DOI: 10.51790/2712-9942-2023-4-2-07

ЕДИНЫЙ ГОСУДАРСТВЕННЫЙ ФОНД ГИДРОМЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ КАК БОЛЬШИЕ ДАННЫЕ. УНИВЕРСАЛЬНЫЙ ПАРСЕР СТРУКТУРЫ ДАННЫХ В ФОРМАТЕ ЯЗЫКА ОПИСАНИЯ ГИДРОМЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ

Л. О. Перетятько

Федеральное государственное бюджетное учреждение «Всероссийский научно-исследовательский институт гидрометеорологической информации — Мировой центр данных», г. Обнинск, Российская Федерация

ORCID: https://orcid.org/0000-0002-9182-2952, ✍ peretyatkol@meteo.ru

Аннотация: данная статья относится к серии статей, посвященной работе с данными, размещёнными в Едином государственном фонде данных Федеральной службы России по гидрометеорологии и мониторингу окружающей среды (Росгидромет), а именно, технологиям и инструментам для работы с этими данными как с большими данными.

Статья посвящена универсальному парсеру структуры данных в формате ЯОД (языка описания гидрометеорологических данных). Структура данных в формате ЯОД также называется ЯОД описанием. Разработанный парсер осуществляет анализ, разбор и сохранение структуры ЯОД описания для последующей конвертации данных в этом формате. Для реализации парсера была разработана система классов на основе описания формата ЯОД. Результатом работы парсера является структура ЯОД в виде набора взаимосвязанных объектов разработанных классов для обеспечения возможности чтения или записи данных в формате ЯОД. Универсальный парсер является ключевой составляющей комплекса программных средств взаимной конвертации данных в формате ЯОД в современные распространённые форматы.

В работе представлены диаграмма классов в нотации UML и их описание, особенности универсального парсера, подробное описание этапов его работы.

Универсальный парсер ЯОД описания разработан на базе Всероссийского научно-исследовательского института гидрометеорологической информации — Мирового центра данных.

Ключевые слова: язык описания гидрометеорологических данных, Единый государственный фонд данных, ЕГФД, парсер, технология обработки данных, иерархическая структура данных.

Для цитирования: Перетятько Л. О. Единый государственный фонд гидрометеорологических данных как большие данные. Универсальный парсер структуры данных в формате языка описания гидрометеорологических данных. *Успехи кибернетики*. 2023;4(2):47–52. DOI: 10.51790/2712-9942-2023-4-2-07.

Поступила в редакцию: 10.04.2023. *В окончательном варианте: 10.04.2023.*

THE NATIONAL HYDROMETEOROLOGICAL ARCHIVE AS BIG DATA. A UNIVERSAL PARSER OF THE HYDROMETEOROLOGICAL DATA DESCRIPTION LANGUAGE

L. O. Peretyatko

All-Russia Research Institute of Hydrometeorological Information, World Data Center, Obninsk, Russian Federation

ORCID: https://orcid.org/0000-0002-9182-2952, ✍ peretyatkol@meteo.ru

Abstract: this paper refers to a series of papers on handling the big data stored in the National Hydrometeorological Archive (NHA) managed by the Russian Federal Service for Hydrometeorology and Environmental Monitoring.

This study presents a universal parser of the data structure in the HDDL format (hydrometeorological data description language). The parser analyzes, parses, and saves the HDDL document structure for subsequent conversion. We developed a number of classes to handle the HDDL format. The parser produces a data source structure as a set of interconnected objects of the classes to read or write data in the HDDL format. The universal parser is a key component of the software package for HDDL data conversion into other common formats.

The paper presents an UML diagram of the classes and their description, features of the universal parser, and a detailed description of its operation.

The universal parser is developed at the All-Russia Research Institute of Hydrometeorological Information, World Data Center.

Keywords: hydrometeorological data description language, National Hydrometeorological Archive, NHA, parser, data processing, hierarchical data structure.

Cite this article: Peretyatko L. O. The National Hydrometeorological Archive as Big Data. A Universal Parser of the Hydrometeorological Data Description Language. *Russian Journal of Cybernetics*. 2023;4(2):47–52. DOI: 10.51790/2712-9942-2023-4-2-07.

Original article submitted: 10.04.2023. *Revision submitted: 10.04.2023*.

Введение

Данная статья относится к серии статей, посвященной работе с данными, размещёнными в Едином государственном фонде данных Федеральной службы России по гидрометеорологии и мониторингу окружающей среды (Росгидромет), а именно, технологиям и инструментам для работы с этими данными как с большими данными [1].

Электронные данные Единого государственного фонда данных о состоянии окружающей среды, ее загрязнении (ЕГФД) хранятся в специальном формате ЯОД (язык описания гидрометеорологических данных). Особенностью этого формата является возможность описывать иерархическую структуру данных, получаемых от наблюдательной сети Росгидромет. Подробное описание формата ЯОД представлено в соответствующей статье [2].

Универсальный парсер структуры данных в формате ЯОД (ЯОД описания) реализован на языке программирования C++ [3], с использованием фреймворка Qt версии 5 [4]. Разработанный парсер является ключевой компонентой системы взаимной конвертации данных различной структуры [5].

Разработана система классов для парсера ЯОД описания, представленная в виде UML диаграммы [6] (см. рис 1).

Класс StringMassive представляет отдельные сегменты ЯОД описания, описывающие отдельные части записей или групп. Хранит текстовое описание сегмента, имя вышестоящего по иерархии сегмента и уровень вложенности.

Класс Element представляет элементы ЯОД описания. Хранит информацию о элементе: имя, тип, формат хранения, формат преобразования, список допустимых значений, комментарий и прочую вспомогательную информацию.

Класс Group представляет группу элементов ЯОД описания. Хранит список элементов, тип, имя, имя вышестоящей по иерархии группы, уровень вложенности, количество экземпляров, описание группы, а также другую вспомогательную информацию.

Класс Record представляет записи ЯОД описания. Хранит информацию о записи: имя, список целых групп, список фрагментов групп, номер и список отдельных сегментов описания записи.

Класс Core представляет ЯОД описание. Реализует методы для получения и сохранения ЯОД описания и его использования в последующей конвертации. Хранит список записей, заголовок каждой записи, название семейства ЯОД описания, тип данных в формате ЯОД, определяемый по ЯОД описанию.

Особенности парсера

Разработанный парсер ЯОД описания обладает следующими особенностями:

- универсальность — работа парсера была успешно протестирована на примере шестидесяти различных ЯОД описаний метеорологических, аэрологических, океанографических и гидрологических данных;

- использование регулярных выражений [7] для упрощения анализа и разбора на составляющие строк ЯОД описания;

- использование рекурсивных алгоритмов;

- сохранение особенностей иерархической структуры ЯОД, которые учитываются при дальнейшей конвертации;

- сохранение текста ЯОД описания в качестве метаданных.

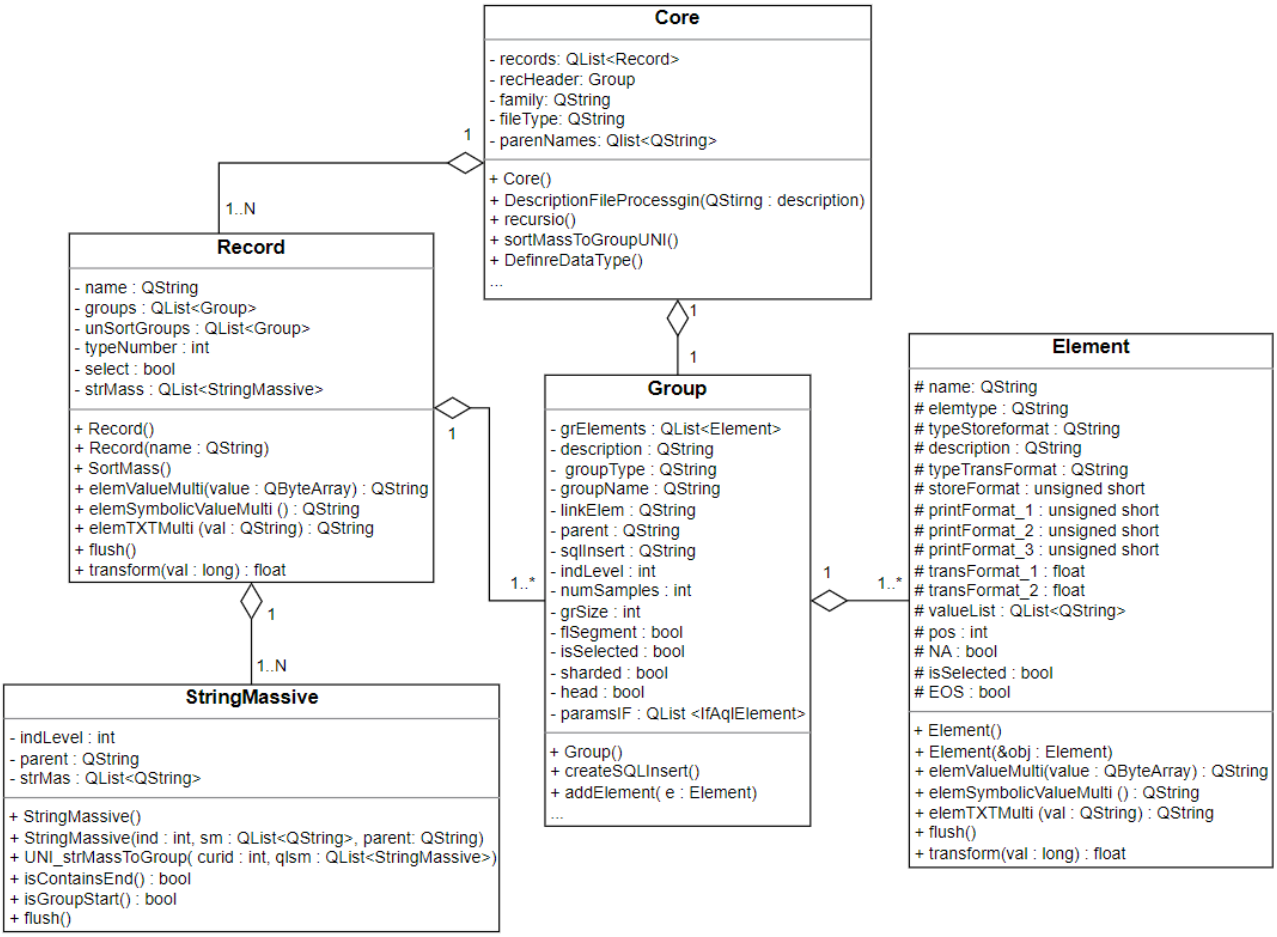

Рис. 1. *UML диаграмма классов парсера ЯОД описания*

Алгоритм работы парсера

Файл с ЯОД описанием считывается построчно. После чтения каждой строки из неё удаляются лишние пробелы в начале и в конце, множественные пробелы или табуляции заменяются на одинарные пробелы.

Работу парсера ЯОД описания можно разделить на четыре основных этапа.

На первом этапе ведётся поиск строки с ключевым словом FAMILY, в которой выделяется и сохраняется имя семейства файлов, указанного после ключевого слова (SI — имя семейства, см. рис 2). Затем выполняется поиск одного из двух возможных ключевых слов — RECORD или RECORDS. Наличие ключевого слова RECORDS означает наличие описания заголовка для множества типов записей в ЯОД описании (см. рис 2) и дальнейшее появление ключевого слова RBODY, обозначающего начало описания каждой отдельной записи. Заголовок каждой записи имеют одинаковую структуру и содержит данные, которые позволяют идентифицировать каждую запись в файле с данными. При обнаружении слова RECODS выполняется сохранение описания заголовка записей для дальнейшей обработки. Ключевое слово RECORD означает, что описывается только одна запись, а описание заголовка отсутствует. При обнаружении слова RECORD выполняется чтение и сохранение описания всей записи.

На втором этапе, если обнаружено ключевое слово RECORDS:

- выполняется обработка текстового описания заголовка записей и сохранение его в виде объекта класса Group (поле recHeader класса Core);

- выполняется поиск ключевого слова RBODY и считывание описания каждой записи до конца файла. Описание каждой записи заканчивается комбинацией ключевого слова END и именем записи, указанного после RBODY;

- выполнение рекурсивной обработки полученного описания записи — разделение на отдельные текстовые сегменты (см. рис 3). Для каждого сегмента определяется уровень вложенности (начинается

```
FAMILY SI V(4044);

RECORDS;
 LNG     ДЛЗАП      B(2) PC(4);  //
 MIT     НУЛИ       B(2) PC(4);  //
 KEY(I,СПГОД)  ГОД             B(2) PC(4);  // Год
 KEY(I,СПМЕСЯЦ ) МЕСЯЦ   B(1) PC(2);  // Месяц
 KEY(U)  СТАНЦИЯ  B(4) PC(7);  // Станция
 MRC(I)   ТИПЗАП    B(1) PC(2);  // Тип записи (1-12)
```

Рис. 2. *Описание семейства ЯОД файлов и заголовка записей в ЯОД описании*

с нуля) и имя вышестоящего сегмента. Последние два шага выполняются для каждой следующей записи.

Если обнаружено ключевое слово RECORD, выполняется рекурсивная обработка и преобразование записи в объект.

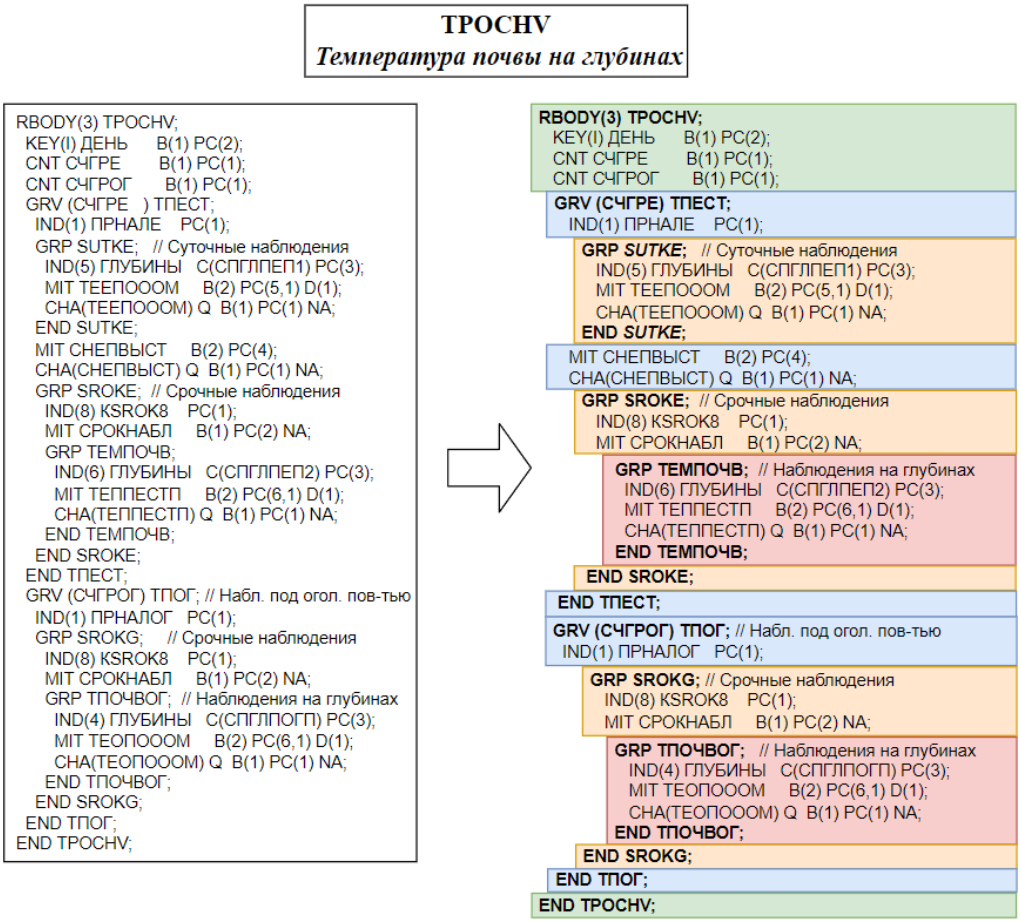

Рис. 3. *Разделение ЯОД описания записи TPOCHV на отдельные тестовые сегменты*

На третьем этапе в записи выполняется объединение текстовых сегментов, описывающих одну и ту же группу или элементы записи на одном и том же уровне. Таким образом получается список объединённых сегментов, который можно представить в виде иерархической структуры. На рисунке 4 представлен пример результата работы третьего этапа.

На четвертом этапе выполняется преобразование списка сегментов записи в список объектов класса Group, каждый из которых содержит полную информацию о группе. Причём преобразованию

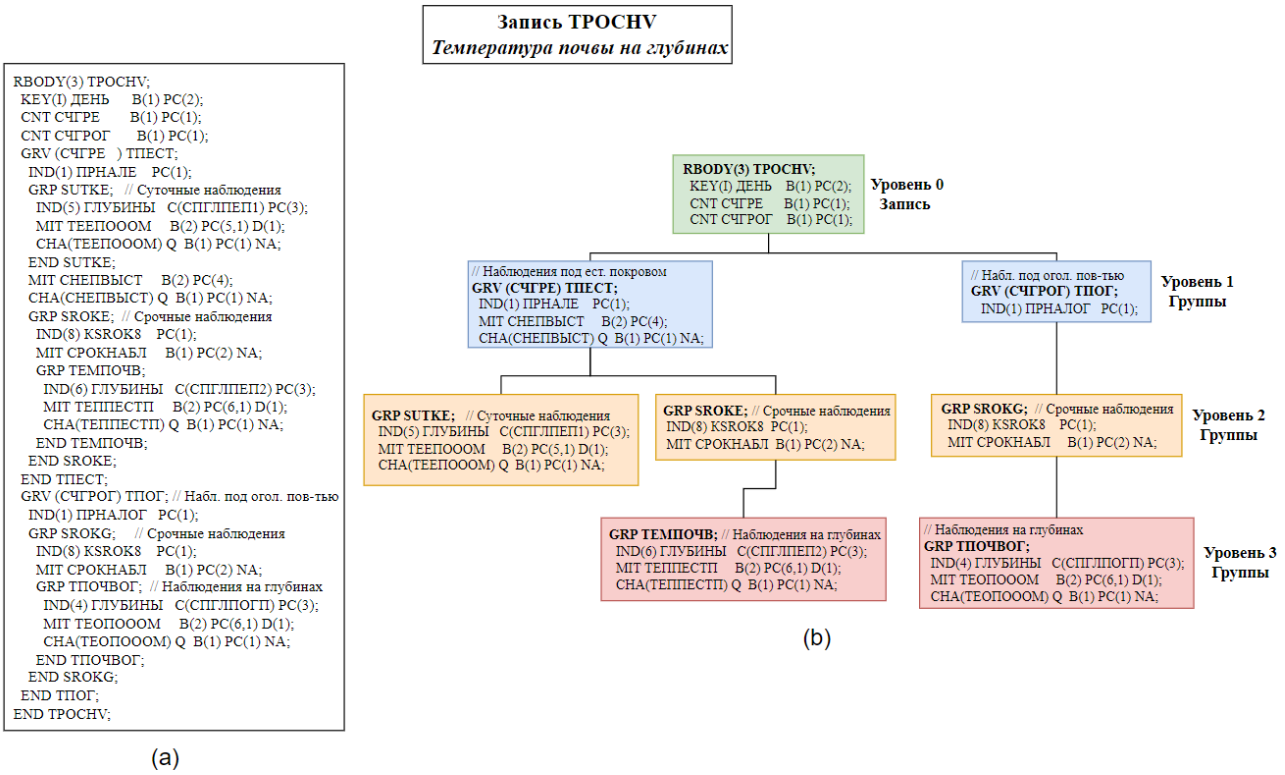

Рис. 4. *Запись TPOCHV в формате ЯОД: (a) описание записи TPOCHV в формате ЯОД; (b) иерархическая структура записи TPOCHV из объединённых сегментов*

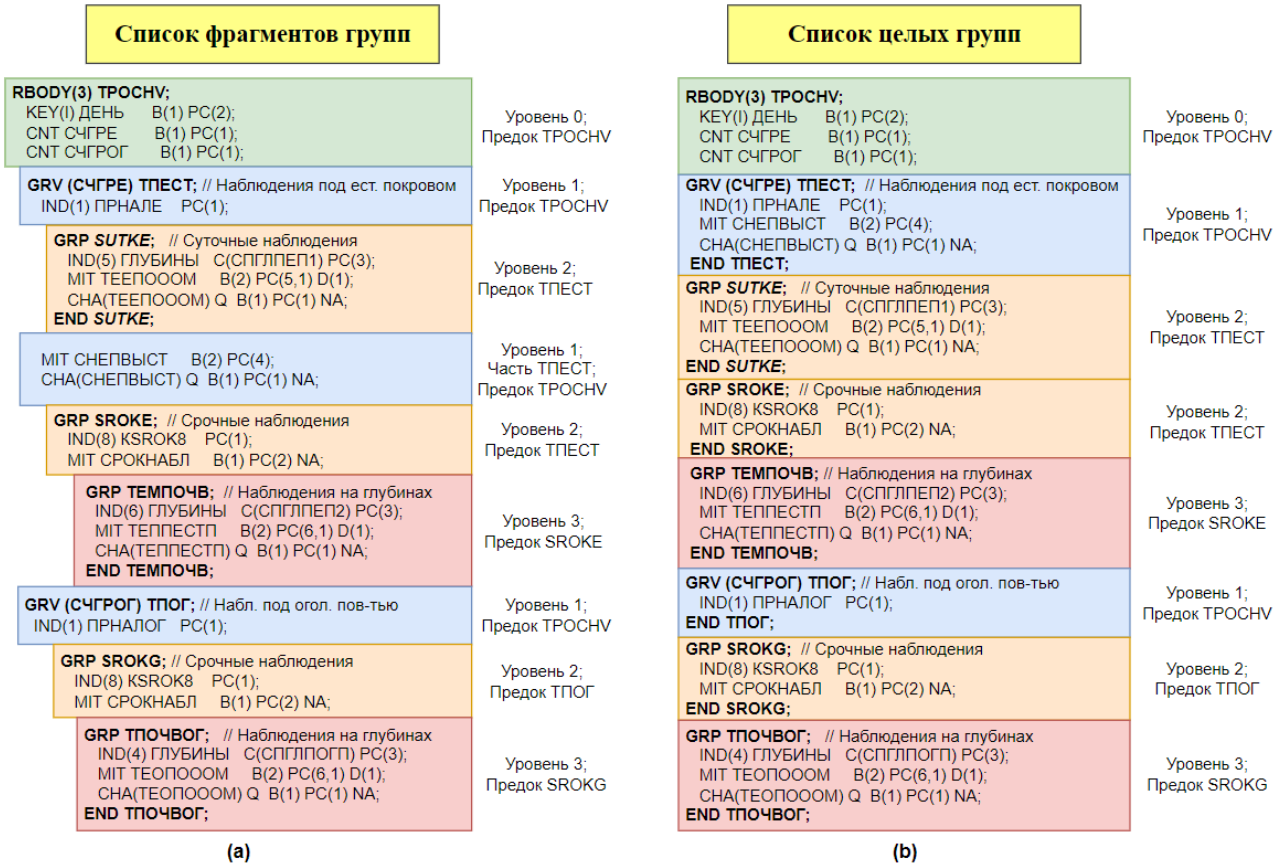

Рис. 5. *Списки описания групп: (a) список частей групп; (b) список целых групп*

подвергаются как список отдельных сегментов записи, так и список объединённых сегментов. В итоге получается список целых групп (см. рис. 5.a) и список частей групп (см. рис. 5.b).

Список целых групп используется для создания структуры, например, в реляционной базе данных, и выстраивания связей между таблицами в ней. Список частей группы отражает порядок ЯОД и используется при чтении данных в формате ЯОД.

Заключение

В настоящей статье были представлены: диаграмма классов в нотации UML с подробным описанием каждого класса, особенности разработанного парсера, поэтапное описание алгоритма работы парсера на примере описания записи TPOCHV (температура почв на глубине), относящейся к восьмисрочным метеорологическим данным. В результате тестирования работы парсера на примере шестидесяти различных ЯОД описаний была доказана универсальность разработанного парсера.

ЛИТЕРАТУРА

1. Перетятько Л. О. Единый государственный фонд гидрометеорологических данных как большие данные. Технологии и инструменты для работы с ним. *Успехи кибернетики*. 2022;3(4):98–101. DOI: 10.51790/2712-9942-2022-3-4-11.
2. Кофтан Ю. Р., Перетятько Л. О. К построению технологии взаимной конвертации баз данных различной структуры для пополнения и верификации данных ЕГФД, а также для обслуживания потребителей. *Труды ВНИИГМИ-МЦД*. 2018;181:162–174.
3. Шлее М. *Qt 5.10. Профессиональное программирование на C++*. СПб.: БХВ-Петербург; 2018. 1072 с.
4. *Документация Qt версии 5.* Режим доступа: https://doc.qt.io/qt-5/qstring.html.
5. Перетятько Л. О., Кофтан Ю. Р. Система взаимной конвертации данных различной структуры для обслуживания потребителей ЕГФД. *Труды ВНИИГМИ-МЦД*. 2020;186:163–175.
6. Арлоу Д., Нейштадт И. *UML 2 и Унифицированный процесс, Практический объектно-ориентированный анализ и проектирование*, 2-е изд. / пер. с англ. СПб: Символ-Плюс; 2007. 624 с.
7. Фицджеральд М. *Регулярные выражения: основы* / пер. с англ. М.: ООО «И.Д. Вильямс»; 2015. 144 с.