID, альтернатива nanoid)
- Поддержка истории изменений справочников и реестров

Вопросы для самопроверки

Актуализировать, когда устоится

Введение (1)

Справочные и реестровые данные критичны для бизнеса.

- Наличие и их качество существенно влияют на ведение бизнеса
- Возникают операционные, управленческие и репутационные риски
- Важно авторитетное ведение реестров и классификаторов
- Раннее был общий термин **НСИ** – нормативно-справочная информация

Реестры ещё называют основными данными или мастер-данными.

Введение (2)

Справочники (классификаторы) служат для анализа данных.

Важна устойчивость и полнота схем классификации
Обобщение и клонирование классификаторов.

Реестры – для идентификации бизнес-сущностей.

Важно отсутствие дублирования экземпляров данных
Проблема идентификации разнородных сущностей в одной роли.

Справочники и реестры возникают при моделировании данных путём декомпозиции сущностей (таблиц) для нормализации отношений.

Определение справочных данных

Справочные данные – систематизированные, структурированные и кодифицированные перечни однородных по своему содержанию или сути данных.

Виды справочников:

Классификаторы – справочники, которые служат для классификации сущностей, содержат названия и описания, например, Виды экономической деятельности ОКВЭД2.

Калькуляционные справочники – справочники, которые служат для расчётов, содержат названия и числа, например: Курсы валют.

Определение реестровых (основных) данных

Реестровые (основные) данные – идентификационные и контекстные данные о бизнесе в форме относящихся к нему сущностей (клиенты, продукты, сотрудники и операции).

Идентификация – различение экземпляра сущности путём его регистрации и дальнейшего указания регистрационного кода в данных.

Обобщённое краткое название – реестры (списки, перечни, регистры)

Не редко путают реестровые и справочные данные, например, классификатор стран мира ОКСМ является по сути реестром. Многие ИТ-спецы называют всё справочниками.

Нормативно-справочная информация, применяемая в Банке России

Область применения	Классификаторы	Реестры	Калькуляционные справочники
Международная НСИ	КЛВ, CFI	SWIFT, LEI	Котировки редких валют из Reuters, Выходные и праздничные дни Еврозоны
Общероссийская (федеральная) НСИ	ОКСМ, ОКВ, ОКВЭД2, ОКИН ОКУД, ОКФИ, ОКАТО, ОКТМО, ОКЭР, ОКПО, ОКФС, ОКДП, ОКОФ, ОКОПФ, ОКЕИ	ЕГРЮЛ, ЕГРИП, ГАР	Справочник выходных и рабочих дней по России
Отраслевая НСИ	ПСБУ КО, ПСБУ НФО, Справочник драгоценных металлов	КГРКО, РУФР, БИК, СВИФТ, АСВ, Реестр ценных бумаг, Реестр эмитентов, Аудиторы	Справочник курсов валют, Справочник учетных цен на аффинированные драгоценные металлы Нормативы достаточности капитала
Ведомственная НСИ	ПСБУ БР	СКП, ТАСБР, Оргструктура	Штатное расписание
Локальная НСИ	Классификаторы отдельных приложений	Реестры отдельных приложений	Калькуляционные справочники отдельных приложений

Примеры реестров

Федеральные реестры

ЕГРЮЛ – Единый государственный Реестр юридических лиц

ЕГРИП – Единый государственный Реестр индивидуальных предпринимателей

ГАР – Государственный адресный реестр

Отраслевые реестры финансового рынка

КГРКО – Книга государственной регистрации кредитных организаций

РУФР – Реестр участников финансовых рынков

БИК – Банковские идентификационные коды

АСВ – Участники Агентства страхования вкладов

Примеры классификаторов и калькуляционных справочников

Общероссийские классификаторы

ОКСМ – страны мира
ОКВ – валюты
ОКЕИ – единицы измерения
ОКАТО – административно-территориальные образования
ОКЭР – экономические регионы
ОКВЭД2 – виды экономической деятельности
ОКПО – предприятия и организации

Калькуляционные справочники

Курсы валют
Котировки драгметаллов

Классификация данных

Классификация данных – распределение экземпляров сущностей (предметов, явлений, процессов, людей, организаций и понятий) по классам согласно определённым признакам для выделения экземпляров с однородными свойствами.

Методы классификации

- *иерархическая* классификация
- *фасетная* классификация
- *дескрипторная* классификация

По-разному определяются и используются классификационные признаки.

Иерархический метод классификации

Метод классификации, при котором заданное множество объектов классификации ***последовательно*** *делится на классификационные группировки по какому-либо признаку.*

Выбранные признаки являются определяющими в конкретной предметной области.

Последовательность признаков зависит от характера информации, частоты и вероятности обращения к признаку.

Количество ступеней классификации определяется характером решаемых задач и обычно фиксировано.

Правила иерархической классификации

- Каждый объект классификации на каждой ступени относится только к одной классификационной группировке
- Деление каждой группировки проводится только на основе одного признака
- Получаемые на каждой ступени классификации группировки не повторяются
- Не остается объектов, не вошедших в состав классификационной группировки

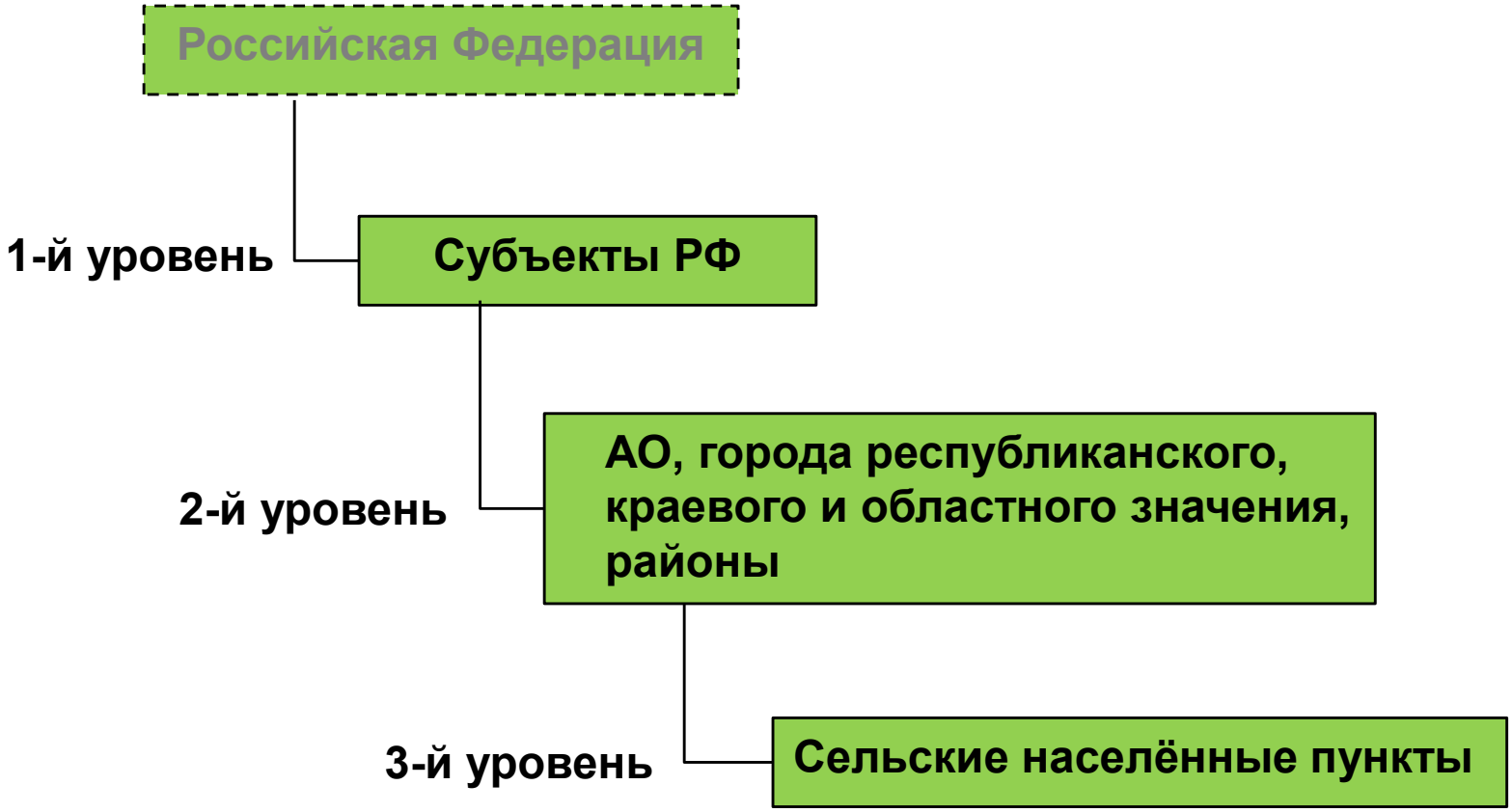
Таксономия

+ A
+ A1
A1.1
A1.2
+ A2
A2.1
A2.1.1

Можно использовать несколько альтернативных иерархий.

Пример иерархической классификации

Общероссийский классификатор административно-территориальных образований РФ ОКАТО



Фасетный метод классификации

Метод классификации, при котором заданное множество объектов классификации ***параллельно*** *делится на непересекающиеся группы по различным признакам классификации.*

Образуются непересекающиеся классификационные группы по различным аспектам или их совокупностям.

Наборы признаков классификации образуют параллельные независимые *фасеты*.

Значения признаков фасетов располагаются в иерархическом порядке или в виде простого перечисления.

Правила фасетной классификации

- Признаки в различных фасетах не повторяются (принцип взаимного исключения фасетов)
- Из всевозможных признаков, характеризующих множество объектов классификации, отбираются и фиксируются только существенные, обеспечивающие решение конкретных задач предметной области.

Пример

Общероссийский классификатор информации о населении
ОКИН (293 фасета)

01	02	03	04	
Пол	Гражданство	Национальность	Родной язык	…

Дескрипторный метод классификации

Метод классификации, при котором с экземплярами сущности связывают дескрипторы (описатели) на естественном языке – ключевые слова.

Применяется для организации поиска информации, для ведения тезаурусов (словарей).

Особенно широко она используется в библиотечной системе поиска.

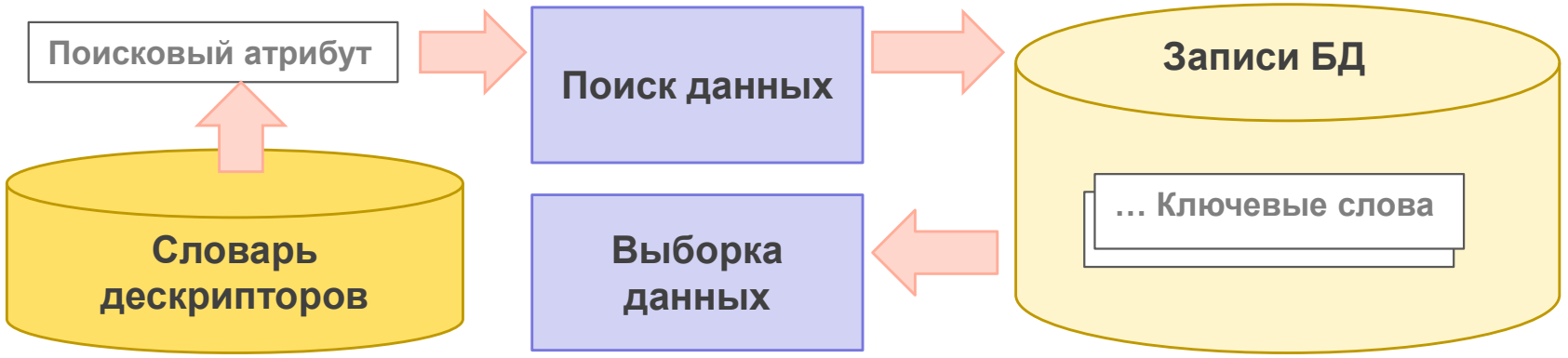
Пример

Ключевые слова этой лекции:

справочники; классификаторы; реестры; методы классификации; методы кодирования

Правила дескрипторной классификации

- Отбирается совокупность *ключевых слов* или словосочетаний, описывающих предметную область или совокупность однородных объектов
- Среди ключевых слов могут быть *синонимы*
- Из совокупности синонимов выбирается один или несколько наиболее употребимых
- Создается *словарь дескрипторов* – словарь отобранных ключевых слов и словосочетаний



Преимущества и недостатки методов классификации данных

Метод классификации	Когда применяется	Преимущества	Недостатки
Иерархический	При соподчиненных признаках классификации и стабильности решаемой задачи	• Логичность • простота построения • удобство логической и числовой обработки • возможность группировки объектов по большинству признаков • высокая информационная насыщенность	• Жестко фиксированные признаки • строгий порядок их следования • коренная переработка при изменении состава объектов, их свойств или характера решаемых задач • требуется предусматривать резерв незанятых кодов
Фасетный	Широко применяется при несоподчиненных признаках классификации и при большой динамичности решаемых задач	• Большая емкость • включение новых объектов без нарушения структуры классификатора • группировки по любому сочетанию фасетов • добавление фасетов и их значений • изменение порядка фасетов • возможность применения ограниченного количества фасетов	• Сложность структуры из-за необходимости учёта всего многообразия признаков классификации • невозможность выделения общих и различных признаков между объектами в разных классификационных группировках
Дескрипторный	Дополнительный механизм поиска и отбора данных, широко применяется для поиска данных, ведения словарей	• Гибкая система • семантическая классификация на естественном языке • расширение области поиска за счёт синонимических, родо-видовых и ассоциативных связей	• Нужна нормализация дескрипторов • требуется вести словарь дескрипторов • определённая сложность установления и ведения связей

Методы кодирования справочной и реестровой информации

Кодирование справочников и реестров – выбор алфавита и формата кода для идентификации и классификации множества объектов при автоматизированной обработке.

Методы кодирования объектов

- *классификационные*
  - последовательное кодирование
  - параллельное кодирование
- ключевые слова и словосочетания
  регистрационные
  - порядковое кодирование
  - серийно-порядковое кодирование

Порядковое кодирование

При порядковом методе кодирования кодами являются числа натурального ряда, каждый из объектов идентификации (классификации) кодируется путем присвоения ему текущего порядкового номера.

Числовой код, типы SMALLINT, INTEGER, BIGINT.

Этот метод широко используется для генерации искусственных – суррогатных ключей. Называется в разных СУБД по-разному: Sequence, Autoincrement.

Серийно-порядковое кодирование

Кодами являются числа натурального ряда с *закреплением отдельных серий этих чисел* (интервалов натурального ряда) за объектами с одинаковыми признаками идентификации (классификации).

Тип ключа CHAR(2), т.к. *код цифровой с ведущим нулём*

В каждой серии предусматривается определенное количество *резервных кодов в середине или в конце*.

Пример серийно-порядкового кодирования

Классификатор сельскохозяйственных культур

Код	Наименование сельскохозяйственной культуры
01	пшеница
02	ячмень
03	рожь
04	овес
11	листовые или стебельные овощи
12	аспарагус
21	лен
22	горчица
...	...

Последовательное кодирование

Код классификационной группы или объекта идентификации (классификации) образуется из кодов последовательно расположенных подчиненных классификационных групп, полученных при иерархической классификации.

Алфавитно-цифровые коды переменной длины, тип данных VARCHAR.

Код нижестоящей группы получается добавлением нужного количества символов к коду вышестоящей группировки.

Пример последовательного кодирования

Общероссийский классификатор административно-территориальных образований ОКАТО

Краснодарский край	03
Абинский район Краснодарского края	03 201
Варнавинский сельский округ Абинского района Краснодарского края	03 201 801
село Варнавинское Варнавинского сельского округа Абинского района Краснодарского края	03 201 801 001

Пример последовательного кодирования видов продукции

Код	Наименование вида продукции
A	продукция сельского хозяйства, охоты и сопутствующие услуги
A. 1	сезонные культуры
A. 101	зерновые культуры, бобовые культуры и масличные семена
A. 10101	пшеница
A. 101011	твердая пшеница
A. 101012	пшеница, за исключением твердой пшеницы
….	…
B	продукция лесоводства, лесозаготовок и связанные с этим услуги
B. 1	услуги лесопитомников и услуги по выращиванию саженцев
B. 101	услуги лесопитомников и услуги по выращиванию саженцев
B. 10101	живые растения лесных пород, семена лесных пород
B. 101011	живые растения лесных пород
B. 101012	семена лесных пород
….	…

Параллельное кодирование

Код классификационной группы или объекта идентификации (классификации) образуется с использованием независимых групп, полученных при использовании фасетного метода классификации.
В этом случае признаки объекта систематизации (классификации) кодируются независимо друг от друга.

Алфавитно-цифровые коды переменной длины,тип данных VARCHAR.

Пример параллельного кодирования (без установки фасетной формулы)

Общероссийский классификатор информации о населении ОКИН (293 фасета)

01 пол	02 гражданство	10 состояние в браке
1 – мужской	1 – гражданин РФ	1 – никогда не состоял(а)
2 – женский	2 – двойное гражданство	2 – состоит в зарегистрированном браке
	3 – иностранный гражданин	3 – состоит в не зарегистрированном браке
	4 – лицо без гражданства	4 – вдовец (вдова)
		5 – разведён (разведена)
		6 – разошёлся (разошлась)

Без установки фасетной формулы указываются коды фасетов: В.И. Артемьев ОКИН: **011 021 102**

Пример параллельного кодирования (с установкой фасетной формулы)

Классификация лекарственных форм

Код лекарственной формы	Наименование лекарственной формы	Основная лекарственная форма	Способ введения	Путь введения	Агрегатное состояние
01010201	аэрозоль для ингаляций дозированный	01	01	02	01
03020302	гель для инъекций	03	02	03	03
04020302	раствор для внутримышечного введения	04	02	03	02
...	...	...	...	...	...

Преимущества и недостатки методов кодирования данных

Метод кодирования	Когда применяется	Преимущества	Недостатки
Последовательное	При стабильном наборе признаков классификации и их последовательности в течение долгого времени	• логичность построения кода • большая емкость кода	• ограниченные возможности идентификации объектов • использовать этот код по частям нельзя, т.к. символы кода зависят от предыдущих • сложно вносить новые признаки • сложно изменять код без перестройки
Параллельное	Для фасетных классификаций	• гибкость структуры кода из-за независимости признаков • большое число кодовых комбинаций из небольшого числа признаков • простота включения новых признаков систематизации • наличие в коде информации о свойствах объекта • приспособленность для машинной обработки информации.	• использование его только для однородных объектов систематизации (классификации), иначе емкость классификатора будет использоваться не полностью
Порядковое	Для простых справочников, суррогатных ключей реестров	• использование коротких кодов • обеспечение однозначности объектов идентификации (классификации) • простое присвоение кодов новым объектам	• отсутствие в коде информации о свойствах объекта • отсутствие возможности размещения кодов объектов в необходимом месте классификатора, так как резерв располагается в конце ряда
Серийно-порядковое	Для объектов, имеющих 2 соподчиненных признака идентификации (классификации)	• простота кодирования	• ограниченная ёмкость кодирования

Универсальные уникальные идентификаторы

Автогенерируемые уникальные коды первичных ключей на основе времени, идентификатора узла или случайного числа.

Универсальный уникальный идентификатор UUID – стандарт *для распределённых систем*, позволяющий уникально идентифицировать данные *без центра координации*.
Созданный UUID с приемлемым уровнем уверенности *можно считать всемирно уникальным*.
Данные, помеченные UUID, могут быть помещены в общую БД *без необходимости разрешения конфликта имен*.

Реализация уникальных идентификаторов

Широко распространён способ реализации GUID фирмы Microsoft:

> UUID представляет собой 16-байтный код.
> Внешнее представление – шестнадцатеричное число в формате 8-4-4-4-12, занимает 36 символов.

Пример:

123e4567-e89b-12d3-a456-426655440000

Альтернатива Nano Id:

> Использует больший алфавит [A-Za-z0-9_-]
> Формат 8_8-3, занимает 21 символ.
> Скорость генерации на 60% выше, чем UUID.

Пример:

V1StGXR8_Z5jdHi6B-myT

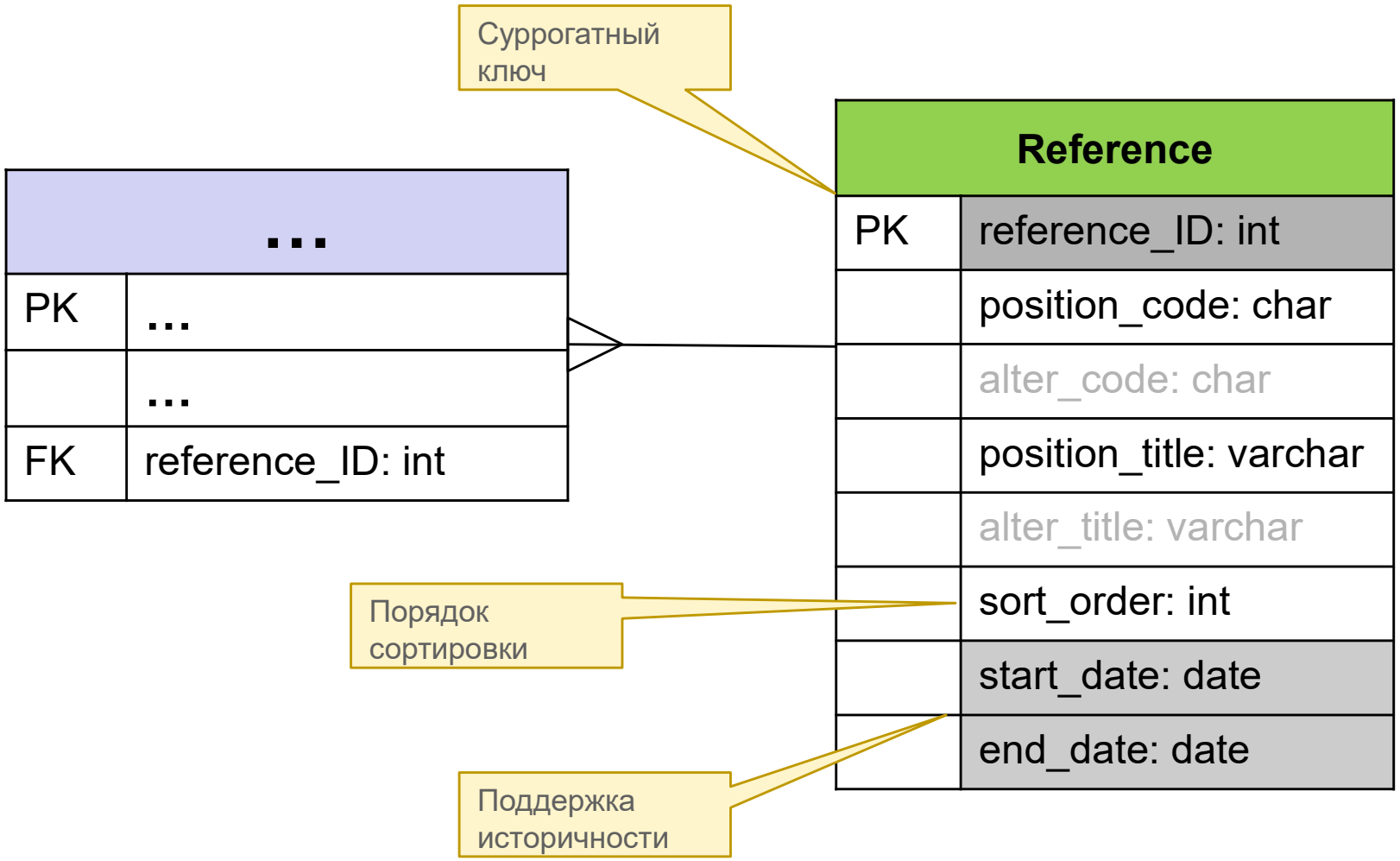
Поддержка истории изменения справочников и реестров

Справочники меняются *редко*.
Реестры могут меняться *довольно часто*.
Нужно поддерживать историю изменений особенно *для задач анализа данных*.

Основные способы для поддержки истории изменений:

1. Длинная история
   - Добавление строки с периодом актуальности данных (start_date, end_date)
   - Обновление периода актуальности старой строки
2. Короткая история
   - Обновление строки с сохранением нового и старого значения важного атрибута и даты обновления

Структура данных линейных справочников



Пример: общероссийский классификатор стран мира (ОКСМ 025-2001)

Идентификация стран мира при обмене информацией и решении задач

На основе ISO 3166-97 "Коды для представления наименований стран"

Объекты классификации – страны, интересные с точки зрения торговли, транспорта и т.д.

Цифровой код (порядковое кодирование) предпочтительнее, на него не влияют изменения в наименованиях. Буквенные коды для визуальной ассоциации с наименованиями

Цифровой код

Буквенные коды

Краткое и полное наименование страны

Country	
PK	country_ID
	num: char(3)
	alfa2: char(2)
	alfa3: char(3)
	short_name: varchar
	full_name: varchar
	start_date: date
	end_date: date

num	alfa2	alfa3	short_name	full_name
112	BY	BLR	БЕЛАРУСЬ	Республика Беларусь
156	CN	CHN	КИТАЙ	Китайская Народная Республика
276	DE	DEU	ГЕРМАНИЯ	Федеративная Республика Германия
643	RU	RUS	РОССИЯ	Российская Федерация
840	US	USA	СОЕДИНЁННЫЕ ШТАТЫ	Соединённые Штаты Америки
...	...	...	...	...

Пример: справочник курсов валют

Таблица CURRENCIES – список курсов валют (Oracle)

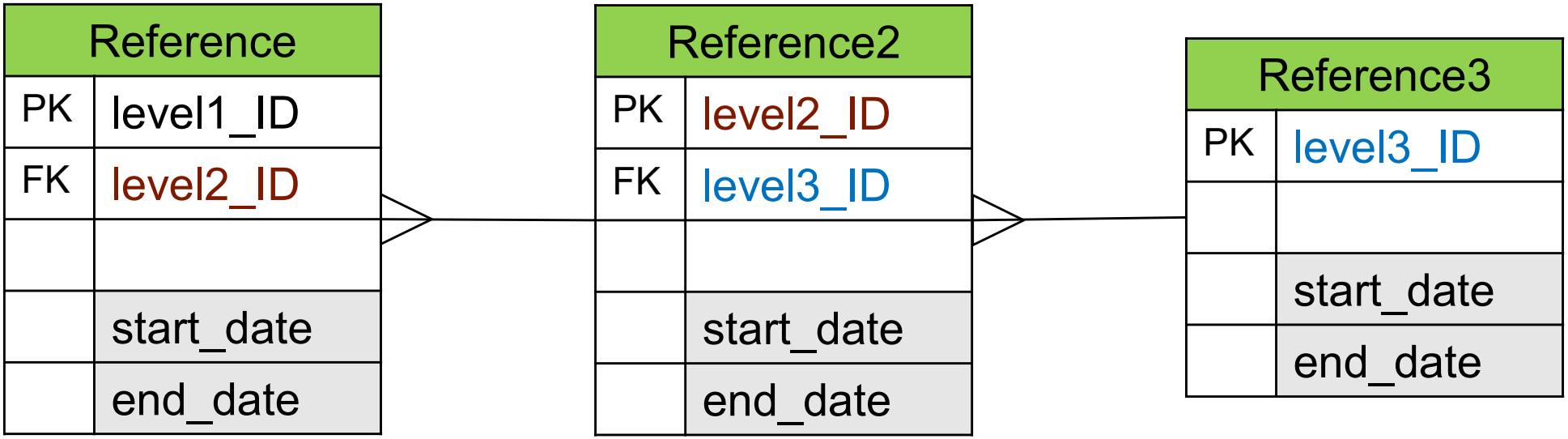
Назначение	Идентификатор	Тип (размер)
Уникальный номер строки	ID	NUMBER(12)
Цифровой код валюты	CURRENCY	NUMBER(3)
Курс валюты	COURSE	NUMBER(20,6)
Масштаб	SCALE	NUMBER(7)
Дата начала действия курса	EFF_DATE	DATE
Номер документа	DOC_NUMB	VARCHAR2(8)
Дата издания документа	DOC_DATE	DATE
Системная дата создания строки	CC_DATE	DATE

Структуры данных иерархических справочников

1-й уровень справочника

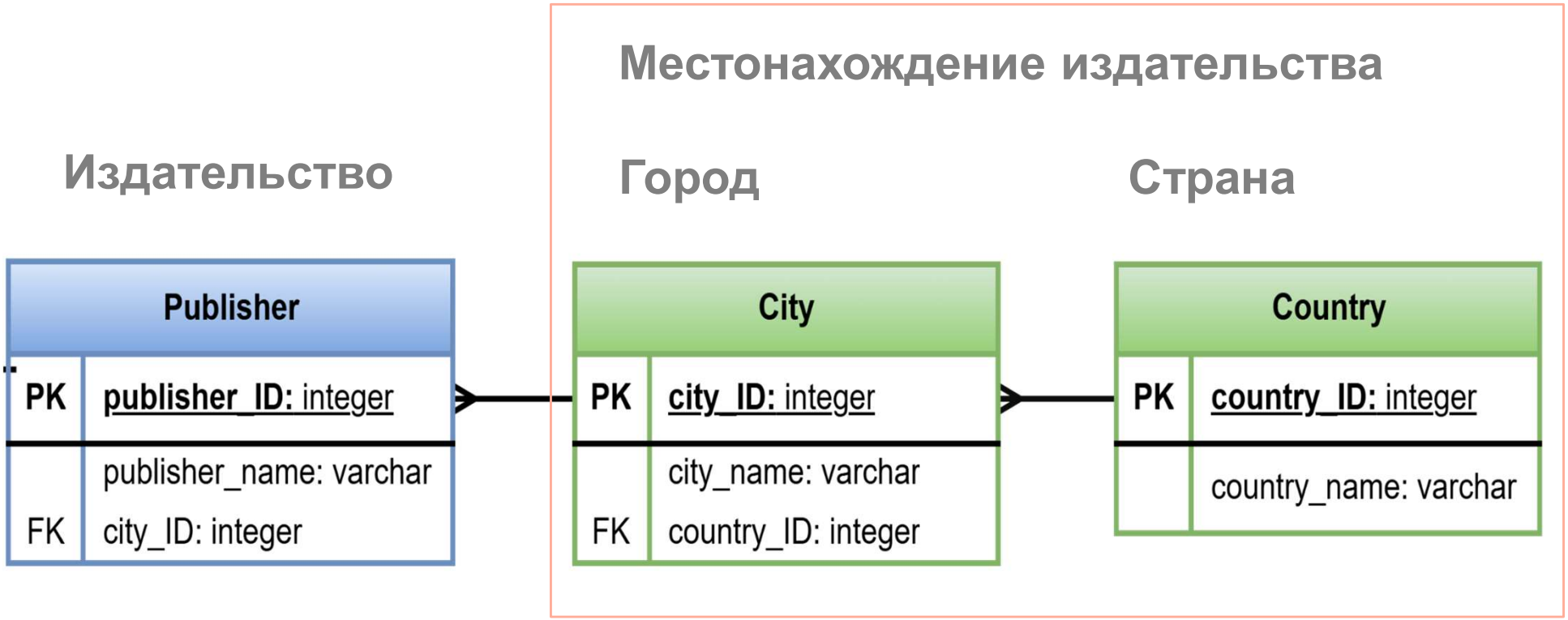
2-й уровень справочника

3-й уровень справочника



Пример иерархического справочника

Фрагмент физической модели данных «Домашняя библиотека»



Этот справочник является частью реестра издательств

В качестве заключения

- Справочники и реестры *критичны для бизнеса*.
- Классификаторы служат *для анализа данных*, калькуляционные справочники – *для расчётов*, реестры – *для идентификации бизнес-сущностей*.
- Существуют *иерархическая, фасетная и дескрипторная классификация* бизнес-сущностей.
- Существуют *последовательное, параллельное, порядковое, серийно-порядковое кодирование* объектов.
- UUID служит *для глобальной идентификации объектов в распределённых системах* (например, сделок).
- Для «правильного» вывода записей справочника нужен *атрибут порядка сортировки*.
- Для поддержки истории изменения справочника и реестра служат *даты актуальности записей*.

Терпения и удачи всем, кто связан с моделированием данных

Спасибо за внимание!

Валерий Иванович Артемьев

Департамент данных, проектов и процессов
Банк России

Тел.: +7(495) 753-96-25

e-mail: avi@cbr.ru