Data integration of the Russian segment of the integrated Arctic observing system

Kristina Belova, RIHMI-WDC (Russian Federation), christy@meteo.ru
Nikolai Mikhailov, RIHMI-WDC (Russian Federation), nodc@meteo.ru

Data are the raw characteristics of information processed and control by computers. Information is the presentation of data in a context that has business value to people. Information is needed by scientific, design and industrial enterprises, government agencies and other participants in maritime activities. Information is especially critical for activities in the Arctic. It is necessary to integrate heterogeneous and distributed data from the Arctic observation networks, databases and systems, to produce relevant information, and to disseminate this information in support regional maritime activities. This problem is being addressed under the INTAROS Project of the HORIZON-2020 Program and the Research Programme of the Ministry of Science and Higher Education of Russia through the creation of an integrated Arctic observation system (iAOS).

Russian segment and node iAOS node

The Russian segment of iAOS includes a network of the Arctic data sources, which is the part of the Unified State System of Information on the World Ocean (ESIMO, www.esimo.ru), and the Russian iAOS node — a software and hardware system that contains subsystems for data integration, for data processing and for provision of services. The ESIMO software and other means have been used as the infrastructure of the node. A multi-component Java-oriented development and operation environment is used as the basic software. Applied software consists of applications and SOAP/ RESTful web services implementing business processes and visual user interfaces of the node subsystems. In the context of computing hardware, a node is a cluster of servers containing virtual machines with applied software components.

Data integration of the Russian segment of iAOS

There are three main approaches to data integration: (1) Consolidation, when data are extracted from multiple sources and integrated into a centralized data warehouse. The technology is called ETL (Extract-Transform-Load); (2) Federalization based on the use of brokers - software applications for transparent access to distributed and heterogeneous data sources, retrieving and transforming data into a common data model, delivering data to a given point in an accepted exchange format. An example of this approach is enterprise information integration (EII) technology; (3) Data distribution based on the transfer of data from the data source to the receiving application. Enterprise application integration (EAI) technology is an example of this approach. Data integration of the Russian segment of iAOS is based on a combination of EAI technology and ETL technology. The data integration subsystem (Figure) of the node includes the following components:
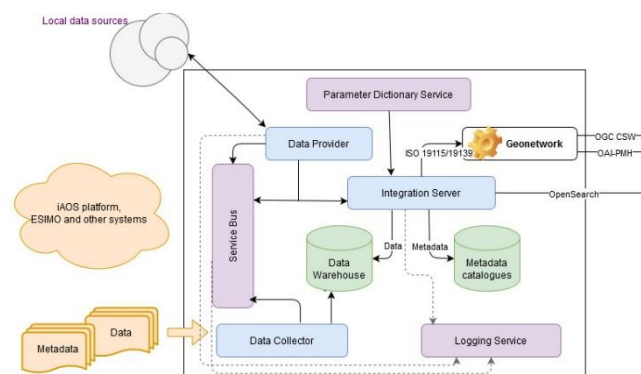

Figure 1 : Structure of the data integration subsystem of the RUSSIAN node of iAOS

• Data Collector, which performs access to data and metadata of the iSNA platform and ESIMO, their transformation if required and inclusion of data messages and flows into the node processes;

• Data Provider is an integration broker. It performs access to data sources using software wrappers, extracting and transforming data into the format of a transport data file (formats supported are NetCDF (2, 3), CSV), and also forms the metadata in the ISO 19115 standard / 19139 and interaction with the Integration Server;

• Integration Server is the data integration adjuster. It maintains a registry of data providers and data sources, can add new objects or delete existing ones. It also collects data and metadata provided by the Data Providers, loads them into the file Data Warehouse, and provides data and metadata for the implementation of internal business processes of the node and for external usage and systems;

• Data Warehouse, implemented as a cache system of transport data files and metadata files for subsequent sharing;

• Service Bus in combination with Logging Service implements uniform, flexible and documented communication between the components listed above. The Parameters Dictionary Service is designed to present information about parameters to consumers in a unified form understandable to humans.

As a result of data integration, the designated data of sources (named as information resources of the node) and metadata are loaded into the file Data Warehouse and the PostgreSQL integrated database. Data Warehouses are replenished automatically and their contents are used to produce new products and services. Note that over 30 percent of the node's information resources are updated in real time, for example, hydrometeorological observation data updated every 3 hours or forecast information updated twice a day. For their integration, special metadata are used that set the schedule and rules for processing of events and organizing the exchange of messages, metadata and data between the Data Providers and the Integration Server. The Node Service Bus plays a significant role in maintaining the node's Data Warehouses in real time. Metadata are an important element of data integration. Several consolidated metadata repositories are used. Examples are the catalog of information resources of the node in the ISO 19115 metadata standard, WMO Core Profile 1.3 and CKAN, available via the OAI-PMH protocols, OpenSearch, the OGC CSW geoservices catalog, the SOAP / RESTful API specifications for web services. The catalogs are available at http: //core.intaros.meteo.ru:8081/geonetwork/ and http://opendata.intaros.meteo.ru.

Development prospects

Data integration of the Russian iAOS segment is based on a combination of consolidation (ETL technology) and data federation (EAI technology) of sources, as well as the widespread use of metadata and service bus tools. This makes it possible to perceive heterogeneous and physically distributed data as a single information space. It is planned to use the means and technologies of the node as a universal node of a distributed system in the context of the tasks of iAOS, ESIMO and the WMO Arctic Distributed Climate Center. The development of data integration tools is associated with the expansion of the use of EAI technology, as well as the study of the capabilities of the EII technology and its testing.