Модели данных

А4. Модели справочников и реестров

Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

сен 2024 года

Москва

А4. 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

Актуализировать, когда устоится

Вопросы для самопроверки

Введение

Справочные и реестровые (основные) данные – критичные для бизнеса данные.

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

Представлены в виде реестров и справочников:

Справочники (классификаторы) – аналитическая ценность

- Важна устойчивость и полнота схем классификации
- Обобщение и клонирование классификаторов.

Реестры – идентификация 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).

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

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

Примечания:

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 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 Справочники форм отчётности	СКП, ТАСБР	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

Ведомственные реестры

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

Методы классификации данных

Классификация – 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.

Методы классификации в Data Science имеют другое содержание.

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

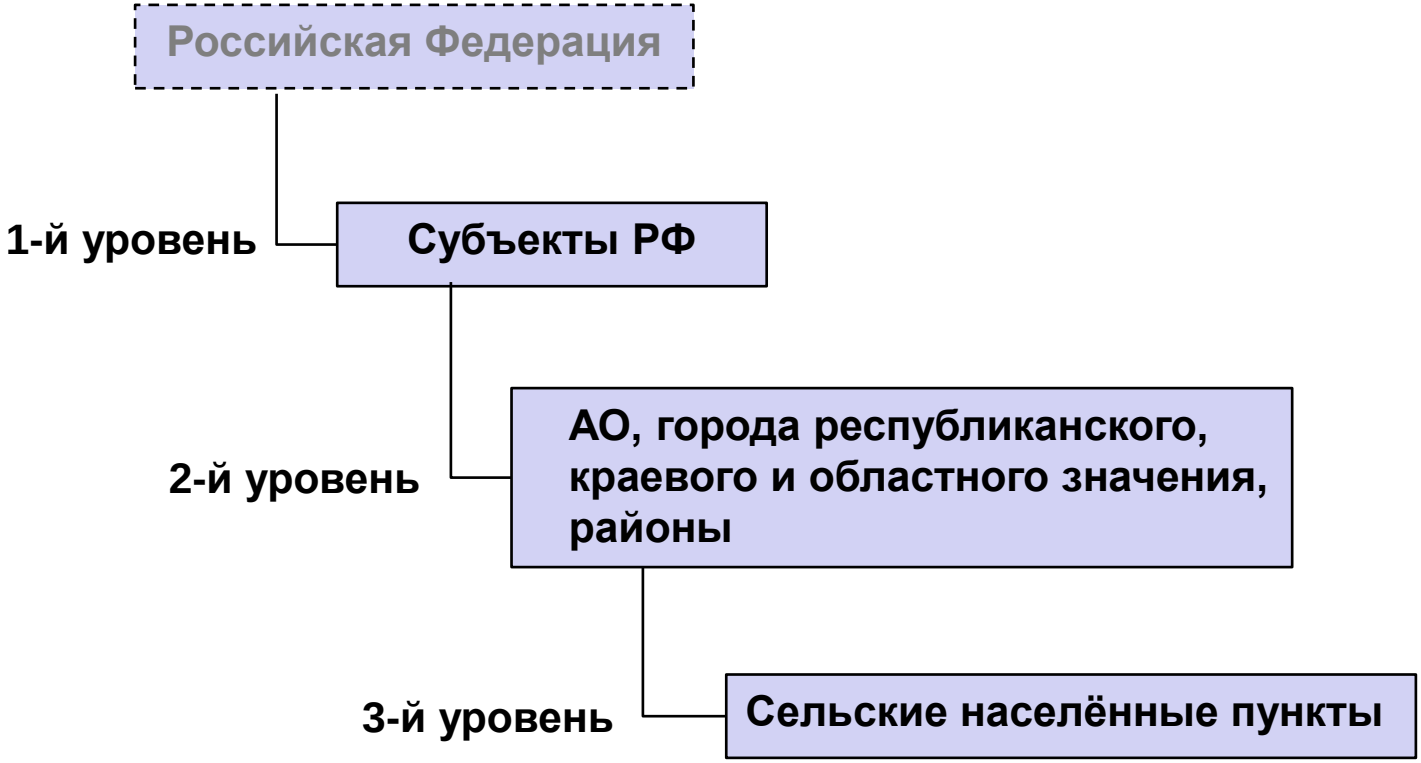
Таксономия

+ 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 ОКАТО



Фасетный метод классификации

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

Образуются непересекающиеся классификационные группировки по различным аспектам или их совокупностям.

Наборы признаков классификации образуют параллельные независимые *фасеты*.

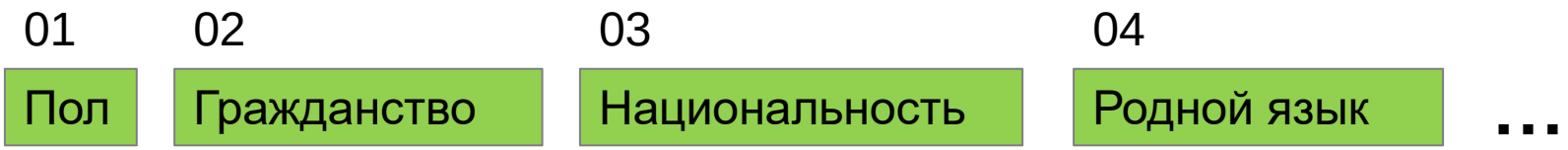
Значения признаков фасетов располагаются в иерархическом порядке или в виде простого перечисления.

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)



Дескрипторный метод классификации

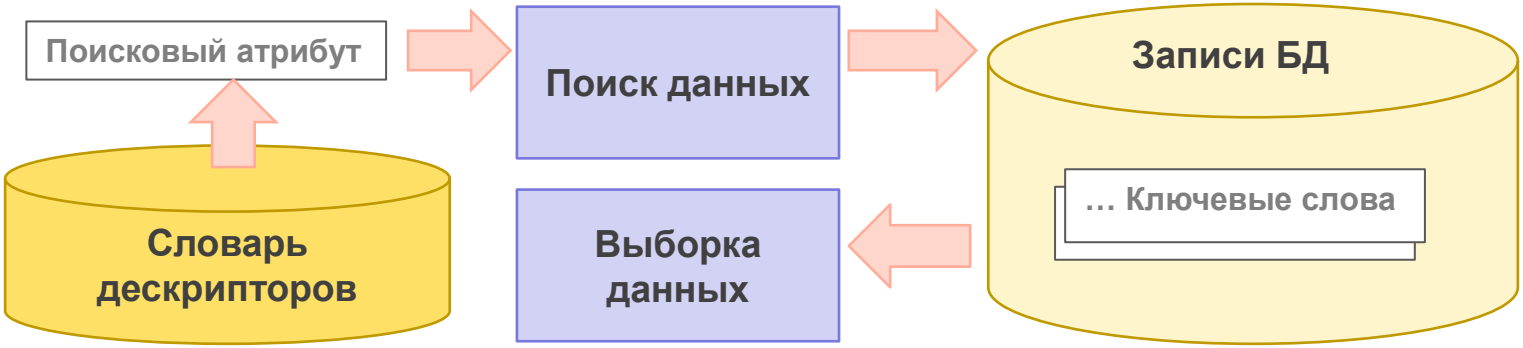
Метод классификации, при котором с экземплярами сущности связывают дескрипторы (описатели) на естественном языке.

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

Особенно широко она используется в библиотечной системе поиска.

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

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

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

Последовательное кодирование

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

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

Алфавитно-цифровые коды переменной длины, тип данных VARCHAR.

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

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 группировки или объекта идентификации (классификации) образуется с использованием независимых группировок, 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

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)

В.И. Артемьев ОКИН: **011 021 102**

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. последующие символы кода зависят от предыдущих • сложно вносить новые признаки классификации • сложно вносить изменения в код без перестройки справочника
Параллельное	Для фасетных классификаций		
Ключевые слова и словосочетания	Для дескрипторной классификации		
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)		

UUID представляет собой 16-байтный (128-битный) номер. В каноническом представлении UUID изображают в виде числа в шестнадцатеричной сис
UUID представляет собой 16-байтный (128-битный) номер. В каноническом представлении UUID изображают в виде числа в шестнадцатеричной с
123e4567-e89b-12d3-a456-426655440000

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.

Широко распространён способ реализации GUID фирмы Microsoft.
UUID представляет собой 16-байтный код, во внешнем представлении это шестнадцатеричное число, разделённое на пять групп 8-4-4-4-12, занимает 36 символов, например: **`123e4567-e89b-12d3-a456-426655440000`**

Альтернатива Nano Id –

использует больший алфавит [A-Za-z0-9_-], занимает 21 символ, например: **`V1StGXR8_Z5jdHi6B-myT`**
Скорость генерации на 60% чем UUID.

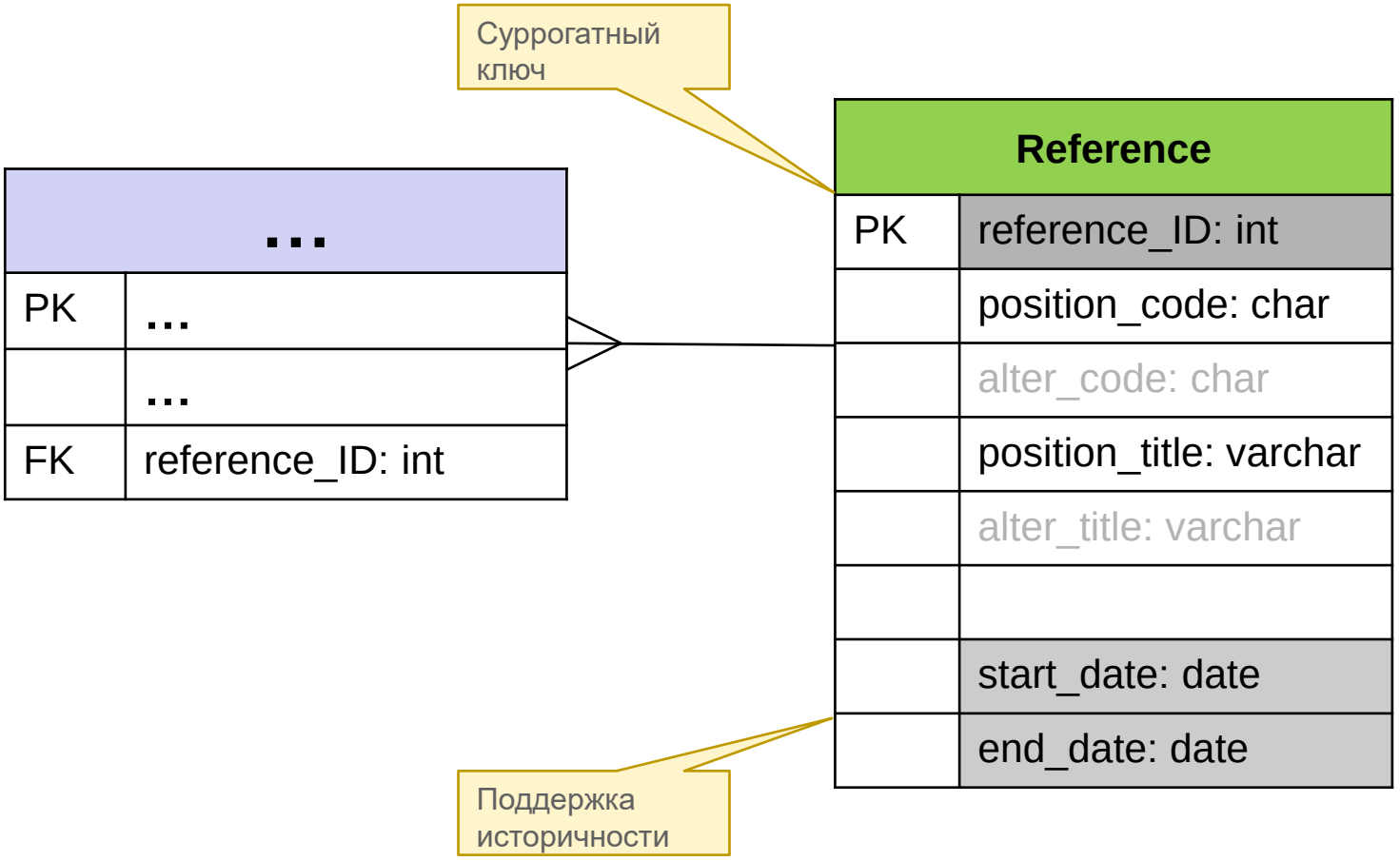
Поддержка истории изменения справочников и реестров

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

Подробнее рассмотрим позже.

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)

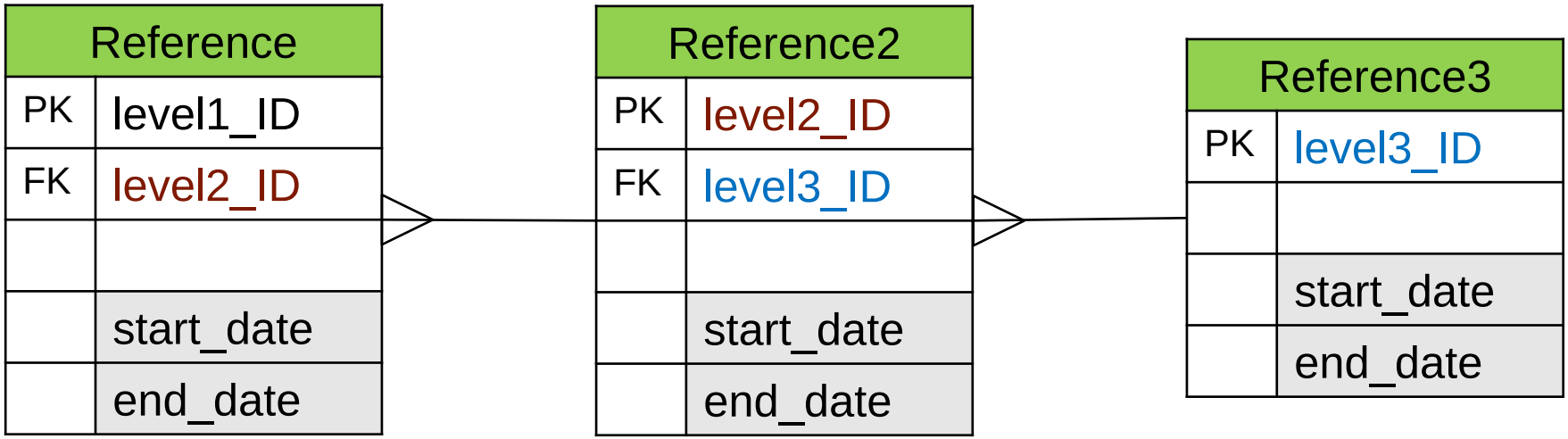
Назначение	- Идентификатор	Тип (размер)
Уникальный номер строки	ID	NUMBER(12)
Цифровой код валюты	CURRENCY	NUMBER(3)
Курс валюты	CURSE	NUMBER(20,6)
Масштаб	SCALE	NUMBER(7)
Дата начала действия курса	EFF_DATE	DATE
Номер документа	ORDR_NUMB	VARCHAR2(8)
Дата издания документа	ORDR_DATE	DATE
Системная дата 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»

Спасибо за внимание!

Валерий Иванович Артемьев

Департамент данных, проектов и процессов
Банк России

Тел.: +7(495) 753-96-25

e-mail: avi@cbr.ru