Geophysical Data Aggregation Center IPE RAS

Stanislav D. Ivanov
Schmidt Institute of Physics
of the Earth RAS
Moscow, Russia
f0ma@ifz.ru

Igor M. Aleshin
Schmidt Institute of Physics
of the Earth RAS
Moscow, Russia
ima@ifz.ru

Vladimir N. Koryagin
Schmidt Institute of Physics
of the Earth RAS
Moscow, Russia
vlad@ifz.ru

Fedor V. Perederin
Schmidt Institute of Physics
of the Earth RAS
Moscow, Russia
crash@ifz.ru

Kirill I. Kholodkov
Schmidt Institute of Physics
of the Earth RAS
Moscow, Russia
keir@ifz.ru

Abstract

Today mostly data collection centers work with certain types of data. And there is no systematic approach to data storage and publication for users, who don't interwork with such centers. We propose a system of hardware, software and organizational procedures at Schmidt Institute of Physics of the Earth RAS towards to reduce the impact of this problem. There we created a centralized operator-friendly data management system that takes care of configuring all parts of data flow chain from instrument configuration to data publication.

The establishment of permanent geophysical monitoring missions fueled the need to store and process the specific data that the mission collects. Data collection, or aggregation, processing, and publishing centers became an essential part of the observational network e.g. IRIS DMC, Geofon (seismology), INTERMAGNET (geomagnetics), IGS (geodesy). Such centres are often tooled towards dealing with contiguous data acquisition. They observe their own data quality and station onboarding standards. Some are aimed for capturing data from temporary experiments when data was produced during a limited period of time e.g. Passcal. Both permanent and temporal data centres are specifically tuned for the corresponding measurement types.

Outside such large projects one can barely note any systematic approach to scientific data storage, processing, publication and distribution. Often those who acquired the data handle the storage and access for

the data. This raises a concern within the scientific community. for data publication, taxonomy and policy. Relatively low demand for such data makes establishment of problem-oriented data centers impractical. The consequences are: complicated and troublesome access or even complete data loss[Vin14].

Geophysical data can be split into two categories based on its definitiveness: preliminary and definitive. Preliminary data can undergo modifications and definitive cannot. Another way to break down into categories is considering how fast the data is published. This way we have archive data that is published after the research or fieldwork is finished, current data that is published during the research process, and real-time data that is published immediately after it is extracted from the instrument. Often archive data is definitive and current and real-time data is preliminary.

We propose the development of system of hardware, software and organizational procedures that would enable the unification of data collection and storage irrelevantly of how fast the data is collected, whether it's preliminary or definitive. A scientific institution is a good place to mount such an effort so we advanced this initiative at Schmidt Institute of Physics of the Earth RAS[Iva18]. Several engineering actions and organizational steps should be made. Among these actions are datacenter maintenance, publication, backup, access control, some common archive-related actions like archive compilation, modifications, additions. In this paper we'd focus on these actions and leave organizational steps for future publication.

IPE RAS already has the required infrastructure for deployment of such centre. It is based on virtual infrastructure back with data storage system that is connected to dedicated autonomous system with fast peer connections to major telecommunication organization. Thus, the backing infrastructure is considered fault-tolerant.

Data sources often use different protocols for access. Some use file-based protocols: FTP, HTTP, SFTP and even encode files in SMTP. Others expose relational and nonrelational (e.g. InfluxDB) database endpoints, or message queuing protocols (MQTT, Apache Kafka), or specialized streaming protocols (Seedlink). Select protocols implement some sort of access control and encryption.

When data is being manipulated manually the selection of protocol and access control is solely a user's decision. By contrast, automated real-time data collection services must implement best suitable protocol and means of access control, preferably on low network stack level, e.g. VPN.

For instance, for time-series data we utilize miniSEED format, mainly because it has lots of tools for collection, transfer, storage and processing. This format is used for collecting data from instruments in experimental observatory "Biryulevo" (Moscow oblast); structure health monitoring system [Ale19] and tiltmeters [Ale18a] situated at IPE RAS main building (Moscow); geomagnetic observatory "Klimovskaya" (Arkhangelsk oblast). In order to transfer miniSEED no additional encoding is required saving resources and lowering latency. Observers that are connected to retransmission node get new data in 5 seconds at the latest. Currently all mentioned observatories are equipped with data collection system based on single board computers (RaspberryPI)[Ale18b]. This solution is proven universal and economic solution for a wide range of data collection tasks.

Additionally, the centre perform collection of retransmitted data from heliophysical space vehicle "Elektro-L N2" (GOMS-3). The data is initially acquired at Fedorov Institute of Applied Physics with autonomous data collection system - a satellite receiver system and then gets retransmitted with Seedlink protocol to the IPE RAS Data Aggregation Center. This data includes particle counter data, x-ray irradiation meters and galactic ray event counters.

The center also performs archiving tasks. *Ad exemplum* the data collected from "Biryulevo" observatory for periods of 2002-2016 and 2011-2017 is available for access in miniSEED format. The center provides coverage plots for achieved data (Figure 1).
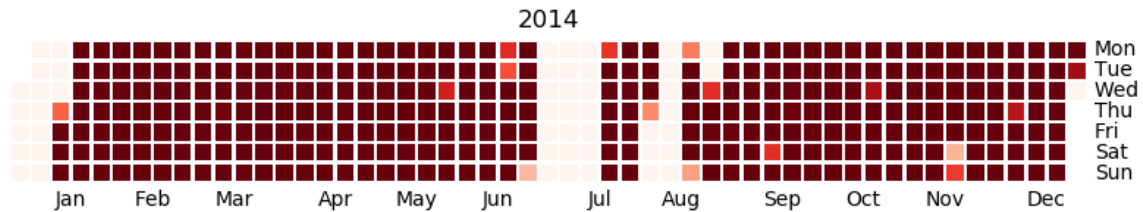

Figure 1: Coverage plot for year 2014 at "Biryulevo" observatory. Color intensity denotes amount of data for the particular day.

If properly set, the amount of stored data does not pose any serious technological and organizational challenge. But the increase of numbers of stored data channels does. To facilitate the management the center shall provide means of assisted configuration of both center-side and instrument-side. Here, at IPE RAS, we are gradually erecting a centralized operator-friendly data site management system that takes care of configuring all parts of data flow chain starting with instrument configuration. The systems relies on relational DBMS and Ansible deployment and maintenance system. The information inside this database is used to create automation scripts for Ansible and configuration files that drive the entire system. The system is capable of both initial configuration and reconfiguration tasks.

Visualization greatly improves the monitoring and accessibility of available data. We utilize specialized database management system InfluxDB acting as intermediate database and feeding data to Grafana (Figure 2) visualization framework. This software stack is used to display real-time data, however, currently, these two components lack centralized automated configuration which is a planned feature update.
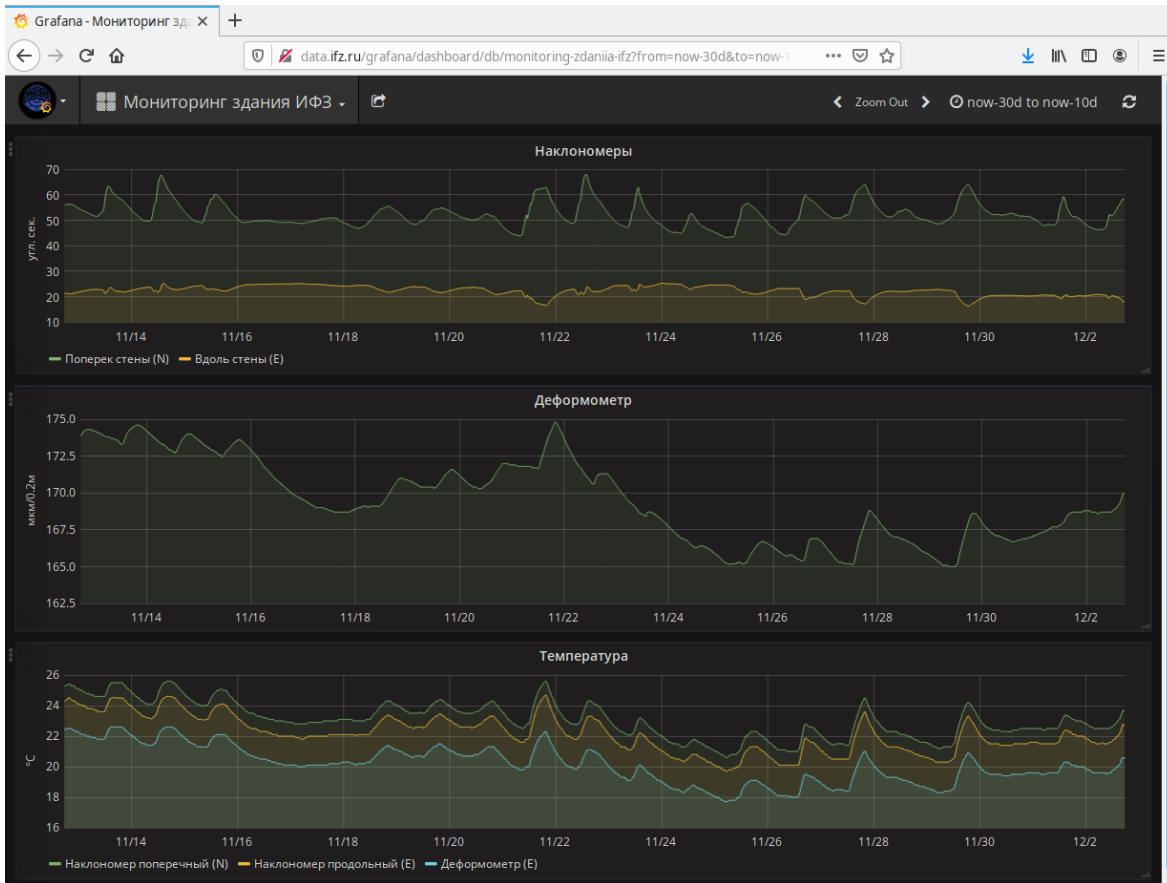
Figure 2: Structural health monitoring instruments show in real-time

The centre also implements way to monitor both data sources and internal components as well by using Shinken open-source monitoring software. The implementation watches over hardware and software vital indicators and provides extensive information should any fault arise. Currently system does not check the values of the incoming data, only its presence. Data sanity check is also a planned feature update.

Another activity of the center is real-time data acquisition from field expeditions. In 2018 we've evaluated the portable high-frequency GNSS data collection system [Per18]. The data acquired from Javad Alpha2 GNSS receiver at 10Hz was collected with portable system and transferred with cellular data connection to the center in realtime. The portable system is also based on single board Raspberry Pi computer. The transmission was performed from a moving road vehicle on Moscow-Arkhangelsk (Fig. 3) and Moscow-Kandalaksha routes.
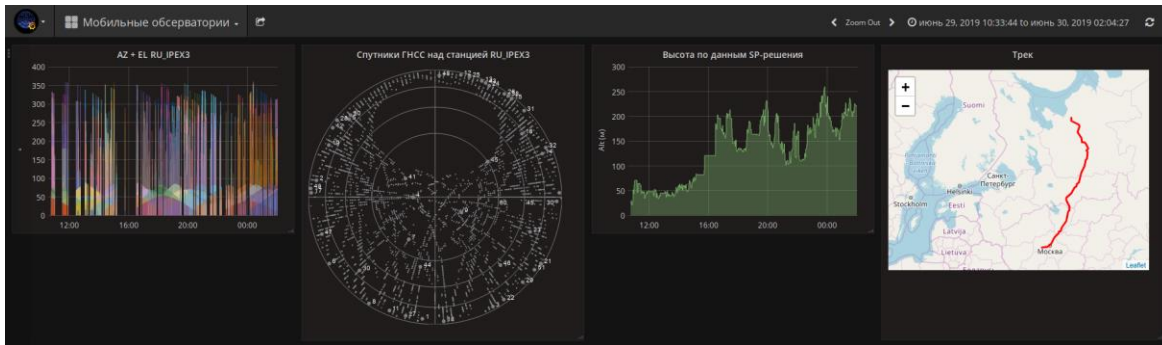
Figure 3: The GNSS measurement data for Moscow-Arkhangelsk experiment is shown in Grafana interface.

References

[Vin14]   Vines T. H. et al. The availability of research data declines rapidly with article age //Current biology. – 2014. – T. 24. – №. 1. – C. 94-97.

[Iva18]   Ivanov S.D., Aleshin I.M., Kholodkov K.I., Perederin F.V. Sistema upravleniya tsentrom agregatsii dannykh IFZ RAN // Nauchnaya konferentsiya molodykh uchenykh i aspirantov IFZ RAN, 23-24 aprelya 2018. Tezisy dokladov. Moscow, 2018.

[Ale19]   Aleshin I. M., Ivanov S. D., Kholodkov K. I. et al. Remote real-time structure health monitoring with mini-smik // Seismic Instruments. — 2019. — Vol. 55, no. 5. — P. 589–595.

[Ale18a]   Aleshin I. M., Ivanov S. D., Koryagin V. N. et al. Online publication of tiltmeter data based on the seedlink protocol // Seismic Instruments. — 2018. — Vol. 54, no. 3. — P. 254–259.

[Ale18b]   Aleshin I.M., Getmanov V.G., Grudnev A.A., Dobrovol'skii M.N., Kholodkov K.I., Koryagin V.N., Krasnoperov R.I., Kudin D.V., Solov'ev A.A., Ivanov S.D. Kompaktnoe energoeffektivnoe ustroistvo sbora i operativnoi peredachi geomagnitnykh dannykh // II Vserossiiskaya nauchno-prakticheskaya konferentsiya «Nauchnoe priborostroenie – sovremennoe sostoyanie i perspektivy razvitiya» 4-7 iyunya 2018 g.Kazan, Russia

[Per18]   Perederin F. V., Aleshin I. M., Ivanov S. D. i dr. Portativnyi kompleks registratsii signalov GNSS s vysokoi chastotoi oprosa: polevye ispytaniya i perspektivy primeneniya // Nauka i tekhnologicheskie razrabotki. — 2018. — T. 97, № 4. — S. 28–40.