

УДК 004.89:004.4

Г. В. Дорохина

Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк  
83048, г. Донецк, ул. Артема, 118-б

## ТРЕБОВАНИЯ К ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ ЦИФРОВОГО СБОРА, ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

G. V. Dorokhina

Public Institution "Donetsk national technical University", Donetsk  
83048, Donetsk, Artyoma str., 118-b

## REQUIREMENTS TO INFORMATION TECHNOLOGY FOR COLLECTING, PROCESSING AND ANALYZING OF DATA IN DIGITAL FORM

Г. В. Дорохіна

Державна установа «Інститут проблем штучного інтелекту», м. Донецьк  
83048, м. Донецьк, вул. Артема, 118-б

## ВИМОГИ ДО ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ЦИФРОВОГО ЗБОРУ ОБРОБКИ І АНАЛІЗУ ДАНИХ

Проведен анализ современных публикаций, связанных с задачей цифрового сбора, обработки и анализа данных. Выделены и обобщены требования к собираемым данным; требования к сбору, хранению данных и способу их представления; требования к регламентации; требования к обработке данных; требования к продуцированию, предоставлению и публикации данных; требования к организации программного обеспечения, реализующего информационную технологию.

**Ключевые слова:** информационная технология; цифровой сбор, обработка и анализ данных; требования.

The analysis of modern publications related to the task of collecting, processing and analyzing of data in digital form is carried out. Highlighted and summarized the requirements for the collected data; requirements for the collection, storage of data and the way they are presented; regulatory requirements; data processing requirements; requirements for the production, provision and publication of data; requirements for the organization of software that implements information technology.

**Key words:** information technology; collecting, processing and analyzing of data in digital form; requirements.

Проведено аналіз сучасних публікацій, пов'язаних із завданням цифрового збору, обробки та аналізу даних. Виділено і узагальнено вимоги до даних, щодо яких виконується збирання; вимоги до збирання, зберігання даних і способу їх подання; вимоги до регламентації; вимоги до обробки даних; вимоги до продукування, надання та публікації даних; вимоги до організації програмного забезпечення, що реалізує інформаційну технологію.

**Ключові слова:** інформаційна технологія; цифровий збір, обробка та аналіз даних; вимоги.

**Актуальность работы.** Успешное развитие предприятий и отраслей экономики зачастую связано с применением современных инструментов поддержки принятия решений. Они работают с оцифрованными структурированными данными, сохраняемыми в процессе работы объектов управления. На уровне государства эти данные формируются из данных статистических наблюдений, плановых мониторингов и разовых запросов, логистических данных. В совокупности они образуют «цифровой портрет государства» – неотъемлемую составляющую цифровой экономики.

**Цель работы.** Для обеспечения совместимости данных статистических наблюдений государства, ведомственных данных, плановых мониторингов и проч. при сборе, обработке и анализе данных необходимо использовать единую информационную технологию. В связи с этим возникла необходимость разработки и формулирования требований к единой (универсальной) информационной технологии цифрового сбора, обработки и анализа данных.

**Анализ последних исследований и публикаций.** Последние годы в Российской Федерации ведутся работы по созданию цифровой платформы сбора и анализа данных нового поколения, нацеленной на переход к новой структурной и функциональной модели производства и распространения статистических данных в России. Данная модель должна реализовывать принципы единого информационного пространства данных, прослеживаемости и взаимосвязи понятий и объектов статистического учета, однократного предоставления первичных статистических данных и их многократного последующего использования. Требования информационной технологии цифрового сбора данных приведены в стратегии развития Росстата до 2024 г. [1].

В настоящее время эксплуатируется множество независимых информационных систем, ориентированных на определённого функционального заказчика. Это приводит к разрозненности управления данными [2].

Требования к информационной технологии цифрового сбора, обработки и анализа данных вытекают из проблем, стоящих на пути цифровизации. Применительно к работе с данными статистических наблюдений выделяют следующие актуальные проблемы [3]: межведомственная рассогласованность статистической деятельности, непрозрачность и противоречивость информации, предоставляемой отдельными органами власти, отсутствие единого методологического надзора; недостаток актуальной, кастомизированной и детализированной информации для бизнеса, науки, гражданского общества, различных групп населения; ограниченность доступа к первичным данным для аналитических и управленческих целей, фрагментарность работ по их деперсонификации.

В Концепции создания цифровой аналитической платформы предоставления статистических данных [4] присутствует требование применения онтологии данных, используемой в модели государственных данных. Онтологические модели и графы знаний уже прочно вошли в сферы промышленного производства [5], инженерии требований, интеграции информационно-коммуникационных систем [6], поддержки жизненного цикла сложных организационно-технических объектов (СОТО) [7], взаимодействия с потребителями продукции. Среди них можно выделить как средства единого формализованного описания предметных областей, ориентированные на людей-специалистов, так и средства для промышленных «безлюдных» технологий [5].

В задаче поддержки жизненного цикла СОТО [7] выделяют такие типы моделей: информационные онтологические модели описания словарей источников данных; поведенческие модели, характеризующие событийно-ориентированные процессы предметной области, связанные с ними модели интеграционных процессов и согласования аналитических вычислительных задач; функциональные модели для описания решения аналитических вычислительных задач. Перспективным считается использование совокупности взаимосвязанных онтологий в виде онтологической системы как способа интеграции и согласования разнородных знаний с использованием мульти-агентного подхода при решении задачи обеспечения семантической интероперабельности элементов распределенного программного комплекса информационно-аналитической поддержки жизненного цикла СОТО.

При разработке автоматизированного интеллектуального программного комплекса «Государственная информационная система Образование» [8] автором решена задача хранения и обработки в единой базе данных детализированную информацию о документах об образовании для всех уровней образования (среднее, среднее профессиональное, высшее профессиональное, и др.) Хранить в общих таблицах и обрабатывать по единому принципу данные значительно различающихся между собой документов об образовании позволило использование «шаблонов». Этот принцип может быть полезен и в технологии цифрового сбора, обработки и анализа данных.

**Основной материал исследования.** Обобщим требования к информационной технологии цифрового сбора, обработки и анализа данных на основе анализа работ [1-7].

*Требования к собираемым данным:* достоверность данных, их высокий аналитический потенциал, в том числе для проведения многомерного комплексного анализа, других исследований [3]; наличие базы данных пользователей [2]; наличие реестра объектов наблюдения [2], [3].

*Требования к сбору, хранению данных и способу их представления:*

- централизация сбора и хранения данных, однократный ввод и многократное использование данных по единой методологии, использование данных в долгосрочной перспективе, непрерывное наблюдение за отдельными сферами [3];
- экономичность и прозрачность процессов сбора и обработки данных [3];
- применение распределённых моделей сбора, обработки, хранения, предоставления и распространения данных [3]; потоковая модель сбора первичных данных, формируемых в автоматизированных системах первичного учета [1], [2];
- открытость, соответствие систем показателей, классификаций, методологических подходов, инструментария международным стандартам [3];
- связанность, сопоставимость, однозначная интерпретируемость сведений, полученных из разных источников [3];
- возможность встраивания альтернативных источников информации [1, 3];
- пригодность для хранения данных об относительно разрозненных, но взаимосвязанных элементах с возможностью их динамической обработки [2].

*Требования к регламентации:* правовая и техническая регламентация работы [2]; унификация и гармонизация определений, классификаций, алгоритмов расчета показателей, способов передачи и методов обработки первичных данных [4], процедур верификации данных [3], в том числе: единые форматы представления данных в электронном виде [1]; единые средства описания (метаданные), позволяющие конструировать необходимые аналитические агрегаты из показателей, связанных на любом уровне иерархической структуры и по разным основаниям [3]; единый порядок организации приема и обработки первичных данных [3]; единый генератор XML-шаблонов форм (собираемых) данных; единый язык описания арифметических и логических контролей; единые форматы транспортных сообщений при обмене электронными документами [1]; формализованное подробное описание всего технологического процесса обработки и формирования информации, включая, при необходимости, описание процессов межформенного согласования показателей [1]; верифицируемость данных и сценариев обработки на предмет их корректности [7].

*Требования к обработке данных:*

- результаты обработки первичных данных должны сохраняться на отдельном от источников данных ресурсе [7];
- возможность обработки средствами искусственного интеллекта [1];
- возможность преобразовывать потоки данных в привязке к классификаторам, отличным от тех, с использованием которых они созданы [2];
- пригодность хранимых данных для использования OLAP-системой специального вида, включающей множество частных задач мониторинга состояния СОТО по содержательно различным видам информации [7];

– пригодность для формирования интегрированной информационной среды с чертами предметной ориентированности, интегрированности, неизменчивости и поддержки хронологии [7];

– пригодность для формирования моделей состояния (электронного паспорта) объектов, с ориентацией на возможность представления их конфигурации [7];

– пригодность для интеграции с формальными системами последовательно уточняющих согласованных спецификаций поведенческих, функциональных и информационных требований [7];

– пригодность для интеграции со спецификациями на основе совокупности декларативно-процедурных моделей представления знаний с графическими нотациями с целью описания предметной области без обращения к алгоритмическому уровню (без программирования в кодах) [7];

– пригодность для выполнения «многоязычного моделирования» и проблемной ориентированности моделей представления знаний для экспертов разных специализаций: бизнес-процессов, информационных структур, аналитических задач, алгоритмов их решения [7];

– использование онтологий как формализма для согласования и верификации собираемых данных и проектируемых на их основе моделей [7];

– задание онтологических отображений на источники данных для обеспечения доступа к ним в терминах предметной области [7];

– применение онтологических моделей для описания сценариев и шаблонов сценариев в терминах последовательности событий и действий, порядка их взаимодействия через сообщения, в том числе для описания порядка выполнения аналитических задач, подготовки данных для них и записи их результатов в терминах предметной области [7].

*Требования к продуцированию, предоставлению и публикации данных:*

– непротиворечивость, независимость и объективность продуцируемых данных, актуальность и своевременность их предоставления, равный и удобный доступ к данным, оперативность представления и адаптивность к запросам пользователей, возможность многократного использования данных, применения к ним интерактивных сервисов и инструментов визуализации данных [3];

– распространение публикация производимой информации в виде логически увязанных последовательностей (например, построенные в единой методологии временные ряды, ранжированные последовательности) в машиночитаемом формате с возможностью преобразования в привычные виды человекочитаемых форматов [1];

– возможность уточнения / кардинального изменения результатов анализа предшествующих наблюдений за счет новых поступающих данных [3];

– пригодность для публикации в открытом доступе не только данных, но и обобщенных моделей (в декларативном виде / в виде инструментальных средств) [3].

*Требования к программному обеспечению, реализующему технологию:* интеграция технологии в существующую инфраструктуру без разрушения существующих прогрессивных механизмов сбора данных [3]; функционирование в режиме реального времени [7]; автоматизированность системы сетевого взаимодействия пользователей [2]; структурная организация в виде платформы и модулей, регламентирование интерфейса [2]; параллелизм, асинхронность и потоковость вычислительных процессов; функционирование на основе принципа управления по данным [7].

## Выводы

На основе анализа современных публикаций выделены и обобщены требования к информационной технологии цифрового сбора, обработки и анализа данных: к собираемым данным; к сбору, хранению данных и способу их представления; к регламентации; к обработке данных; требования к продуцированию, предоставлению и публикации данных; к организации программного обеспечения, реализующего информационную технологию.

## Список литературы

1. Стратегия развития Росстата и системы государственной статистики Российской Федерации до 2024 года [Электронный ресурс] – 2019. – 48 с. – URL: <https://rosstat.gov.ru/storage/mediabank/Strategy.pdf>
2. Пашков А. М. Новые горизонты российской статистики и упорядочение процессов цифровой аналитической платформы [Текст] / А. М. Пашков, Т. И. Ларинина // Российские регионы: взгляд в будущее. – vol. 7, № 3, 2020. – С. 113–123.
3. Перспективная модель государственной статистики в цифровую эпоху [Текст]: докл. к XIX Апр. междунар. науч. конф. по проблемам развития экономики и общества, Москва, 10–13 апр. 2018 г. / науч. ред. Л. М. Гохберг ; Нац. исслед. ун-т «Высшая школа экономики». — М.: Изд. дом Высшей школы экономики, 2018. – 35, [1] с.
4. Распоряжение Правительства Российской Федерации от 17 декабря 2019 г. № 3074-р. [Текст]
5. Муромцев Д. Индустриальные графы знаний – интеллектуальное ядро цифровой экономики [Текст] / Дмитрий Муромцев, Алексей Романов, Дмитрий Волчек // Control Engineering Россия. – № 5 (83). – октябрь 2019. – С. 23–39.
6. Родионцев Н. Н. Анализ применения онтологий при разработке информационно-коммуникационных систем нефтегазовой отрасли [Текст] / Н. Н. Родионцев // Московский экономический журнал. – 2019. – № 2. – С. 695–699.
7. Охтилев П.А. Алгоритмы и онтологические модели информационно-аналитической поддержки процессов создания и применения космических средств [Текст]: Диссертация на соискание ученой степени кандидата технических наук. Специальность 05.13.01 – Системный анализ, управление и обработка информации (технические системы) Дата защиты: 19.12.2019 г./ Петр Алексеевич Охтилев. – Санкт-Петербург, 2019. – 408 с.
8. Дорохина Г. В. Разработка автоматизированного интеллектуального программного комплекса «Государственная информационная система Образование» [Текст] / Г. В. Дорохина, В. И. Финаев // Проблемы автоматизации. Региональное управление. Связь и автоматика (ПАРУСА-2018) : Сборник трудов VII Всероссийской научной конференции молодых ученых, аспирантов и студентов, г. Геленджик, 2018 : в 2 т. / сост. Ю. Б.Щемелева, С. В. Кирильчик ; Южный федеральный университет. – Ростов-на-Дону ; Таганрог : Издательство Южного федерального университета, 2018. Т. 2. – С. 20–25.
9. Дорохина Г.В. Формализованные онтологии и задача построения компьютерной информационной технологии цифрового сбора, обработки и анализа данных [Текст] / Г.В. Дорохина // Сборник тезисов Международного круглого стола «Искусственный интеллект: теоретические аспекты и практическое применение». – 2020. – С. 229–234.

## References

1. *Strategiya razvitiya Rosstata i sistemy gosudarstvennoy statistiki Rossiyskoy Federatsii do 2024 goda*. [Development strategy of Rosstat and the system of state statistics of the Russian Federation until 2024] 2019, 48 s. URL: <https://rosstat.gov.ru/storage/mediabank/Strategy.pdf>
2. Pashkov A. M., Larinina T. I. *Novyye gorizonty rossiyskoy statistiki i uporyadocheniye protsessov tsifrovoy analiticheskoy platform* [New horizons of Russian statistics and streamlining the processes of the digital analytical platform]. *Rossiyskiye regiony: vzglyad v budushcheye* [Russian regions: a look into the future], vol. 7, No 3, 2020, S. 113-123.
3. *Perspektivnaya model' gosudarstvennoy statistiki v tsifrovuyu epokhu: dokl. k XIX Apr. mezhdunar. nauch. konf. po problemam razvitiya ekonomiki i obshchestva* [A promising model of state statistics in the digital age: reports. to 19th Apr. int. scientific. conf. on the problems of economic and social development], Moskva, 10–13 apr. 2018 g. / nauch. red. L. M. Gokhberg ; Nats. issled. un-t «Vysshaya shkola ekonomiki». M., Izd. dom Vysshey shkoly ekonomiki, 2018. 35, [1] s.
4. *Rasporyazheniye Pravitel'stva Rossiyskoy Federatsii ot 17 dekabrya 2019 g. No 3074-r*. [Order of the Government of the Russian Federation of December 17, 2019 No. 3074-r]
5. Muromtsev D., Romanov A., Volchek D. *Industriallye grafy znaniy – intellektual'noye yadro tsifrovoy ekonomiki* [Industrial knowledge graphs - the intellectual core of the digital economy]. *Control Engineering Rossiya* [Control Engineering Russia], No. 5 (83) oktyabr' 2019. S. 23-39.
6. Rodiontsev N. N. *Analiz primeneniya ontologiy pri razrabotke informatsionno-kommunikatsionnykh sistem neftegazovoy otrasli* [Analysis of the application of ontologies in the development of information and communication systems of the oil and gas industry]. *Moskovskiy ekonomicheskij zhurnal* [Moscow economic journal], No 2, 2019, S. 695 – 699.
7. Okhtilev P.A. *Algoritmy i ontologicheskiye modeli informatsionno-analiticheskoy podderzhki protsessov sozdaniya i primeneniya kosmicheskikh sredstv* [elektronnyy resurs]: Dissertatsiya na soiskaniye uchenoy stepeni kandidata tekhnicheskikh nauk. Spetsial'nost' 05.13.01 – Sistemnyy analiz, upravleniye i obrabotka informatsii (tekhnicheskkiye sistemy) Data zashchity: 19.12.2019 g [Algorithms and ontological models of information and analytical support of the processes of creation and use of space vehicles], Sankt-Peterburg, 2019, 408 s.
8. Dorokhina G.V., Finayev V.I. *Razrabotka avtomatizirovannogo intellektual'nogo programmno kompleksa «Gosudarstvennaya informatsionnaya sistema Obrazovaniye»* [Development of an automated intelligent software complex "State information system Education"]. *Problemy avtomatizatsii. Regional'noye upravleniye. Svyaz' i avtomatika* (PARUSA-2018) [Problems of automation. Regional administration. Communication and automation (SAIL-2018)] *Sbornik trudov VII Vserossiyskoy nauchnoy konferentsii molodykh uchenykh, aspirantov i studentov, g. Gelendzhik, 2018 : v 2 t. / sost. YU. B. Shchemeleva, S. V. Kiril'chik ; Yuzhnyy federal'nyy universitet, Rostov-na-Donu ; Taganrog : Izdatel'stvo Yuzhnogo federal'nogo universiteta, 2018, T. 2, s. 20-25.*

9. Dorokhina G.V. Formalizovannyye ontologii i zadacha postroyeniya komp'yuternoy informatsionnoy tekhnologii tsifrovogo shora, obrabotki i analiza dannykh. [Formalized ontologies and the problem of building a computer information technology for digital collection, processing and analysis of data] *Sbornik tezisev Mezhdunarodnogo kruglogo stola «Iskusstvennyy intellekt: teoreticheskiye aspekty i prakticheskoye primeneniye»* [Collection of Abstracts of the International Round Table "Artificial Intelligence: Theoretical Aspects and Practical Application"] 2020, S. 229-234.

## RESUME

*G. V. Dorokhina*

### *Requirements to Information Technology for Collecting, Processing and Analyzing of Data in Digital Form*

To ensure the compatibility of the data of statistical observations of the state, departmental data, scheduled monitoring, etc. when collecting, processing and analyzing data, it is necessary to use a unified information technology. In this regard, it became necessary to develop requirements for a unified (universal) information technology for digital collection, processing and analysis of data.

The analysis of modern publications related to the task of digital collection, processing and analysis of data is carried out. Highlighted and summarized the requirements for the collected data; requirements for the collection, storage of data and the method of their presentation; regulatory requirements; data processing requirements; requirements for the production, provision and publication of data; requirements for the organization of software that implements information technology.

It was proposed to use the principle of "templating" to solve certain problems of storing disparate data, tested by the author, and the problem of storing, processing and managing detailed information in a single database of documents on education for all levels of education was solved.

As a result, the author developed requirements for a unified information technology for digital collection, processing and analysis of data: for the collected data; collection, storage and presentation of data; to regulation; to data processing; requirements for the production, provision and publication of data; to the organization of software that successfully implements information technology.

## РЕЗЮМЕ

*Г. В. Дорохина*

### *Требования к информационной технологии цифрового сбора, обработки и анализа данных*

Для обеспечения совместимости данных статистических наблюдений государства, ведомственных данных, плановых мониторингов и проч. при сборе, обработке и анализе данных необходимо использовать единую информационную технологию. В связи с этим возникла необходимость разработки требований к единой (универсальной) информационной технологии цифрового сбора, обработки и анализа данных.

Проведен анализ современных публикаций, связанных с задачей цифрового сбора, обработки и анализа данных. Выделены и обобщены требования к собираемым данным; требования к сбору, хранению данных и способу их представления; требования к регламентации; требования к обработке данных; требования к продуцированию, предоставлению и публикации данных; требования к организации программного обеспечения, реализующего информационную технологию.

Предложено для решения определенных проблем хранения разрозненных данных использовать принцип «шаблонизации», апробированный автором и была решена задача хранения, обработки и управления детализированной информацией в единой базе данных о документах об образовании для всех уровней образования.

В результате автором были разработаны требования к единой информационной технологии цифрового сбора, обработки и анализа данных: к собираемым данным; к сбору, хранению данных и способу их представления; к регламентации; к обработке данных; требования к продуцированию, предоставлению и публикации данных; к организации программного обеспечения, успешного реализующего информационную технологию.

Статья поступила в редакцию 02.11.2020.