# Методика оценки качества данных реестра операторов персональных данных

С.Е. Духовенский, П.Ю. Пушкин, Е.В. Никульчев

***Аннотация* — Рассмотрена проблема наличия пропусков и неточностей в данных реестра операторов, осуществляющих обработку персональных данных, опубликованном на сайте Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций. Сформирован комплекс инструментов и разработана методика оценки качества данных реестра. Предложенная методика реализована в информационно-аналитической системе, включающей модуль получения данных, репозиторий метаданных и модуль выполнения проверок.**

**Проведены экспериментальные исследования разработанной системы на выборке, включающей данные по 1671 действующему на момент исследования оператору. Проведенная оценка качества данных показала наличие неточностей в данных 12% рассмотренных операторов, включая неполноту данных, наличие признаков несогласованности в данных, присутствие дублей и выход отдельных значений за пределы допустимого диапазона. Информация по обнаруженным ошибкам может быть использована для улучшения процесса взаимодействия сообщества операторов персональных данных и Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций на основе предложенных в статье рекомендаций, в том числе в формате доработки электронной формы и самопроверки операторов во время ее заполнения.**

***Ключевые слова* — информационно-аналитические системы, качество данных, реестр операторов персональных данных, репозиторий метаданных.**

## I. Введение

В соответствии с федеральным законом [1] операторы персональных данных (далее – Операторы) обязаны с помощью специального уведомления передавать Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций (далее – Роскомнадзор) информацию для ведения реестра операторов персональных данных [2] (далее – Реестр), включая операции добавления, изменения или исключения записей Реестра.

В работах [3]-[5] авторы обратили внимание на наличие определенных неточностей при заполнении Операторами уведомлений на основе общедоступных полей Реестра. Для анализа указанной проблемы была разработана архитектура информационно-аналитической системы мониторинга [3], реализован опытный стенд и проведен глубокий анализа двух полей Реестра: «перечень действий с персональными данными» и «срок или условие прекращения обработки персональных данных» [4]. На основе анализа были даны рекомендации по заполнению указанных полей университетским операторским сообществом [5]. Стоит отметить, что в исследованиях были приведены выдержки из Отчета о деятельности Роскомнадзора за 2019 год [6], которые подтверждают наличие проблем, связанных с получением некорректных данных от Операторов.

На момент данного исследования, отмеченный выше вопрос не потерял актуальности. Так, в Отчете о выполнении Плана и показателей деятельности Роскомнадзора за 2022 год [7] (далее – Отчет) указано, что в течение года было выявлено 457 нарушений в рамках наблюдения за соблюдением требований законодательства Российской Федерации в области персональных данных (далее – ПДн). При этом как минимум 24% от выявленных нарушений были связаны с ведением Реестра, а именно:

- представление в уполномоченный орган уведомления об обработке ПДн, содержащего неполные и (или) недостоверные сведения – в 61 случае (13%);
- непредставление в уполномоченный орган сведений о прекращении обработки ПДн или об изменении информации, содержащейся в уведомлении об обработке ПДн – в 51 случае (11%).

Также в Отчете отмечено, что за 2022 год было подготовлено 4792 приказов на внесение сведений в Реестр, 3065 приказов на внесение изменений (включая информацию о местонахождении баз данных, об обеспечении безопасности, о контактных данных лица, ответственного за организацию обработки, и о наличии трансграничной передаче данных), а также 31 приказ на устранение ошибок. Всего в течение 2022 года в Реестр были внесены сведения о 427 520 Операторах.

На основе представленного Отчета можно сделать вывод, что в настоящее время наблюдается существенное увеличение количества записей Реестра при одновременной работе регулятора по его актуализации, исправлению и дополнению новыми атрибутами, что в совокупности с большим количеством накопленных в реестре данных (более 930 тыс. записей) естественным

Статья получена 28 декабря 2023.
Духовенский С.Е., РТУ МИРЭА, Москва, Россия (dukhovenskiy.s.e@edu.mirea.ru).
Пушкин П.Ю., к.т.н., доцент, РТУ МИРЭА, Москва, Россия (pushkin@mirea.ru).
Никульчев Е.В., д.т.н., профессор, РТУ МИРЭА, Москва, Россия (nikulchev@mirea.ru).

образом приводит к появлению неточностей и пропусков в данных.

С учетом вышесказанного представляется перспективным расширение системы мониторинга, предложенной в [3], с целью покрытия большего количества атрибутов Реестра. Для решения указанной задачи может быть использована разработанная научным сообществом методология оценки качества данных.

Общие методы оценки качества данных, а также подходов к их реализации были подробно исследованы в работах [8] и [9]. Кроме этого, исследованию качества данных посвящено множество зарубежных источников, включая [10]-[12]. Указанные работы, в том числе, освещают различные показатели качества данных (data quality dimensions), которые применяются для оценки качества данных. В текущем исследовании, в первую очередь, были использованы показатели, описанные в одном из самых актуальных обзоров современных методов оценки качества данных [12].

Целью данной работы является анализ текущего уровня качества данных, размещенных в Реестре, и формирование на основе проведенного анализа рекомендаций по повышению качества данных Реестра.

Для достижения указанной цели исследования в работе решались следующие задачи:

- определение перечня доступных для анализа таблиц и атрибутов Реестра, определение применимых показателей качества данных и формулировка методики его оценки (раздел II),
- разработка информационно-аналитической системы для экспериментального исследования предложенной методики (раздел III),
- анализ результатов эксперимента и формирование рекомендаций (раздел IV).

Научная новизна работы заключается в применении методов оценки качества данных для решения задачи выработки рекомендаций по улучшению процесса взаимодействия операторского сообщества и регулятора. Практическая значимость работы заключается в разработке работоспособной методики оценки качества данных Реестра, формировании расширяемой информационно-аналитической системы для оценки качества данных, а также выработке рекомендаций для операторского сообщества.

## II. Описание данных Реестра и методы оценки их качества

Детальная информация по Операторам представлена в Реестре в формате трех таблиц:

1) Основная таблица, содержащая общие атрибуты
2) Таблица «Список информационных систем и их параметры» (далее – список ИС), где для Операторов перечислены информационные системы обработки ПДн и параметры их функционирования
3) Таблица «Цели обработки ПДн» (далее – цели обработки), где для Операторов перечислены возможные цели обработки с указание параметров обработки

Информация по выбранному Оператору обязательно представлена в основной таблице, а также в одной из двух дополнительных таблиц («Список ИС» или «Цели обработки»), причем для одного Оператора количество информационных систем или целей обработки может быть больше 1.

Таблица I. Перечень атрибутов Реестра (с сокращениями)

| |
|---|
| **Основная таблица**<br>• Регистрационный номер<br>• Дата и основание внесения Оператора в Реестр<br>• Статус<br>• Наименование<br>• ИНН<br>• Адрес<br>• Дата регистрации уведомления<br>• Субъект РФ<br>• Наличие шифровальных средств<br>• Трансграничная передача<br>• Местонахождение БД<br>• Цели обработки данных<br>• Правовое основание обработки данных<br>• Описание мер, предусмотренных ст. 18.1 и 19 Закона<br>• ФИО ответственного за обработку данных<br>• Контакты ответственного<br>• Дата начала обработки данных<br>• Срок или условие прекращения обработки ПДн<br>• Дата и основание внесения записи в Реестр<br>• Дата и основание исключения записи из Реестра |
| **Таблица «Список ИС»**<br>• Номер таблицы<br>• Категории данных<br>• Категории субъектов<br>• Перечень действий<br>• Тип обработки<br>• Трансграничная передача<br>• Местонахождение БД |
| **Таблица «Цели обработки»**<br>• Номер таблицы<br>• Цель обработки данных<br>• Правовое основание обработки данных<br>• Категории данных<br>• Категории субъектов<br>• Перечень действий<br>• Тип обработки |

Следуя [12], была рассмотрена применимость показателей качества данных в контексте решаемой задачи; легко видеть, что не все перечисленные в указанном обзоре показатели качества данных подходят для оценки качества данных Реестра: например, технически трудно определить актуальность данных (timeliness) и их правдоподобность (believability). Поэтому в настоящем исследовании были выбраны следующие показатели каче-

ства данных:

1) Отсутствие дубликатов (unambiguous)
2) Полнота данных по отдельным атрибутам (completeness)
3) Вхождение значений атрибутов в допустимый диапазон (accuracy)
4) Согласованность значений разных атрибутов, включая ссылочную целостность (consistency)

На основе указанных показателей качества данных и с учетом структуры хранения данных Реестра была разработана методика оценки его качества данных, состоящая из следующих этапов:

1) Формирование целевой выборки Операторов.
2) Получение данных из источника по выборке.
3) Подготовка данных: приведение значений к подходящим форматам данных, разложение информации по трем плоским связным таблицам.
4) Проверка записей в трех таблицах на наличие дубликатов по определенным кортежам атрибутов.
5) Проверка ссылочной целостности между таблицами.
6) Последовательная проверка каждого значения определенных атрибутов на полноту данных и вхождение значений в допустимый диапазон.
7) Построчная проверка трех таблиц на предмет выполнения определенных критериев согласованности данных.
8) Анализ выявленных на этапах 4-7 записей Реестра, не подходящих под заданные критерии, корректировка проверок при необходимости.

Итоговые проверки, определенные в рамках настоящей работы для использования на этапах 4-7 методики, перечислены в таблице II (всего 59 уникальных проверок).

Таблица II. Реализованные проверки качества данных Реестра

| **Показатель качества данных** | **Область применения:** основная таблица | **Область применения:** таблица «Список ИС» | **Область применения:** таблица «Цели обработки» |
|---|---|---|---|
| Отсутствие дубликатов | • По «Рег. номер»<br>• По «ИНН»<br>• По кортежу: «Наименование», «Адрес» | • По кортежу: «Рег. номер», «Номер таблицы»<br>• По «Категории субъектов» | • По кортежу: «Рег. номер», «Номер таблицы»<br>• По «Цель обработки данных» |
| Ссылочная целостность | «Рег. номер» Оператора должен обязательно присутствовать либо в таблице «Список ИС», либо в таблице «Цели обработки» | | |
| Полнота данных | Все атрибуты в таблице должны содержать непустые значения (кроме «не заполнено»), за исключением атрибутов:<br>• Статус<br>• Наличие шифровальных средств<br>• Трансграничная передача<br>• Местонахождение БД<br>• Цели обработки данных<br>• Правовое основание обработки данных<br>• Дата и основание исключения записи из Реестра | Все атрибуты в таблице должны содержать непустые значения (кроме «не указано») | Все атрибуты в таблице должны содержать непустые значения |
| Допустимый диапазон | • «ИНН» содержит от 10 до 12 цифр<br>• «Наличие шифровальных средств» содержит значения «Используется» или «Не используется»<br>• «Трансграничная передача» принимает значение «да» или «нет»<br>• «Дата включения в Реестр» не превышает текущую дату и указана не ранее 01.01. 2007 г.<br>• «Дата начала обработки» не превышает текущую дату и указана не ранее 01.01. 1900 | «Трансграничная передача» принимает значение «да» или «нет» | |
| Согласованность | • Если указано «ФИО ответственного», то должны быть указаны его «Контакты»<br>• «Дата регистрации уведомления» не должна превышать «дату включения в Реестр»<br>• «Дата начала обработки» не должна пре- | Если Оператор включен в эту таблицу, то в основной таблице должны быть заполнены атрибуты «Цели обработки» и | Если Оператор включен в эту таблицу, то в основной таблице должны быть заполнены атрибуты |

| | | | |
|---|---|---|---|
| | вышать «дату включения в Реестр» более чем на один год | «Правовое основание обработки» | «Трансграничная передача» и «Местонахождение БД» |

Важно отметить, что подбор подходящих проверок качества данных преимущественно опирается на предъявляемые к этим данным требованиям. Разработка точных и корректных проверок качества данных является нетривиальной задачей и требует вовлечения экспертов по выбранной предметной области. Более того, поскольку со временем возможны изменения как в требованиях к данным, так и в самой структуре данных, разработанные проверки требуют регулярного мониторинга и актуализации.

## III. Архитектура системы оценки качества данных

Для оценки качества данных Реестра на основе предложенных выше проверок была разработана информационно-аналитическая система, состоящая из трех компонентов (рис. 1): модуль получения данных (ETL модуль), репозиторий метаданных (metadata repository), модуль выполнения проверок (execution модуль).

Указанное разделение системы на компоненты позволяет быстро адаптировать решение для применения на данных из любых источников, гибко управлять реализованными проверками качества данных, а также масштабировать систему путем добавления новых показателей оценки качества данных.

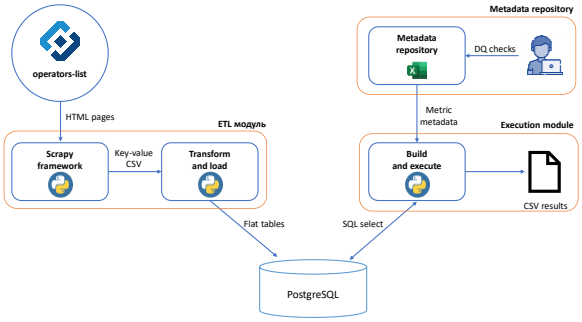


Рис. 1. Архитектура системы

### A. *ETL модуль*

Extract-Transform-Load модуль предназначен для получения данных из системы-источника, преобразования данных в удобный для дальнейшей обработки формат и загрузки в целевую БД.

По состоянию на декабрь 2023 года Реестр доступен только посредством запроса через веб-форму на сайте Роскомнадзора [2]. Представленная форма позволяет задать критерии выборки Операторов (наименование, ИНН, регистрационный номер) и получить перечень уникальных ссылок с детальной информацией по удовлетворяющим критериям выбора Операторам (один Оператор – одна ссылка). Соответственно, для получения полного набора данных по N Операторам необходимо получить HTML код N веб-страниц и извлечь из полученного кода значения целевых атрибутов.

Для реализации данной идеи был использован open source фреймворк Scrapy [13], автоматизирующий запросы к веб-страницам и позволяющий обработать полученный HTML код с целью получения данных. С учетом того, что значения уникальных ссылок для каждого Оператора заранее не известны, а структура веб-страницы с детальной информацией зависит от типа Оператора, веб-скрейпинг (web scraping) выполнялся в три этапа:

1) Получение некоторого перечня доступных Операторов ПДн с помощью выборки по ИНН (например, по первым двум цифрам, которые отвечают за номер Субъекта РФ).
2) Переход по уникальным ссылкам, полученным на шаге 1, извлечение всех значимых значений из HTML кода и выгрузка результатов в CSV в формате «ключ-значение» (key-value).
3) Транспонирование данных из полученных CSV в плоские таблицы и загрузка таблиц в БД PostgreSQL.

### B. *Metadata repository*

Репозиторий метаданных предназначен для добавления или изменения выбранных показателей качества данных с обязательной привязкой к физической структуре таблиц, полученных модулем ETL. На основе формализованных таким образом метаданных можно в автоматическом режиме сформировать машиночитаемые проверки качества данных целевых таблиц, потому структура репозитория метаданных и формат описания показателей качества зависят от реализации алгоритма формирования проверок.

Для целей оценки качества данных Реестра был реализован репозиторий метаданных, состоящий из двух плоских таблиц и позволяющий управлять показателями качества данных в виде SQL выражений.

1) Таблица «Сущности» содержит информацию о проверяемых таблицах: физическая реализация, применяемые фильтры и признак группировки результатов на основе одного из атрибутов.
2) Таблица «Атрибуты» содержит информацию о проверяемых атрибутах: физическая реализация, включенность в уникальный ключ, нижний предел заполняемости, условие вхождения в заданный диапазон, условие согласованности с другими атрибутами.

### C. *Execution module*

Данный модуль позволяет сформировать проверки качества данных в машиночитаемом виде на основе репозитория метаданных, а также выполнить полученные проверки в целевой БД. На выходе модуль формирует таблицу с детальной информацией по выполненным проверкам с обязательным указанием статуса проверки (успешно, некорректная проверка или ошибка в дан-

ных), а также ТОП-10 примеров самых частых ошибочных данных по указанному показателю (в случае нахождения ошибок). Также доступна группировка полученных результатов по одному из атрибутов в целевой таблице.

Для реализации данной идеи был разработан Python-модуль, который последовательно по каждому показателю качества, внесенному в репозиторий метаданных, формирует проверки в виде SQL запроса с типом SELECT и выполняет полученные запросы в БД.

Идея автоматизации формирования проверок качества данных основана на том, что все реализованные типы проверок проходят по одному сценарию: выборка данных из целевой таблицы, фильтрация данных (при необходимости), формирование проверяемых кортежей значений, проверка соответствия кортежей значений показателю качества данных, расчет количества ошибок, получение ТОП-10 наиболее часто встречающихся ошибочных значений.

Тогда для формирования SQL запросов можно использовать следующий шаблон:

```
with step1_step2_filtering as (
      select {column_list} from {table_name}
      where {filter_condition}
   ),
step3_dq_cases as (
      select * from step1_step2_ filtering
      where {dq_case_condition}
      ),
step4_dq_condition as (
      select   cases.*,
         case when {dq_cond} then 1 else 0 end as passed
         from step3_dq_cases as cases
         ),
step5_group_totals as (
      select
         count(*) as total_count,
         count(*) – sum(passed) as failure_count,
         sum(passed)/count(*)::FLOAT as success_rate
      from step4_dq_condition order by 1
   ),
step5_top_failed_cases as (
      select {column_list}, count(*) as failure_count
      from step4_dq_condition
      where dq_condition_passed = 0
      group by {column_list}
      order by failure_count desc
      limit {top_failure_count}
   )
select * from {step5_group_totals | step5_top_failed_cases}
```

В указанном шаблоне внутри фигурных скобок указаны параметры проверок, которые формируются на основе информации из репозитория метаданных.

## IV. Анализ результатов оценки качества данных

В рамках экспериментальных исследований была отобрана случайная выборка, состоящая из 1671 действующего Оператора (признак действия оператора определяется по атрибуту «Статус»).

Применение к выборке предложенной выше методики оценки качества данных Реестра показали наличие ошибок в данных 200 Операторов, что составляет 12% от общего количества Операторов в выборке. При этом записи по 36 Операторам имели две и более ошибок (рис. 2).

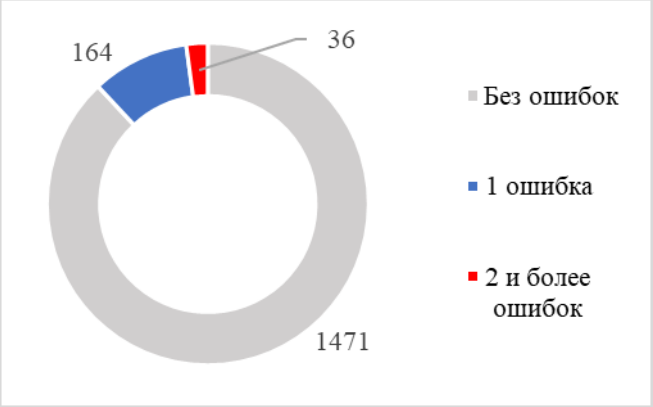


Рис. 2. Количество Операторов в разбивке по числу найденных ошибок

Среди выявленных ошибок чаще всего встречались ошибки полноты данных (51% случаев) и согласованности данных (39%), а также обнаружены дубликаты и несоответствия допустимым диапазонам (9% и 1% соответственно) – см рис. 3. Подобное распределение ошибок в первую очередь обусловлено количеством проводимых проверок каждого типа: детальнее всего рассматривалась полнота данных и согласованность атрибутов внутри одного объекта.


Рис. 3. Количество Операторов в разбивке по показателю качества данных

Углубленный анализ ошибок, связанных с неполнотой данных, показал, что чаще всего были пропущены значения атрибута «Тип обработки ПДн» (50% от всех ошибок данного типа) и информация по ответственному лицу (ФИО или контакты, совокупно 38%).


Рис. 4. Количество ошибок по неполноте данных

Результаты по неполноте данных, выявленной прямыми проверками по указанным атрибутам, можно дополнить результатами проверок на согласованность атрибутов: у 13 Операторов указаны либо только ФИО, либо только контакты, а также была обнаружена 21 запись, где атрибут «Местонахождение БД» не был заполнен ни в основной, ни в дополнительных таблицах. Отметим, что данный результат подтверждает актуальность работ по внесению изменений в Реестр, отмеченных в Отчете.

Наряду с ошибками, связанными с отсутствием значений в определенных атрибутах, стоит отметить ошибки, обусловленные некорректным заполнением значений (таблица III). Например, проверки на несогласованность выявили случаи, когда «Дата начала обработки» превышала «Дату включения в Реестр» более чем на один календарный год. При этом в атрибуте «Дата начала обработки» были найдены несоответствия допустимому диапазону: два Оператора указали несуществующие значения («0009-02-20» и «0028-07-20»). Также стоит обратить внимание на наличие дубликатов: соответствующая проверка определила Операторов, которые указали в таблице «Список ИС» повторяющиеся значения по атрибуту «Категории субъектов».

Таблица III. Примеры ошибок в данных, обусловленные некорректным заполнением значений

| Описание проверки | Кол-во записей | Примеры ошибок |
|---|---|---|
| «Дата начала обработки» указана не ранее 1 января 1900 года | 4 | '0009-02-20', '0028-07-20', '1883-01-01', '1894-01-01' |
| «Дата начала обработки» не должна превышать «дату включения в Реестр» более чем на один год | 69 | Регистрация ‘2008-12-15’, начало обработки ‘2020-11-10’<br>Регистрация ‘2012-05-23’, начало обработки ‘2022-09-01’<br>Регистрация ‘2018-08-27’, начало обработки ‘2023-07-14’ |
| Отсутствие дубликатов в таблице «Список ИС» по атрибуту «Категории субъектов» | 24 | У одного Оператора для атрибута «Субъекты ПДн»<br>3 раза указано значение «работники ООО ‘Транснефть – Порт Приморск’» |

На основе проведенного анализа можно дать следующие рекомендации по повышению качества данных Реестра:

1) В процессе заполнения формы уведомления Оператору необходимо обратить внимание на одновременное указание непустых значений для атрибутов «ФИО» и «Контакты ответственного лица».
2) Также Оператору рекомендуется внимательно проверять корректность заполнения атрибута «Дата начала обработки».
3) В процессе заполнения таблицы «Список ИС» Оператору рекомендуется объединять информацию на основе атрибута «Категории субъектов» с целью исключения дублирования данных.
4) Поскольку на текущий момент в Реестре для отдельных Операторов наблюдаются пропуски значений в атрибутах «Тип обработки» и «Местонахождение БД», можно рекомендовать провести дополнительную работу по актуализации данных сведений.
5) Дополнительно можно рекомендовать добавление в электронную форму уведомления автоматических проверок на заполняемость связанных полей «ФИО» и «Контакты ответственного лица», на нижнюю и верхнюю границы атрибута «Дата начала обработки», а также на наличие дублей по атрибуту «Категории субъектов», что позволит исключить опечатки в процессе заполнения формы Операторами.

## V. Заключение

Проведенные исследования подтвердили успешность применения методологии оценки качества данных для выявления неточностей в Реестре Операторов ПДн и разработке рекомендаций по повышению качества данных Реестра.

Предложенная в статье методика проверки качества данных на случайной выборке из 1671 Оператора позволила «подсветить» наличие ошибок в данных 12% Операторов, связанных в первую очередь с неполнотой данных по отдельным атрибутам, в том числе в формате несогласованной заполненности данных. Также требуют внимания обнаруженные в данных дубликаты и несоответствия некоторых значений допустимым диапазонам.

Информация по обнаруженным ошибкам может быть использована для улучшения процесса взаимодействия сообщества операторов персональных данных и Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций на основе предложенных в статье рекомендаций, включая доработку электронной формы и самопроверку Операторов во время ее заполнения.

Будет справедливым отметить, что полученная в рамках исследования оценка качества данных Реестра напрямую зависит от разработанных для этой цели проверок. Для уточнения полученной оценки и выработки более полных рекомендаций по повышению качества данных представляется целесообразным добавление новых проверок в рамках предложенных показателей качества данных, равно как и расширение перечня рассматриваемых в таблице I показателей. Данная особенность была учтена при разработке применяемой в рамках исследования информационно-аналитической системы, которая позволяет гибко изменять или расширять перечень используемых проверок. Поскольку реализованный образец системы был разработан с помощью свободно распространяемых программных компонентов, он может быть легко интегрирован в общую схему предложенной в [3] системы мониторинга выполнения Операторами требований законодательства.

Наряду с совершенствованием перечня проверок возможным перспективным развитием исследования является проведение оценки качества данных на большей выборке записей Реестра с группировкой по различным разрезам, например по Субъектам РФ. В этом случае предположительно можно будет дать рекомендации по повышению качества данных Реестра с учетом регионального расположения Операторов.

## Библиография


[1] *О персональных данных,* Федеральный закон от 27.07.2006 №152-ФЗ, http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261

[2] *Реестр операторов, осуществляющих обработку персональных данных*, https://pd.rkn.gov.ru/operators-registry/operators-list/

[3] В. П. Лось, Е. В. Никульчев, П. Ю. Пушкин, А. М. Русаков. Информационно-аналитическая система мониторинга выполнения операторами персональных данных требований законодательства // Проблемы информационной безопасности. Компьютерные системы. 2020. № 3. С. 16-23.

[4] П. Ю. Пушкин, А. М. Русаков. Результаты автоматического интеллектуального анализа отдельных полей реестра операторов персональных данных // International Journal of Open Information Technologies. 2021. Т. 9. № 1. С. 37-47.

[5] Е. В. Никульчев, П. Ю. Пушкин, А. М. Русаков. Рекомендации по заполнению университетским операторским сообществом отдельных разделов уведомления об обработке персональных данных // В кн.: Информационная безопасность личности субъектов образовательного процесса в цифровой информационно-образовательной среде: сборник научных статей. – М.: РГУ нефти и газа (НИУ) имени И.М. Губкина, 2021, с. 318-325.

[6] *Отчет о деятельности Уполномоченного органа по защите прав субъектов персональных данных за 2019 год*, https://rkn.gov.ru/docs/Otchet_UO-2019_new.pdf

[7] *Отчет о выполнении Плана и показателей деятельности Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций в 2022 году*, https://rkn.gov.ru/docs/doc_3806.pdf

[8] А. А. Ильин. Автоматизированная технология проектирования модели данных при построении информационно-аналитической системы // Вестник российских университетов. Математика. – 2008. Т. 13. №. 1. С. 89-90.

[9] M. Borisyak, A. Ryzhikov, A. Ustyuzhanin, D. Derkach, F. Ratnikov, O. Mineeva. (1+ ε)-class classification: an anomaly detection method for highly imbalanced or incomplete data sets // The Journal of Machine Learning Research. 2020. Т. 21. №. 1. 2768-2789.

[10] W. Fan. Data quality: From theory to practice // ACM SIGMOD Record. 2015. Т. 44. №. 3. С. 7-18.

[11] P. Oliveira, F. Rodrigues, P. R. Henriques. A formal definition of data quality problems // in *Proceedings of the 2005 International Conference on Information Quality*. – MIT, 2005.

[12] J. Wang, Y. Liu, P. Li, Z. Lin, S. Sindakis, S. Aggarwal. Overview of data quality: examining the dimensions, antecedents, and impacts of data quality // Journal of the Knowledge Economy. 2023 https://doi.org/10.1007/s13132-022-01096-6.

[13] D. S. Sirisuriya. A comparative study on web scraping. 2015. http://ir.kdu.ac.lk/handle/345/1051.

# The Data Quality Assessment Technique of Personal Data Operators Register

S. E. Dukhovenskiy, P. Pushkin, E. Nikulchev

***Abstract* — The work considers the issue of completeness and accuracy of personal data operators register published by Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications. The set of tools was designed and the data quality assessment technique was determined. The suggested technique was implemented via the information and analytical system including ETL module, metadata repository and check execution module.**

**An experimental test of the implemented system covered a random sample of 1671 currently active operators. The resulting data quality assessment highlighted inaccuracies in records of 12% operators, including data incompleteness, data inconsistency, the presence of duplicates and some outliers beyond acceptable range. Revealed failures can be used to improve the electronic document interchange between operator community and Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications taking into account the recommendations, presented in the paper, such as electronic form improvements and operators' self-check as part of document workflow.**

***Key words* — information and analytical systems, data quality, personal data operators register, metadata repository.**

## REFERENCES

[1] *About Personal Data*, Federal law of 27.07.2006 No. 152-FZ, http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261 (in Rus)

[2] *Personal data operators register*, https://pd.rkn.gov.ru/operators-registry/operators-list/ (in Rus)

[3] V.P. Los, E.V., Nikulchev P.Y. Pushkin, A.M. Rusakov, "Information and analytical system for monitoring the compliance of personal data operators with the requirements of the legislation," *Problems of information security. Computer systems*, no. 3, pp. 16-23, 2020. (in Rus)

[4] P.Y. Pushkin, A.M. Rusakov, "Results of automatic mining individual fields of personal data operators register," *International Journal of Open Information Technologies*, vol. 9, no. 1, pp. 37-47, 2021. (in Rus)

[5] E.V. Nikulchev, P.Y. Pushkin, A.M. Rusakov, "Recommendations for certain attribute filling of the personal data processing notifications by the university operator community" in *Information security of the educational process subject personality in the digital information and educational environment: collection of scientific articles*, Moscow: Gubkin Russian State University of Oil and Gas, pp. 318-325. 2021. (in Rus)

[6] *Report on the activities of the Authorized Service for the Protection of the Rights of Personal Data Subjects for 2019*, https://rkn.gov.ru/docs/Otchet_UO-2019_new.pdf (in Rus)

[7] *Report on execution of the Plan and performance indicators by Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications in 2022*, https://rkn.gov.ru/docs/doc_3806.pdf (in Rus)

[8] A. A. Ilyin, "Automated technology for designing a data model when building an information and analytical system ," *Bulletin of Russian Universities. Mathematics*, vol. 13, no. 1, pp. 89-90, 2008. (in Rus)

[9] M. Borisyak, A. Ryzhikov, A. Ustyuzhanin, D. Derkach, F. Ratnikov, O. Mineeva, "$(1+\varepsilon)$-class classification: an anomaly detection method for highly imbalanced or incomplete data sets," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 2768-2789, 2020.

[10] W. Fan, "Data quality: From theory to practice," *ACM SIGMOD Record*, vol. 44, no. 3, p. 7-18, 2015.

[11] P. Oliveira, F. Rodrigues, P. R. Henriques, "A formal definition of data quality problems," in *Proceedings of the 2005 International Conference on Information Quality*, MIT, 2005

[12] J. Wang, Y. Liu, P. Li, Z. Lin, S. Sindakis, S. Aggarwal, "Overview of data quality: examining the dimensions, antecedents, and impacts of data quality," *Journal of the Knowledge Economy*, 2023 https://doi.org/10.1007/s13132-022-01096-6.

[13] D. S. Sirisuriya, "A comparative study on web scraping," 2015. Available: http://ir.kdu.ac.lk/handle/345/1051.

**About Authors**

**S. E. Dukhovenskiy**, graduate student, MIREA – Russian Technological University, Moscow, Russia (dukhovenskiy.s.e@edu.mirea.ru).

**Pavel Pushkin**, Director of the Institute Institute of Advanced Technologies and Industrial Programming, MIREA – Russian Technological University, Moscow, Russia (pushkin@mirea.ru).

**Evgeny Nikulchev**, Professor of Department of Digital Data Processing Technologies, MIREA – Russian Technological University, Moscow, Russia (nikulchev@mirea.ru).