DOI: 10.51790/2712-9942-2022-3-4-11

ЕДИНЫЙ ГОСУДАРСТВЕННЫЙ ФОНД ГИДРОМЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ КАК БОЛЬШИЕ ДАННЫЕ. ТЕХНОЛОГИИ И ИНСТРУМЕНТЫ ДЛЯ РАБОТЫ С НИМ

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: статья посвящена вопросу рассмотрения данных, содержащихся в Едином государственном фонде данных (ЕГФД) Федеральной службы России по гидрометеорологии и мониторингу окружающей среды (Росгидромет), как больших данных (“Big Data”), а также описанию технологий и инструментов для работы с данными ЕГФД. В ЕГФД хранятся данные как на бумажных носителях, так в электронном виде, по гидрометеорологии и смежным с ней областям (метеорологии, аэрологии, гидрологии, океанологии и др.). Особый исследовательский интерес для аналитиков данных и специалистов по обработке данных могут представлять первичные данные в электронном виде. Электронные первичные данные распределены по слоям. В работе представлено обоснование классификации данных ЕГФД как больших данных, также описаны особенности хранящихся в ЕГФД данных. Кроме того, в качестве примера больших данных приведены данные специализированных массивов для климатических исследований, расположенные в открытом доступе, хранимые и пополняемые одним из учреждений Росгидромета — Федеральным государственным бюджетным учреждением «Всероссийский научно-исследовательский институт гидрометеорологической информации — Мировой центр данных» (ФГБУ «ВНИИГМИ-МЦД»). Приведен перечень предполагаемых для разработки технологий и инструментов для повышения эффективности работы с данными ЕГФД.

Ключевые слова: Госфонд Росгидромета, большие данные, язык описания гидрометеорологических данных, технологии обработки данных, технологии конвертации данных.

Для цитирования: 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.

THE NATIONAL HYDROMETEOROLOGICAL ARCHIVE AS BIG DATA. APPROPRIATE TECHNOLOGIES AND TOOLS

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: we considered the records of the National Hydrometeorological Archive (NHMA) as Big Data and presented the technologies and tools for processing the NHMA data. The NHMA stores information on hydrometeorology and related areas (meteorology, aerology, hydrology, oceanology, etc.) both on paper and digitally. Digital primary data are of particular interest to data analysts. The data are divided into layers. This paper presents the rationale for considering the USDF records as big data and describes their features. We used the publicly available climate research records stored and replenished by the All-Russia Research Institute of Hydrometeorological Information, World Data Center as an example of big data. The paper also lists the technologies and tools to be applied for more efficient handling of the NHMA data.

Keywords: Russian National Hydrometeorological Archive, big data, data description language for hydrometeorological data, data processing technologies, data conversion technologies.

Cite this article: Peretyatko L. O. The National Hydrometeorological Archive as Big Data. Appropriate Technologies and Tools. *Russian Journal of Cybernetics*. 2022;3(4):98–101. DOI: 10.51790/2712-9942-2022-3-4-11.

Введение

Единый государственный фонд данных о состоянии окружающей среды, ее загрязнении (ЕГФД) является фондом, в рамках которого обеспечивается долговременное (в том числе и постоянное) хранение государственных информационных ресурсов Росгидромета и обслуживание потребителей. Единицами хранения в ЕГФД являются архивные документы [1], хранящиеся по правилам Росархива, и динамически изменяемые электронные массивы данных наблюдений (синтаксически и семантически обработанных первичных данных). Головной организацией, осуществляющей ведение ЕГФД, является Федеральное государственное бюджетное учреждение «Всероссийский научно-исследовательский институт гидрометеорологической информации — Мировой центр данных» (ФГБУ «ВНИИГМИ-МЦД», далее ВНИИГМИ-МЦД).

Согласно постановлению Правительства РФ от 21.12.1999 № 1410 ЕГФД представляет собой упорядоченную, постоянно пополняемую совокупность документированной информации о состоянии окружающей среды, ее загрязнении, получаемой в результате работы наблюдательной сети Росгидромета, а также информацию общего назначения и специализированную информацию в области гидрометеорологии и смежных с ней областях [2]. Данные в ЕГФД хранятся на фото- и бумажных носителях, а также в электронном виде на магнитных лентах и жестких дисках.

Данные первичных наблюдений являются уникальными и наиболее ценными для использования. По этой причине в ЕГФД данные первичных наблюдений хранятся в электронном виде с разной степенью обработки и организации [3]. Электронные первичные данные подразделяются на следующие виды: оперативные (используются для формирования оперативных прогнозов), режимные (полные данные наблюдений, обрабатываются с задержкой, используются для формирования режимно-справочных банков данных) и исторические (хранятся в неизменном виде, представляют собой непрерывные временные ряды данных по точке наблюдения). Данные первичных наблюдений являются уникальными и наиболее ценными.

В ЕГФД выполняется многослойное хранение электронных данных первичных наблюдений, поскольку, во-первых, необходимо хранить исходные данные первичных наблюдений, во-вторых этапы формирования данных повышенной достоверности подразумевают изменение исходных данных первичных наблюдений, получаемых из наблюдательной сети.

В первом слое находятся первичные данные в электронном виде, полученные с пунктов наблюдения. Данные этого слоя проходят минимальную обработку и хранятся в исходном виде, т. е. в виде, максимально приближенном к виду данных, получаемых из системы наблюдения, которые описываются иерархической моделью данных.

Во втором слое находятся электронные данные, полученные в результате синтаксической обработки и проверки на полноту исходных данных. Также в этом слое данных хранятся нормализованные архивы исходных данных, представленных в виде рядов наблюдений по точке наблюдения, и сгруппированные в режимно-справочные банки данных (РСБД) по видам наблюдений, выстроенных по точке наблюдения.

Технологии создания третьего слоя данных повышенной достоверности находятся в разработке. Этот слой данных будет формироваться на основе данных второго слоя путем устранения семантических ошибок.

Особенностью хранимых электронных данных в ЕГФД является формат хранения данных — ЯОД (язык описания гидрометеорологических данных). Данный формат позволяет сохранить иерархическую структуру данных, получаемых от сети наблюдения Росгидромета. Детальное описание формата представлено в статье [4].

Данные ЕГФД — это большие данные (“Big Data”)

В книге «Большие данные: Новый рубеж в производительности, инновациях и конкурентной борьбе» Джеймс Маньика совместно с соавторами дает следующее определение большим данным: «Большие данные относятся к наборам данных, размер которых превышает возможности обычных программных средств для сбора, хранения, управления и анализа баз данных» [5].

В качестве определяющих характеристик больших данных традиционно выделяют «три V»: *volume* — физический объем, *velocity* — скорость прироста, *variety* — разнообразие.

Изначально «три V» были выработаны в 2001 году вне контекста представлений о больших данных. В дальнейшем возникли различные интерпретации признака «трех V». Кроме того, появились

расширения изначальных признаков:

- «четыре V» (*volume, velocity, variety, veracity*) [6] — добавился признак *veracity* — достоверность;
- «пять V» (*volume, velocity, variety, viability, value*) [7] — были добавлены признаки *viability* — жизнеспособность и *value* — ценность;
- «семь V» (*volume, velocity, variety, veracity, value, variability, visualization*) [8] — этот набор признаков является комбинацией признаков «четыре V» и «пять V», где был исключен признак *viability (жизнеспособность)* и были добавлены признаки *variability* — переменчивость и *visualization* — визуализация.

Во всех случаях подчеркивается, что физический объем больших данных не является определяющей характеристикой. Важны и другие характеристики, существенные для задач обработки и анализа данных.

Рассмотрим данные ЕГФД с позиции следующих «шести V» признаков:

- *volume* — объем. На 31 декабря 2021 года в ЕГФД хранится информация на электронных носителях, которая записана в роботизированную библиотеку в объеме 2299,23 Гб (включая дубли), а также 2781536 единиц хранения документов на бумажном носителе информации и 865393 единицы хранения документов на фотоносителях [3];
- *velocity* — скорость прироста. Данные ЕГФД, в зависимости от вида информации, пополняются ежемесячно или ежегодно;
- *variety* — разнообразие. Объем информации ЕГФД Росгидромета, хранящейся во ВНИИГМИ-МЦД на электронных носителях, по состоянию на 31 декабря 2021 года по видам информации, без учета информации ИСЗ и зарубежной информации, составляет: метеорологическая информация (34,30 %), морская гидрометеорологическая (22,01 %), аэрологическая (19,36 %), гео- и гелиофизическая (11,83 %). Остальные виды информации составляют менее 5 % от объема фонда каждый [3];
- *veracity* — достоверность. Данные второго слоя ЕГФД проходят синтаксическую верификацию и проверку на полноту, что в совокупности обеспечивает повышенную достоверность этих данных;
- *value* — ценность. Данные ЕГФД представляют исключительную историческую, исследовательскую и практическую ценность для науки и органов власти;
- *viability* — жизнеспособность. Данные ЕГФД являются уникальными историческими данными.

На основе рассмотренных выше признаков и подтверждающей эти признаки информации можно однозначно классифицировать данные, хранящиеся в ЕГФД, как большие данные.

Примером больших данных также являются специализированные массивы для климатологических исследований, созданные на основе первичных электронных данных ЕГФД и выстроенные во временные ряды по точкам наблюдения (метеорологическим станциям). Специализированные массивы находятся в открытом доступе на web-сайте ВНИИГМИ-МЦД, с возможность получения выборок, предназначены для решения научно-прикладных задач в области исследования климата. Они регулярно пополняются, а обнаруживаемые в них ошибки исправляются [9].

Специализированные массивы составлены по следующим климатическим параметрам [10]:

- средняя месячная температура воздуха на станциях России;
- месячные суммы осадков на станциях России;
- среднемесячное давление воздуха на уровне станции;
- суммарная за месяц продолжительность солнечного сияния на станциях России;
- среднемесячное парциальное давление водяного пара на станциях России;
- суточная температура воздуха и количество осадков на станциях России и бывшего СССР.

Технологии и инструменты для работы с ЕГФД

Принимая во внимание особенности и специфичность формата хранения электронных данных ЕГФД, для эффективной работы с этими данными (обработки и распределения данных по слоям) предполагается создать следующие технологии и инструменты:

- технология и система формирования и хранения базы данных метаописаний для понимания фактического состояния файлов ЕГФД (полноты, наличия ошибок и их физического/логического расположения);
- технология и система поиска и выборки данных из больших массивов данных ЕГФД;
- комплекс программных средств взаимной конвертации данных в формате ЯОД в современные распространенные форматы (реляционные базы данных, HDF/NetCDF, XML и др.), реализующий тех-

нологию взаимной конвертации данных различной структуры, а также содержащий методы проверки адекватности выполненной конвертации, с поддержкой различных систем управления реляционными базами данных (СУБД), например, таких как PostgreSQL, MySQL, SQLite, Microsoft SQL Server. Первая версия системы взаимной конвертации описана в соответствующей статье [11];

- технология сокращения времени подготовки (преобразования) данных ЕГФД с использованием параллельных вычислений;
- технология формирования третьего слоя данных наблюдений повышенной достоверности;
- комплекс программных средств формирования третьего слоя данных повышенной достоверности, реализующий технологию формирования третьего слоя данных, включающий в себя систему формирования заявок от потребителей на исправление данных, с рассмотрением каждой заявки экспертами, определяющими обоснованность внесения изменений;
- модернизация инструментов составления режимно-справочных банков данных (второго слоя данных), а также пополнения специализированных массивов данных для климатологических исследований ЕГФД.

Перечисленные технологии и инструменты в данный момент находятся на разных стадиях разработки. Последующие статьи будут посвящены каждому разрабатываемому инструментальному средству в отдельности, а также их составляющим.

ЛИТЕРАТУРА

1. *РД 52.19.143 –2019 Перечень документов архивного фонда данных о состоянии окружающей среды, ее загрязнении.* Режим доступа: http://meteo.ru/egfd/145-metodical.
2. *Постановление Правительства РФ от 21.12.1999 № 1410 «О создании и ведении Единого государственного фонда данных о состоянии окружающей природной среды, ее загрязнении».* Режим доступа: https://www.meteorf.gov.ru/services/gosuslugi/29/doclist29/142/11985/.
3. *Ведение ЕГФД, архивация данных, обслуживание данными и информацией ЕГФД. Сведения о составе Госфонда Росгидромета.* Режим доступа: http://meteo.ru/egfd/142-about-egfd.
4. 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.
5. Manyika J. et al. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute; 2011.
6. Alba Diaz. *The Four V's of Big Data.* Режим доступа: https://opensistemas.com/en/the-four-vs-of-big-data.
7. Niel Biehn. *The Missing V's in Big Data: Viability and Value.* Режим доступа: https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/.
8. Eileen McNulty. *Understanding Big Data: the Seven V's.* Режим доступа: https://dataconomy.com/2014/05/seven-vs-big-data.
9. Булыгина О. Н., Коршунова Н. Н., Разуваев В. Н. Специализированные массивы данных для климатических исследований. *Труды ВНИИГМИ-МЦД*. 2014;177:136–148.
10. *Специализированные массивы*. Режим доступа: http://meteo.ru/data.
11. Перетятько Л. О., Кофтан Ю. Р. Система взаимной конвертации данных различной структуры для обслуживания потребителей ЕГФД. *Труды ВНИИГМИ-МЦД*. 2020;186:163–175.