Standardization of Forms and Tools for Inter-machine Interaction in Exchange of Hydrometeorological Data

E D Viazilov, D A Melnikov and A S Mikheev

RIHMI-WDC, 6, Koroleva, 249035, Obninsk, Kaluga Region, Russia, http://meteo.ru.

E-mail: vjaz@meteo.ru

Abstract. The necessity of organizing a machine-to-machine data exchange in the Earth sciences for their subsequent interdisciplinary analysis is shown. The barriers that hinder interaction between various information systems are indicated. The requirements, FAIR, for datasets and the principles, TRUST, for data repositories that must be met to effectively exchange data are presented. An ocean data exchange scheme is described, which is used within a pan-European project, SeaDataCloud, and can be proposed for data exchange with the World Data System. International formats of data exchange services and data visualization tools are briefly described. Examples of integration of distributed and heterogeneous data, metadata bases, and information about classifiers used in ESIMO (http://esimo.ru) are considered. An idea of standardizing a format for storing and exchanging data in the form of time series is proposed. In addition to standardization of formats, it is necessary to standardize the common codes and classifiers used in the data exchange and the parameter names from a unified vocabulary. Some prospects of inter-machine interaction are the creation of tools of data exchange with the existing international and national systems and of conveyor data processing ranging from hydrometeorological data collection to their use in business processes.

1. Introduction

In recent years, an urgent need has emerged for organizing interdisciplinary data analysis in the Earth sciences. For interdisciplinary data analysis, it is necessary to organize exchange of data between different systems at the national and international levels. In the field of hydrometeorology, international exchange of data has always been an important aspect of replenishing the state fund of data on hydrometeorology and environmental monitoring (State Fund). Initially data exchange was carried out on paper (tables, reports, punched cards), then on magnetic media (magnetic tapes, magnetic cartridges, floppy disks), further on compact disks (CD-ROM, DVD), by e-mail (Figure 1). And now it is required to exchange data between software applications located on different Internet nodes (servers). The World Data Centers (WDC) system of the International Council of Scientific Unions (ICSU) is currently being transformed into the World Data System (WDS), where the main mode of exchange is machine-to-machine communication between information systems.

There are already examples of such interaction within the pan-European projects EUDAT (https://www.eudat.eu/), SeaDataCloud (http://www.seadatanet.org), and EMODNet Ingestion (https://www.emodnet-ingestion.eu/). In Russia, there are the Gosuslugi portal...
From the point of view of the users, there are the following barriers to obtaining data:

- it takes calendar time to discover data;
- there is great fragmentation in many types of data;
- users are unable to access the data, because of no permission to use them;
- there are restrictions on the use of the data, for example, "for research purposes only";
- difficulty of combining data from different sources;
- there is no information about the origin of the data;
- measurement accuracy, quality, and completeness of data are not always known;
- the spatial and temporal resolution of the measurement data is insufficient.

Implementation of machine-to-machine interaction between various information systems requires standardization of metadata, codes and classifiers being used, parameter names, formats for collecting, exchanging, and storing data, forms for issuing information products, data access services, user interfaces, developing requirements to data exchange formats, increasing relevance and availability of data, integration of distributed and heterogeneous data. Only if these barriers are removed, effective interdisciplinary data analysis may be organized.

2. Requirements to data exchange

Requirements to data exchange are as follows:

- organizations involved in the exchange of data must be accredited, for example, as is now done for the ICSU WDC system [1].
- exchanged databases (sets) of data must be certified;
- data must be transmitted in open or encrypted form, or with an electronic digital signature;
- data formats should provide the ability of data transfer (exchange) and be known and widespread;
- formats must be cross-platform;
- any system that recorded the file must subsequently read it and restore all information;
- information carriers must have technological compatibility at the read-write level (any system must read a file written by another system, and at the same time receive all information that it is able to process);
- any system will be able to read the file and restore all information from it if this file was originally written by the system itself, but was later changed by another system and saved again (the property of saving other people's data);
- the exchanged datasets should be documented and have metadata, including a semantic description of the data, information about platforms, tools, data quality, and information about the data exchange formats;
- all records must have key attributes;
- the list of parameters should correspond to the composition of observations defined by the current manuals and guidelines;
- parameters of observations and generalizations should have a quality attribute.

The main characteristics of data quality are defined in the ISO 9126-1991 standard [2]. These are volume, completeness, relevance, retrospective of data, validity, reliability, accuracy of data attributes, etc. The exchanged datasets must meet the requirements of Findable, Accessibility, Interoperability, and Reuse (FAIR) [3].

Findable data:
- metadata (MD) and data are assigned a globally unique and permanent identification (ID);
- data is described by using several MD objects;
- MDs clearly and explicitly include the ID of the data described;
- MDs and data are recorded in a searchable Internet resource.
Accessibility of data and MD:
- MD and data can be obtained by their IDs by using a standardized transferring protocol, which is open, free and universal, allows for authentication and authorization;
- MDs must be available even if data is no longer available.

Interoperability of date sets:
- MD and data must use vocabularies that also must follow FAIR requirements;
- MD and data must include qualified references to other MD objects;
- MD and data must be in accordance with the existing standards on used classifiers, formats of storage attributes (date and time, place coordinates).

Reuse of data:
- MD and data must be well described by a variety of attributes;
- MD and data must have a license for use;
- MD and data must associate with detailed descriptions of data sources, information of lifecycle of data (history of converting of data sets).

For repositories (data stores) in the WDCs system, the TRUST – principles are used [1]:
- transparency - repositories must be transparent about specific data repository services;
- responsibility - repositories must be responsible for ensuring the authenticity and integrity of data stores, and the reliability and consistency of maintenance by data;
- user focus repositories must ensure a data governance compliance with the users’ expectations;
- sustainability - repositories must maintain services and preserve data in the long term;
- technology - repositories must provide an infrastructure and capabilities to support secure, consistent, and reliable services.

The ocean data exchange scheme, which is partially used within the pan-European SeaDataCloud project and can be proposed for exchange with the WDS, is shown in Figure 2. The Ocean Data Portal (International Oceanographic Data Exchange of Intergovernmental Oceanographic Commission, http://www.oceandataportal.org/) should become the main element of the oceanographic data exchange system, which should assimilate national data from pan-European systems and deliver them to various applications.

![Image](image_url)

**Figure 2.** Developable ocean data exchange.
3. Formats of data exchange

Formats, depending on their application, can be divided into formats for exchange, collection, storage of initial, inverted, and calculated data. International formats (ODV, NetCDF), [https://www.seadatanet.org/Standards/Data-Transport-Formats](https://www.seadatanet.org/Standards/Data-Transport-Formats); (GRID, GRIB, BUFR) [4]; or eXtensible Markup Language (XML) and JavaScript Object Notation (JSON / GeoJSON) can be used for transport data exchange formats.

The collection formats are used to present observations. They should contain all observable data and information related to methods of obtaining, data sources, and measuring systems. The international formats are used to collect real time data by the global telecommunication system. For hydrometeorological data collected in a deferred mode, universal self-describing formats are proposed (in the first lines of the file, a list of parameter names is given, and then the data itself is written).

The requirement, which is imposed on the data storage format is the presence of key metadata attributes in the format of data (for example, name of the country, organization of the ship owner, name of the vessel for ocean data). The State Fund for storage uses the Language for the description of hydrometeorological data [5].

The main requirement for inverted data storage formats is to bring them as close as possible to further processing. Other requirements for such formats are extreme simplicity of their presentation, high access speed, and the availability of information about the sources of these data and the history of their transformations. Data can be presented as a point, profile, grid, object files catalog. The structure of these data should be such that data from these data sets can be used by most data bases management systems, i.e. they must have a simple structure and a fixed record length.

For storing, interpolated values of parameter formats of the "Vertical" type are distinguished: the results of interpolation of parameters to standard horizons (heights) for vertical profiles; "Grid" - the results of spatial interpolation of data measured at fixed points in space into the nodes of a regular grid. The interpolated data at the nodes of a regular grid are presented in the form of three tables (information about the grid, information about each field and data matrix - latitude, longitude, value) [6].

The data on time series are drawn up in the form of two tables - information about the time series, values of parameters for the time series (Table 1). The data catalogue about object files with text or graphic information (documents, pictures, images, photos, video, sound) must include the date, time, type and name of the platform, store links to files with unstructured data.

![Table 1](https://example.com/table1.png)

| ID_time series | Lat | Long | Begin date | End date | Time resolution | Parameter | Author |
|----------------|-----|------|------------|----------|-----------------|-----------|--------|

4. Data integration

For distributed, heterogeneous data to be available for use from one or more Internet nodes, through one software interface, so that data can be easily exchanged, delivered to any consumer by automatic means, they must be integrated before use. A description of the integration system for ESIMO is given in [7]. A technological scheme of integration, application processing, and use of data is shown in Figure 3.

The advantages of such a data processing scheme after their integration is the ability to quickly configure applications for new data, processing tools, necessary parameters, calculating statistical characteristics, indicating parameter values by threshold values of their dangerous level [http://portal.intaros.meteo.ru/portal/intaros/services/climate_monitoring](http://portal.intaros.meteo.ru/portal/intaros/services/climate_monitoring).
5. Data access services

The results of processing of the observed and processed data can be presented on the site by using stand-alone applications, map services, and web services.

A stand-alone application is a type of product that has a permanent link on the web and involves interaction between the user and the visual interface. It needs to standardize request types and their parameters, styles, user interface elements, legends, naming and forms of significant interface elements.

Service is a type of product characterized by the presence of a permanent link in the network and interaction between two software components. The format of the link is subject to standardization; allowed request types and their parameters; legend for images, naming and storage formats of attribute values should be in accordance with the accepted standards.

A web service is identified by a web address; it is a program with a standardized interface. Web services can interact with each other and with third-party applications through messages based on registers and protocols, Universal Description Discovery & Integration (UDDI), Web Services Description Language (WSDL), Simple Object Access Protocol (SOAP), XML-Remote Procedure Call (XML-RPC), thematic schemes of XML.

An applied programming interface (API) is a set of ready-made classes, procedures, functions, structures, and constants provided by an application (library, service) for use in external software products. Access to databases can be organized by API.

Map services are standard protocols for serving spatial data created by a map server. The main types of map services are Open Geospatial Consortium standards Web map service (WMS), Web feature service (WFS), Web coverage service (WCS); Google standard - Keyhole Markup Language (KML). In addition, the following ones are used to represent spatial information:
- GeoTIFF format - an open format for representing raster data in the TIFF format together with georeferencing metadata;
GeoPDF format is an enhanced PDF format containing spatial information that can be viewed by Adobe Reader. GeoPDF allows one to compile multiple layers in one PDF file, insert coordinates, hyperlinks; add notes to maps.

6. Thematic metadata

When exchanging data, metadata should be widely used, both for describing data and for information products obtained on their basis. In addition, some information products may require additional metadata objects. For example, information on observation platforms is necessary to describe most of the observed data, as they usually contain only the main identifying attributes (synoptic index, call sign of the vessel, buoy, etc.). The metadata objects that are needed when processing data are information about sets and databases, organizations - data sources, custodians and providers of data, of projects and programs, within the framework of which certain data were obtained, of observation platforms - research vessels, drifting surface or moored buoys, hydrometeorological stations and posts, satellites; devices; expeditions of research vessels; models and calculation methods, etc.

Metadata objects can be described by using the XML extensible markup language. To do this, it is necessary to develop an XML schema that includes all parameters of metadata and information products, as well as to develop universal forms for describing and presenting information products.

All metadata objects are visualized according to the standard scheme (for example, http://esimo.ru/portal/portal/esimo-user/metadata). On the left, there is a possibility of selection by synchronized attribute values. Above the table there is an opportunity to set a keyword search. Only the main attributes are placed in the output table. To obtain complete information about an instance of an MD object, press the “i” sign in the first column of the table, after which all attribute values will be revealed. Individual attributes during rendering can disable, as well as export of the table attributes presented on the screen. To obtain a copy of the MD, the selected information can be exported in a Comma-Separated Values (CSV) format.

All MD objects are interconnected (Figure 4). For each developed technology, its own composition of used or generated data sets and databases is formed. Each data set or database must have its own instances of metadata objects (information about data sets, organizations, formats, observational platforms, methods, and software tools, regulatory and methodological documents). Such a model for describing metadata objects allows one to get all reference information about data and other MD objects much faster. There remains a separate search for each MD object to select instances of these objects for various conditions. This method of organizing MDs and access to them is implemented in ESIMO (http://www.esimo.net/meta/).

![Figure 4. Model of description of metadata objects.](image-url)
7. General codes and classifiers

When exchanging data, it is necessary to use the highest level of standardization of classifiers – international ones, and only in the absence of such classifiers can national or industry classifiers be used. To organize and search for classifiers that can be used in the exchange of data, databases of common codes and classifiers should be created. For example, ESIMO has a database "Classifiers" (http://portal.esimo.net/portal/portal/tech/), which contains more than 500 classifiers; there are tools for their search and viewing.

For the interaction of information systems operating in different countries and organizations, it is necessary to have the same understanding of the local names of the attributes included in the integrated database from different local systems. That is, mapping from local names to system-wide codes is required (Table 2). Currently, the most advanced vocabulary of parameters has been created at the British Oceanographic Data Center (https://www.bodc.ac.uk/resources/vocabularies/), which includes more than 35 thousand names. To unify the names of parameters in various information resources, ESIMO uses the Unified Dictionary of Parameters (http://esimo.ru/portal/auth/portal/esimo-user/metadata). Here it can find information about 2500 parameters. The parameter code is represented according to the following scheme: ANNNN_SS_K, where, A - section of the parameter dictionary (P - basic parameters, M - metadata attributes or R - parameters characterizing the content of data for object file; NNNN - parameter numeric code; SS - numerical code reflecting the name of the statistical characteristic; K - parameter properties (indicator, data quality attribute, used method for determining parameters).

Table 2. Example of Mapping Parameter Names.

| Attribute | Name                        | Parameter code |
|-----------|-----------------------------|----------------|
| U_Id      | Identification              | M4000          |
| O_Name    | Organization: name          | M4107          |
| Lat       | Latitude                    | M4311          |
| Ta        | Temperature of air: measured| P0001_00       |

8. Data visualization

For universal visualization of integrated data using the example of ESIMO, the following are used:
- map, graph, table - regular products are visualized for each information resource by default;
- interactive map - a combination of thematic and basic layers based on the use of cartographic services, the interactive map has the ability to get information about the values of parameters at a point, as well as a graph of the time course of parameters;
- directory of object files - allows one to quickly find the storage address of the required instance (file) and, if necessary, download it to a local computer;
- finished information products in the form of regularly updated ("live") web pages - represents a dynamic page with a permanent address on the Internet for each information resource;
- matrix of information products - prepared for several geographic regions (for example, individual seas), sets of parameters and generalizations in the form of a triad - observations, forecast, climate;
- the results of the indication of parameter values - exceeding the threshold values of the parameters (local, national, or climatic values);
- combining the observed and forecast information on one graph;
- data export - for structured data formats the following are used: NetCDF, ASCII flat structure files, CSV, XML, Excel, PDF files; for unstructured data - graphs and maps in the form of files such as gif, tif;
- presentation of information resources in an applied task - a specially programmed interface for analytics of different data types, disciplines.
9. Conclusions
The FAIR requirements for data and TRUST- principles for data warehouses were presented. For a one-time exchange of hydrometeorological data, the NetCdf format and for permanent regular exchange the XML, JSON formats are used. For the first time in hydrometeorology, an idea of standardizing formats for storing and exchanging data in the form of time series and grid data was discussed. A mechanism of standardizing the common codes and classifiers used in data exchange and the parameter names by a unified vocabulary was realized.

Based on the above-proposed approaches, data exchange was implemented using the web services for automatic delivery of information resources of ESIMO to the users; and to provide information about observing stations for the Spatial Planning System of the Ministry of Economic Development of Russia. A service was designed to provide information about the Russian meteorological stations of international exchange for the OSCAR system (World Meteorological Organization), https://oscar.wmo.int/surface/index.html#/.

Some prospects are the development of tools of machine-to-machine interaction with the existing international and national systems and conveyor data processing for all processes ranging from data collection to their use in business processes.

Acknowledgments
This work was supported by the Ministry of Science and Higher Education of the Russian Federation under project RFMEFI61618X0103.

References
[1] Dawei Lin et al 2020 The TRUST Principles for digital repositories. Scientific Data 7(1) DOI: 10.1038/s41597-020-0486-7.
[2] ISO 9126 1991 Software engineering — Product quality ISO/IEC JTC 1/SC 7 13 p
[3] Wilkinson M D et al 2016 The FAIR Guiding Principles for scientific data management and stewardship Sci. Data 3 September 160018
[4] Jean Claude Bergès 2002 Support of WMO binary format (BUFR and GRIB) Proceedings of the Open source GIS − GRASS user conference (Trento, Italy, 11−13 September 2002) pp 1-6 http://www.ing.unitn.it/~grass/conferences/GRASS2002/proceedings/proceedings/pdfs/Berges_Jean_Claude.pdf.
[5] Veselov V M 1995 Language for the description of hydrometeorological data for IBM PC (Obninsk: RIHMI-WDC) proceedings 160 pp 41 – 54
[6] Viazilov E D 2018 Creation and use of databases (Germany: Palmarium Academic Publishing) ISBN: 978-620-2-38198-7 545 p
[7] Viazilov E D, Mikhailov N N 2013 Data integration for marine environment and marine activities Infrastructure of satellite geoinformation resources and their integration Collection of scientific articles ed Doctor of Technical Sciences M A Popov and D Sc E B Kudashev (Kiev: Carbon-Service) pp 174 -81