

**Европейская экономическая комиссия****Конференция европейских статистиков****Группа экспертов по переписям населения
и жилищного фонда****Девятнадцатое совещание**

Женева, 4–6 октября 2017 года

Пункт 4 предварительной повестки дня

Оценка переписи и измерение качества данных**Измерение качества в государственных
информационных системах и базах данных****Записка Департамента статистики Эстонии¹***Резюме*

Организация первой регистровой переписи в Эстонии требует относительно большого объема подготовительной работы в целях повышения качества данных, укрепления семантического потенциала и т.д.

Была выдвинута новая инициатива, которая обеспечивает на национальном уровне локализацию институциональных знаний и навыков, необходимых для обеспечения удовлетворяющего определенным функциональным требованиям результата при проведении регистровой переписи в Эстонии в 2021 году. Департамент государственной информационной системы (ДГИС) поручил провести в 2016 году исследование с целью подготовки руководства по измерению качества данных для держателей регистров.

Качество данных в регистровых базах данных непосредственно влияет на функционирование государства. Чем выше качество баз данных, тем более удобными и доступными являются их услуги для широкой общественности и тем более точные решения могут приниматься в статистических целях.

Настоящий документ посвящен описанию вышеупомянутой новой инициативы общенационального уровня.

¹ Авторы: Кристи Лехто и Диана Белтадзе.



Содержание

	<i>Пункты</i>	<i>Стр.</i>
I. Введение	1–6	3
II. Является ли качество регистровых данных только предметом озабоченности для статистиков?	7–13	4
III. Руководство по качеству данных для регистров	14–27	4
IV. Результаты пилотного исследования	28–34	
V. Выводы	35–38	7
Справочные материалы		9
Приложение		
Вопросник для определения степени зрелости системы обеспечения качества данных		10

I. Введение

1. Регистровая перепись населения и жилищного фонда (REGREL) требует наличия полных, регулярно обновляемых и поддающихся проверке государственных регистров, которые охватывают основные признаки переписи.

2. С 2009 года Департамент статистики Эстонии занимается подготовкой к проведению переписи населения и жилищного фонда 2021 года в регистровом формате. Регистровая перепись опирается на благоприятный контекст: данные по почти всем обязательным признакам переписи ЕС (Регламент ЕС № 763/2008) могут быть получены из эстонских регистров; имеется действующая система личных идентификационных кодов, позволяющая идентификацию во всех регистрах персональных данных; и во многие базы данных интегрирована система адресных данных (САД), которая была создан и введена в действие в 2008 году.

3. В 2015 году Департамент статистики Эстонии издал требования для держателей статистических регистров, которым они должны удовлетворять до начала регистровой переписи:

- по меньшей мере, 97-процентный охват населения;
- информация по требуемым признакам переписи должна обновляться на регулярной основе, и этот процесс должен документироваться;
- по меньшей мере, 95% записей должны иметь идентификатор стандартного формата;
- по меньшей мере, 95-процентный охват соответствующих признаков переписи;
- доля значительных или технических ошибок в значениях соответствующих признаков переписи не может превышать 1%.

4. Меры, необходимые для обеспечения соответствия всех регистров установленным требованиям, включают в себя: повышение и контроль качества данных; замена других источников адресных данных административной системой САД; предоставление данных с классификациями. Гармонизация классификаций и терминов в государственных регистрах является одним из ключевых направлений деятельности в рамках подготовки регистровой переписи.

5. Согласно статье 50 (2) Закона об официальной статистике, Департамент статистики Эстонии осуществляет оценку качества данных в базах данных. Требования к обеспечению качества данных в рамках переписи являются следующими:

- регистры должны учитывать результаты регулярной оценки качества, проводимой Департаментом статистики Эстонии;
- данные, получаемые из регистров, должны соответствовать признакам, согласованным с Департаментом статистики Эстонии, быть полными и точными;
- регистры должны проверять внутреннюю непротиворечивость своих данных (например, выявлять и устранять дублирующие записи и ошибки, которые могут выявляться путем комбинирования различных признаков и т. д.).

6. Основное внимание в рамках подготовки регистровой переписи уделяется с 2015 года содержанию и качеству данных:

- в случае одной пятой части населения зарегистрированное место жительства отличается от фактического места жительства;

- качество данных в государственном регистре строительных работ является недостаточным для статистических целей. Низкое качество регистра строительных работ обусловлено недостаточным охватом зданий и жилищ, неполнотой данных о технических характеристиках, а также отсутствием обновлений.

II. Является ли качество регистровых данных только предметом озабоченности для статистиков?

7. Проведение REGREL основывается на положениях статьи 50 Закона об официальной статистике (ЗОФ). Согласно им, Департамент статистики Эстонии осуществляет оценку качества базовых данных баз данных и направляет предложения ответственному обработчику базы данных по повышению качества данных. В этом контексте не имеется неразрешимых правовых проблем, если только мы не будем учитывать тот факт, что разработчику официальной статистики поручено осуществлять надзор за качеством данных в базах данных для статистических целей, однако сфера охвата этой задачи не определена. Неясно, кто должен нести ответственность за разработку стандартов качества для государственных регистров.

8. Также возникают вопросы в отношении формата отзывов по качеству данных, которые должны направляться регистрам. В настоящее время не существует никаких действующих стандартов, регулирующих эти вопросы.

9. В качестве примера мы можем взять базу данных о жилищах. Для повышения качества данных регистратор Эстонского регистра зданий должен провести работу, которая требует четкого правового регулирования. Так, например, если собственник здания представил данные в регистр на основе юридического документа начала прошлого века, а нынешний собственник проживает в другой стране, собственник не обязан проверять и обновлять данные о своей собственности в Эстонском регистре зданий. В результате этого база данных содержит ряд элементов данных, которые не соответствуют нынешним требованиям к качеству данных, что в свою очередь создает проблемы для разработки статистики, даже если данные обладают законной силой.

10. Другая сторона медали заключается в том, что базы данных не располагают достаточными ресурсами и необходимые работы по повышению качества данных не интегрированы в их повседневную деятельность.

11. На совещании Рабочей группы по базам данных для REGREL 2016 года многие регистраторы поднимали вопрос о том, как повысить качество данных и какие другие выгоды это принесет, помимо разработки статистики. Универсальный ответ Департамента статистики Эстонии заключался в том, что повышение качества данных содействует принятию более эффективных решений.

12. Высококачественные решения могут опираться только на проверенные базы данных высокого качества. Проверка качества данных не будет продуктивной, если она останется разовым мероприятием. Контроль качества данных должен быть всеобъемлющим и постоянным процессом. Он требует планирования, оценки, анализа, тестирования и, в случае необходимости, (правовых) мер по улучшению данных.

13. Для улучшения общей ситуации в стране была создана универсальная модель управления качеством данных.

III. Руководство по качеству данных для регистров

14. Департамент государственной информационной системы (ДГИС) постановил провести в 2016 году исследование с целью подготовки руководства по

измерению качества данных для держателей регистров. Данное исследование было проведено фирмой AS PricewaterhouseCoopers Advisors.

15. Цель ДГИС заключалась в разработке всеобъемлющей, простой и применимой методологии оценки текущего положения, позволяющей определить оптимальный требуемый уровень качества, планировать изменения на индивидуальном уровне и установить процедуры перехода с одного уровня на другой. Работа по подготовке руководства по качеству данных привела к спецификации методов, приемов, организационных структур, ролей и ключевых функций в области использования баз данных и управления ими, которые облегчают техническое и семантическое усовершенствование качества регистровых данных.

16. Разработанное руководство по качеству данных для государственных регистров содержит методологии оценки и обеспечения качества данных в государственной информационной системе в целом. В руководстве описывается методология управления контролем и надзором за качеством базы данных и даются рекомендации в отношении параметров для использования в целях контроля качества данных.

17. Разработанная рамочная основа управления качеством данных включает в себя три элемента:

- модель зрелости качества данных, разработанная для оценки и совершенствования категорий, связанных с управлением качеством данных. Разработанная модель охватывает пять категорий управления на пяти уровнях зрелости. Эта модель также включает в себя описание методологии реализации;
- набор показателей качества данных, которые могут использоваться для тестирования различных аспектов качества данных. Данный набор состоит из девяти показателей, каждый из которых содержит требования к качеству данных, меры по обеспечению качества и тестовые вопросы;
- рамочная основа управления качеством данных, которая представляет собой набор итерационно реализуемых мер по обеспечению качества данных на основе модели управления качеством OPDCA.

18. Эта рамочная основа объединяет перечисленные элементы в единую систему управления качеством.

19. Цель изучения показателей качества данных заключалась в определении универсального набора показателей качества данных, которые могли бы применяться всеми базами данных государственной информационной системы для измерения и обеспечения качества.

20. Источники показателей были определены с помощью веб-исследования с целью нахождения в государственной информационной среде описаний гипотетических или реализованных рамочных основ обеспечения качества данных, которые содержат определенные показатели для оценки качества данных.

21. Результатом этого исследования стал анализ около 20 описаний рамочных основ качества данных, включая стандарты качества данных, практические методы обеспечения качества в международных банках и компаниях, практические методы управления качеством данных в государственных органах различных стран, статьи, опубликованные экспертами-практиками, а также посты в блогах широко признанных онлайн-исследований и публикаций по тематике ИТ.

22. Анализ показателей позволил определить набор из девяти показателей, которые были признаны пригодными для оценки качества данных в базах данных государственной информационной системы. Описание отобранных показателей приводится в таблице 1.

Таблица 1
Набор показателей качества данных

Точность	Под точностью данных понимается точное представление истинного значения термина или события атрибутами данных в конкретном контексте использования.
Полнота	Полнота данных означает наличие введенных значений данных по всем атрибутам данных всех объектов, по которым требуются значения.
Непротиворечивость	Под непротиворечивостью данных понимается отсутствие расхождений между данными и непротиворечивость с другими данными.
Достоверность	Под достоверностью данных понимаются атрибуты данных, которые могут считаться пользователями истинными и надежными в конкретном контексте.
Своевременность	Под своевременностью данных понимается период времени между записью данных и основополагающим событием.
Конфиденциальность	Конфиденциальность данных означает предоставление доступа к данным только авторизованным пользователям (лицам или техническим системам) и недоступность их для всех остальных.
Единичность	Под единичностью данных понимается соблюдение применимых принципов единичного сбора данных от поставщиков.
Неизбыточность	Под избыточностью понимается планирование и разработка мер, направленных на предотвращение избыточности в структурах данных базы данных и мер по недопущению избыточности в обработке данных.
Нормативность	Под нормативностью данных понимается наличие атрибутов данных, которые соответствуют стандартам качества данных, правилам, применимому законодательству или аналогичным нормам в конкретном контексте использования.

23. Цель исследования по управлению качеством данных заключалась в разработке методологии измерения качества данных для баз данных государственной информационной системы. Задача исследования состояла в анализе практических методов, применяемых организациями для измерения процессов обеспечения качества данных, и в разработке методологии для измерения качества данных в базах данных государственной информационной системы.

24. Эти аспекты качества были сгруппированы по пяти основным категориям обеспечения качества данных:

- управление и планирование. К этой категории относятся аспекты деятельности и инструменты комплексного управления и планирования качества данных, включая осуществление правил, политики, норм и планов и стратегического подхода к обеспечению качества данных, принятого организацией;

- организация и обязанности. Эта категория охватывает аспекты, связанные с созданием организации, отвечающей за управление качеством данных, и распределением обязанностей;
- процессы. Эта категория касается осуществления планов в области качества данных, включая контроль, измерение и обеспечение качества;
- знания и компетенции. Эта категория включает в себя аспекты, относящиеся к управлению знаниями по вопросам качества данных (включая распространение знаний и приобретение компетенций);
- технические инструменты. Эта категория охватывает аспекты, связанные с внедрением и совершенствованием инструментов ИТ в поддержку процессов обеспечения качества данных.

25. Следующий шаг заключался в определении надлежащей методики для оценки положения с качеством данных в каждой отдельной категории. Согласно выдвинутой гипотезе, в качестве такого инструмента могли бы использоваться узконаправленные модели зрелости, поскольку модель зрелости является инструментом, широко используемым для оценки людей/культур, процессов/структур и объектов/технологий. Модель зрелости является инструментом, который облегчает шкальную оценку текущего положения в той или иной конкретной области и установление поэтапных целей развития.

26. Отобранная модель зрелости опиралась на следующие пять уровней:

- a) реактивный уровень. Процессы обеспечения качества данных непредсказуемы, слабо контролируются, и действия, касающиеся процессов, носят реактивный характер;
- b) контролируемый уровень. Важность управления качеством данных признана, и внедрены повторяемые процедуры;
- c) стандартизованный уровень. Процессы обеспечения качества данных стандартизованы, и качество данных проверяется на соответствие;
- d) управляемый уровень. Процессы стали устойчивыми. Результаты измерения качества данных используются для управления существующими процессами;
- e) оптимизированный уровень. Процессы обеспечения качества данных регулярно пересматриваются и обновляются.

27. Данное исследование привело к разработке модели зрелости качества данных для оценки зрелости каждой категории качества данных на основе пяти уровней зрелости. Эта модель позволяет пользователям измерять зрелость управления качеством данных по 5-балльной шкале (см. приложение).

IV. Результаты пилотного исследования

28. После завершения подготовки руководства по качеству данных для держателей регистров было проведено исследование по изучению качества данных баз данных для проверки применимости данного руководства и его усовершенствования с учетом отзывов по результатам его практического применения.

29. Исследование по изучению применимости руководства по качеству данных, заказанное ДГИС, проводилось в трех государственных регистрах: регистре населения (РН), системе адресных данных (САД) и системе администрирования государственной информационной системы (САСИС).

30. Зрелость системы обеспечения качества данных оценивалась с использованием инструмента оценки зрелости (см. приложение). Оценка производилась путем сопоставления описаний уровней зрелости в разных категориях с фактическим положением.

31. Оценка проводилась в два раунда. В ходе первого раунда некоторые описания в модели зрелости оказались непонятными для участников, после чего в инструмент оценки были внесены изменения для улучшения понимания, а затем был проведен второй раунд оценки. Что касается регистра населения, то первый и второй раунды дали различные оценки уровня зрелости.

32. По каждому регистру оценка уровней зрелости предусматривала определение:

- уровней зрелости в разбивке по категориям, иллюстрирующих зрелость системы управления качеством данных в каждой категории, включенной в модель зрелости (шкала от 1 до 5);
- суммарный рейтинг по зрелости, который можно использовать для подробного анализа эффективности мер по повышению качества данных.

Таблица 2

Уровни зрелости системы обеспечения качества данных

<i>Категория зрелости</i>	<i>САСИС</i>	<i>РН (1)</i>	<i>РН (2)</i>	<i>САД</i>
1. Управление и планирование	1	1	3	5
2. Организация и обязанности	1	3	3	5
3. Процессы	1	1	4	5
4. Знания и компетенции	1	2	5	5
5. Технические инструменты	1	5	5	5
Суммарный рейтинг по зрелости	1,0	2,4	4,0	5,0

33. Были сформулированы следующие общие рекомендации в отношении применения руководства по качеству данных и оценки качества:

- В случае баз данных государственной информационной системы индивидуальные базы данных должны постоянно следить за качеством данных с охватом всех данных;
- при зонировании организации следует изучать возможность включения всех заинтересованных сторон, которые влияют на качество данных (включая авторизованных обработчиков баз данных и специалистов по информационным технологиям), в зону управления качеством данных;
- при оценке применимости описаний уровней зрелости следует немедленно указывать по каждому описанию причины вынесения соответствующей оценки. Это служит важным шагом в направлении составления плана усовершенствований. Это также помогает сократить время, требуемое для разработки плана, и предотвратить последующие расхождения с первоначальными оценками;
- следует привлекать специалистов по информационным технологиям организации к оценке для предоставления компетентной оценки состояния технических инструментов.

34. Среди прочего, для держателей регистровых данных были определены требования к Эстонскому национальному каталогу метаданных: он должен быть всеобъемлющим, обязательным и служить базовым регистром для ряда процессов. Новая концепция предусматривает: распределенную разработку и публикацию метаданных; публикацию в машиночитаемом формате; загрузку в центральное хранилище; открытие внутренних хранилищ метаданных учреждений; более реалистичный подход к требованиям к разработке метаданных.

V. Выводы

35. Цель данного исследования состояла в разработке методологии для измерения качества данных и подготовке структурированного руководства по ее реализации. ДГИС также провел пилотное исследование и оценку качества данных регистра населения (РН), системы адресных данных (САД) и системы администрирования государственной информационной системы (САСИС).

36. Ключевые выводы по использованию руководства по качеству данных касаются:

- признания важности качества данных администрацией;
- создания постоянно действующей устойчивой системы;
- четкого распределения обязанностей и обеспечения охвата.

37. Качество исходных данных имеет весьма важное значение для всех стран, проводящих регистровую перепись. Крайне важно, чтобы качество данных являлось одним из приоритетов для держателей регистров.

38. Инициатива «Качество данных», направленная на повышение качества данных в регистрах, опирается на хорошо известные модели качества данных (параметры качества) и рамочные основы совершенствования процессов (5-уровневая модель зрелости). В ее рамках был разработан руководящий документ по качеству данных и проведено в 2016 году обследование качества данных в трех эстонских регистрах.

Справочные материалы

Руководство по качеству данных (2016)

<https://www.ria.ee/ee/andmekvaliteedi-tagamine.html> (только на эстонском языке).

Регламент ЕС № 763/2008

<http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32008R0763>.

Закон об официальной статистике

<https://www.riigiteataja.ee/en/eli/ee/506012015002/consolide>.

Приложение

Вопросник для определения уровня зрелости системы обеспечения качества данных

УПРАВЛЕНИЕ И ПЛАНИРОВАНИЕ КАЧЕСТВА ДАННЫХ

- 1 Требования к качеству данных для базы данных определены и задокументированы (1, 2).
- 2 Качество данных подвергается проверке на соответствие требованиям (3).
- 3 Все важнейшие данные, являющиеся объектом политики, определены и задокументированы (2).
- 4 Политика в области качества данных задокументирована и обнародована (3).
- 5 Система управления качеством данных реализована единообразно в соответствующей зоне (4).
- 6 Качество данных измеряется и совершенствуется на основе анализа результатов измерений (4).
- 7 Показатели качества данных увязаны со стратегиями или политикой более высокого уровня (4).
- 8 Процессы обеспечения качества данных подлежат постоянному совершенствованию (5).
- 9 Управление качеством данных осуществляется путем регулярного обзора, пересмотра и обнародования политики (5).

ОБЯЗАННОСТИ ПО УПРАВЛЕНИЮ КАЧЕСТВОМ ДАННЫХ В ОРГАНИЗАЦИИ

- 10 Процесс управления качеством данных имеет идентифицируемого владельца (1, 2).
- 11 Обязанности владельца процессов обеспечения качества данных четко определены в описании должностных функций (3).
- 12 Специфицированы владельцы данных (1, 2, 3).
- 13 Создана группа старших руководителей, отвечающих за обеспечение качества данных (4).
- 14 В состав этой группы входят представители других заинтересованных сторон соответствующего регистра (4).
- 15 Группа регулярно пересматривает и обновляет обязанности по управлению качеством данных (5).

ПРОЦЕССЫ ОБЕСПЕЧЕНИЯ КАЧЕСТВА ДАННЫХ

- 16 Специфицирован профиль качества, который используется для выявления недостатков процесса (1, 2).
- 17 Поддающиеся измерению показатели качества данных определены и задокументированы (1, 2).
- 18 Качество данных измеряется, и недостатки выявляются на ранней стадии процесса (3).

- 19 Организация ведет поиск путей предотвращения возникновения проблем (1, 2).
- 20 Все выявленные проблемы регистрируются, и процесс их решения поддается отслеживанию (1, 2).
- 21 Процессы обеспечения качества данных определены и задокументированы (1, 2).
- 22 Определено воздействие недостатков на качество данных (2).
- 23 Измерения качества данных проводятся на регулярной основе (4).
- 24 Результаты измерений предоставляются администрации (4).
- 25 План действий по улучшению качества данных определен и задокументирован (4).
- 26 Изучение коренных причин недостатков в качестве данных является обычной практикой (4).
- 27 Показатели качества данных регулярно изучаются в целях выявления дополнительных возможностей совершенствования процессов (5).

ЗНАНИЯ И КОМПЕТЕНЦИИ В ОБЛАСТИ КАЧЕСТВА ДАННЫХ

- 28 Проводится учеба по вопросам качества данных с целью повышения осведомленности о проблемах качества данных (1, 2).
- 29 Ключевые сотрудники разрабатывают планы и дают рекомендации по управлению качеством данных (1, 2).
- 30 Работники, которые могут влиять на качество данных, прошли обучение по вопросам качества данных (3).
- 31 Используется практика неформального наставничества для повышения осведомленности о проблемах качества данных (3).
- 32 Передовая практика в области обеспечения качества данных документируется в виде общедоступной базы знаний (3).
- 33 Учеба по вопросам качества данных организуется на регулярной основе (4).
- 34 Содержание программы обучения по вопросам качества данных регулярно пересматривается и обновляется по мере необходимости (5).

ТЕХНОЛОГИЧЕСКИЕ ИНСТРУМЕНТЫ ОБЕСПЕЧЕНИЯ КАЧЕСТВА ДАННЫХ

- 35 Стандарты качества данных для инструментов ИТ разработаны и задокументированы (1, 2).
- 36 Предпринимаются усилия по созданию баз данных, соответствующих модели «единый источник достоверных данных» (2).
- 37 Внедрены и используются инструменты для измерения и совершенствования качества данных (1, 2, 3).
- 38 Соответствие стандартам качества данных контролируется на этапе выдвижения новых предложений по проектам (3).
- 39 Отчетность по качеству данных содержит аналитические функции для оказания помощи в принятии управленческих решений, касающихся качества данных (4).

40. Согласован целевой портфель мер по развитию ИТ-систем в поддержку управления качеством данных (4).
 41. Внедрены автоматизированные процедуры проверки и корректировки данных на основе профиля качества данных (4).
 42. При разработке баз данных используются принципы первичных данных (5).
 43. Целевой портфель мер по развитию ИТ-систем в поддержку управления качеством данных регулярно пересматривается и обновляется по мере необходимости (5).
-