Методические рекомендации по обеспечению качества данных, предназначенных для межведомственного обмена

Как правило, данные, передающиеся между ведомствами (участвующие в межведомственном обмене) на стороне получателя подвергаются машинной обработке. Получатели таких данных вправе рассчитывать на то, что получаемые данные достоверны, актуальны и документированы и не расходовать время и ресурсы на их дополнительную проверку и очистку. В этой связи к качеству данных, участвующих в межведомственном обмене предъявляются повышенные требования.

> Например, при необходимости поиска по названию организации, пользователь вправе рассчитывать, что название содержит буквы одного алфавита, а не представляет собой смесь кириллических и латинских символов.

Более того, если данные являются эталонными, пользователь эталонных данных не вправе вносить в какие-либо изменения даже в случае обнаружения ошибок, и вынужден направлять запрос на коррекцию в адрес обладателя эталонных данных.

Обращаем внимание, что ошибки в данных могут не нарушать требования нормативных актов, но, тем не менее, существенно усложнять и удорожать обработку таких данных на стороне пользователя.

> Например, требование НПА может указывать, что в данном атрибуте должна содержаться дата документа, но не устанавливать, что какой тип данных следует использовать для хранения даты в информационной системе.

В настоящих методических рекомендациях приводятся наиболее распространенные ошибки в данных, а также рекомендации по их раннему выявлению и недопущению.

1. Типы данных

Тип данных характеризует одновременно множество допустимых значений, которые могут принимать данные, принадлежащие к этому типу и набор операций, которые можно осуществлять над данными, принадлежащими к этому типу. Типизация данных позволяет на ранних этапах не допускать некорректных операций с данными.

> Например, если числа 5 и 100 имеют тип string (строка), то при простой сортировке их по возрастанию 100 будет считаться меньше чем 5.

Крайне важно присваивать данным правильные типы. Хранение числовых данных либо данных дат в текстовых полях способствуют ошибкам и часто приводят к трудноразрешимым проблемам несовместимости.

- ✓ Если в поле содержатся только числа – тип поля должен быть числовым: byte, short, integer, long, big_decimal для перечислимых значений либо с плавающей точкой: float, double для действительных чисел.

> Коды и идентификаторы, в том случае, если установлено, что они могут состоять только из чисел (например, ИНН, СНИЛС) необходимо хранить как числовые поля.

- ✓ Если в поле содержатся только значения даты или даты и времени – тип поля должен быть date или timestamp соответственно.

2. Атомарность данных

Объединение в одном данных различной природы в разы усложняет любые операции с такими данными.

> Например, хранение в одном поле всех реквизитов адреса объекта не позволяет надежно определить какие из слов в строке относятся к населенному пункту, какие - к улице и т.п., соответственно, не позволяют ни обеспечить фильтрацию, ни сортировку, ни контроли заполнения поля на соответствие справочникам.

Информацию о различных атрибутах объектов следует размещать в различных полях. Объединение в одном поле различных атрибутов допустимо только если такое содержание поля явно требуется нормативными актами.

3. Текстовые поля

По содержимому текстовых полей пользователи осуществляют поиск;

Содержимое текстовых полей может подвергаться дополнительной машинной обработке: анализироваться, размечаться, сопоставляться с другими полями, переводиться на другие языки.

В этой связи для текстовых полей рекомендованы следующие проверки:

- ✓ Кодировку: текст должен соответствовать требованиям Unicode и быть представлен в кодировке UTF-8.
- ✓ Наличие непечатных символов: текст должен содержать только печатные символы и символы разметки: знаки пробела, табуляции, конца строки и конца абзаца; при допустимости в тексте иных символов (например, лигатур) – это должно быть явно указано в метаданных соответствующего поля.
- ✓ Язык текста: текст должен быть на русском языке; если в тексте допускается использование других языков – это должно быть явно указано в метаданных поля.
- ✓ Поддержку основного алфавита: текст должен содержать только символы основных алфавитов допустимых языков, а также цифр и знаков

пунктуации; если в тексте допускается использование неалфавитных символов (например, диакритических знаков) – это должно быть явно указано в метаданных атрибута.

- ✓ Отсутствие знаков разметки: текст в атрибуте должен быть представлен в виде простого текста без использования специальных знаков разметки; если текст содержит символы разметки (например, xml) – это должно быть явно указано в метаданных поля.
- ✓ Уместность знаков препинания: в тексте должно быть исключено неуместное либо избыточное использование знаков препинания и иных допустимых символов разметки (например, удвоение пробелов, знаки переноса, подчеркивания между словами и т.п.)
- ✓ Грамотность текста: там, где это применимо, текст должен удовлетворять грамматическим требованиям языка.
- ✓ Наличие аббревиатур и сокращений: там, где это применимо, любые несловарные аббревиатуры и сокращения, в тексте должны быть пояснены в комментарии к полю.
- ✓ Использование специальных шрифтов: в случае, если текст должен отображаться специальным шрифтом (например, моноширинным) – это должно быть явно указано в метаданных поля.

4. Текстовые идентифицирующие поля

Примерами текстовых идентифицирующих полей могут быть фамилии, имена, отчества людей, наименования организаций, объектов, товаров, объектов адресации.

Помимо операций, применимых для любого текста: поиска, машинной обработки, значения текстовых идентифицирующих полей используются для сортировки и фильтрации данных, их агрегации, дедупликации и построения выборок.

Текстовые идентифицирующие поля должны соответствовать всем требованиям к текстовым полям, а также дополнительным требованиям:

- ✓ Поле должно содержать только символы русского алфавита, символы основного латинского алфавита, цифры, а также знаки препинания и пробела; при допустимости в поле иных символов (например, знака подчеркивания) – это должно быть явно указано в метаданных соответствующего поля.
- ✓ Первый символ поля не может быть знаком пробела либо знаком препинания.
- ✓ Не допускается смешение в одном слове русских и латинских символов, если такая возможность явно не указана в метаданных соответствующего поля.

- ✓ Не допускается смешение в одном слове букв и цифр, размещение знаков препинания не последним символом слова, если такая возможность явно не указана в метаданных соответствующего поля.

> Такая проверка рекомендована для исключения возможности замены буквы схожей по написанию цифрой, например, «О» и «0», «З» и «3».
> В случае, если возможность использования в одном слове русских и английских букв, цифр, знаков препинания допускается в редких случаях - рекомендуется помечать такие записи подозрительными и направлять на решение человеку.

- ✓ Не допускается дополнение идентифицирующих полей посторонней информацией, служебными пометками (например, «бывш. …» и т.п.

5. Специальные идентифицирующие поля

К ряду полей, содержащих специальные сведения, применяются дополнительные требования:

5.1. Фамилия, имя, отчество

Поля, содержащие фамилию, имя или отчество гражданина России должны соответствовать всем требованиям к текстовым идентифицирующим полям, а также дополнительным требованиям:

- ✓ Допускаются только символы русского алфавита, дефис и знак пробела.
- ✓ Требуется выделение отдельных полей для фамилии, имени и отчества.
- ✓ Необходимо обеспечить проверки на недопустимость размещения информации в несоответствующих полях (имени в фамилии и т.п.),

> Поскольку полностью автоматически такие проверки обеспечить сложно, рекомендуется отмечать подозрительные записи и передавать на решение человеку.
> Примерами таких подозрительных записей могут быть:
> - окончания имени «ов/ова», «ев/ева», «ин/ина», «ич», «нко»;
> - окончание отчества, отличное от «ич/вна/чна»;
> - фамилии, совпадающие с распространенными именами

- ✓ Рекомендуется обеспечить проверки на соответствие имени/фамилии/отчества полу гражданина.

5.2. Дата

Поля, содержащие дату, должны, как указывалось выше, иметь тип «дата», в этом случае все проверки на форматную допустимость хранимой даты осуществляются автоматически. Дополнительно рекомендуется проверять:

- ✓ Диапазон даты: дата должна попадать в заданный диапазон.

Например, если дата соответствует свершившемуся факту - она не может быть позднее сегодняшнего дня; если дата является частью реквизитов документа - она не может быть ранее даты ввода такого документа в оборот.
Ряд проверок, например, дата рождения гражданина, могут отмечать подозрительные записи (например, даты ранее XX века) для передачи на решение человеку.

- ✓ Последовательность дат: если в наборе данных содержатся даты хронологически последовательных событий, должна проводиться проверка такой же хронологической последовательности соответствующих дат.

Такие же проверки диапазонов и хронологии необходимо осуществлять и для полей «дата-время» (timestamp).

5.1. Числа

- ✓ Числа рекомендуется проверяться на допустимый диапазон.

Например, должна проверяться уместность указания нулевых либо отрицательных значений в числовом поле.
Особое внимание необходимо уделять полям, имеющим нестандартную размерность (например, в тысячах) - для таких полей должны быть реализованы проверки на максимальное значение.

5.2. Коды

Если правила либо ограничения значения поля заданы нормативно, необходимо обеспечить проверку таких правил и/или ограничений, например:

- ✓ Проверка длины и контрольного числа для ИНН, СНИЛС, ОГРН.
- ✓ Проверка длины, допустимых символов и контрольного знака для идентификационного номера транспортного средства.

6. Обязательные поля

Поля, обязательные к заполнению, должны проверяться на наличие значимого значения.

Например, текстовое поле должно содержать как минимум одну букву

7. Использование справочников и классификаторов

Во всех случаях, когда используются внешние (с т.ч. установленные нормативно) либо внутренние справочники, регистры, классификаторы – такие справочники следует выделить явно и обеспечить дополнительные проверки соответствия их структуры и значений эталонным.

> Например, коды стран мира должны соответствовать Общероссийскому классификатору стран мира (ОКСМ)

Также рекомендуется поддерживать внутренние классификаторы в случаях отсутствия нормативно установленного эталона, например, справочник сокращений организационно-правовых форм

8. Дедупликация данных

Одной из причин, приводящих к наиболее серьезным проблемам с качеством данных, является дублирование данных – в этом случае любые изменения данных, включая устранения ошибок, могут не учитываться процедурами и операциями, использующими другие копии данных.

Устранение дублирующих записей (дедупликация) является сложным многоэтапным процессом. Ниже приводятся рекомендации по осуществлению этапов такого процесса.

8.1. Идентификация данных

Для каждого объекта должны быть определены поля, обязательные к заполнению, рекомендованные к заполнению, а также комбинация полей, обеспечивающая уникальность сведений.

> Например, для сведений о гражданине России обязательными являются поля «Имя» и «Фамилия», рекомендованным к заполнению – поле «СНИЛС»;
> Уникальной комбинацией может являться набор «Имя»+«Фамилия»+«СНИЛС».

Наличие уникальной комбинации полей часто не запрещает ввод дублирующих записей, но свидетельствует о проблемах качества данных для таких записей, например, отсутствие значений в полях, рекомендованных к заполнению, либо наличия дублирующих записей.

8.2. Определение золотой записи

Должны быть разработаны и документированы правила определения «золотой записи» – записи, содержащей эталонную информацию об объекте.

> Зачастую, если данные получаются из внешнего «эталонного» источника, создается новая золотая запись, соответствующая эталонному источнику.

8.3. Очистка данных

Самым трудозатратным этапом дедупликации является очистка данных, на котором собственно дедупликация и происходит.

Как правило, очистка данных происходит в полуавтоматическом режиме и состоит из следующих этапов:

- ✓ автоматическое определение неполных и дублирующих записей, связывание таких записей с «золотыми»;
- ✓ ручной анализ каждого дублирующей либо неполной записи и принятие решения о возможности пополнения золотой записи информацией из дублирующей;
- ✓ перенос необходимой информации в золотые записи;
- ✓ пометка «к удалению» всех обработанных дублирующих и неполных записей;
- ✓ автоматическое определение всех зависимостей, ссылающихся на записи, помеченные к удалению, перенаправление таких зависимостей на золотую запись;
- ✓ исключение всех записей, помеченных «к удалению» (определение таких записей не действующими);
- ✓ проверка результата и принятие решения о возможности физического удаления таких исключенных записей.

8.4. Недопущение дупликации

Как правило, в процессе дедупликации данных также реализуются механизмы управления основными и ссылочными данными, минимизирующие возникновение дублей в будущем, и, следовательно, трудозатраты на их устранение.

Часто такие механизмы включают в себя:

- ✓ автоматическое обновление данных из эталонных источников;
- ✓ автоматическое приведение имеющихся данных в соответствие с эталонными;
- ✓ регулярные проверки количества дублирующих и неполных записей.