Standardization of forms and tools for inter-machine interaction in the exchange of hydrometeorological data

Evgenii Viazilov, Denis Melnikov, Alexander Mikheev
(RIHMI-WDC, e-mail: vjaz@meteo.ru)

Outline

- Introduction

- Data exchange

- FAIR Dataset Requirements and TRUST Principles for Repositories

- Storage and exchange formats

- Product types and data delivery services

- Export (visualizations)

Introduction

From the point of view of users, there are barriers to obtaining data:

–   There is great fragmentation in many types of data.
–   It takes calendar time to discover data
–   Unable to access the data, because no permission to use them
–   There are restrictions on the use of the data, "for research purposes only"
–   Difficulty 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.

■   These barriers affect the relevance of the data.
■   We need metadata, accessible data.
■   Integration of distributed and heterogeneous data is required.
■   Requires regular communication between applications on different servers
■   There is a need for multi-disciplinary data analysis.

Data exchange

| 1957–1973 | 1974–1990 | 1990–2000 | 1990–2016 | 2020 |

- WDS
- Data exchange within the framework of national and international projects
- GTS data collection
- Using data in applications on different Internet sites
- Data exchange between individual organizations
- Automated delivery of data to departmental information systems
- For examples: EUDAT, SeaDataNet. Gosuslugi, ESIMO

Developable ocean data exchange
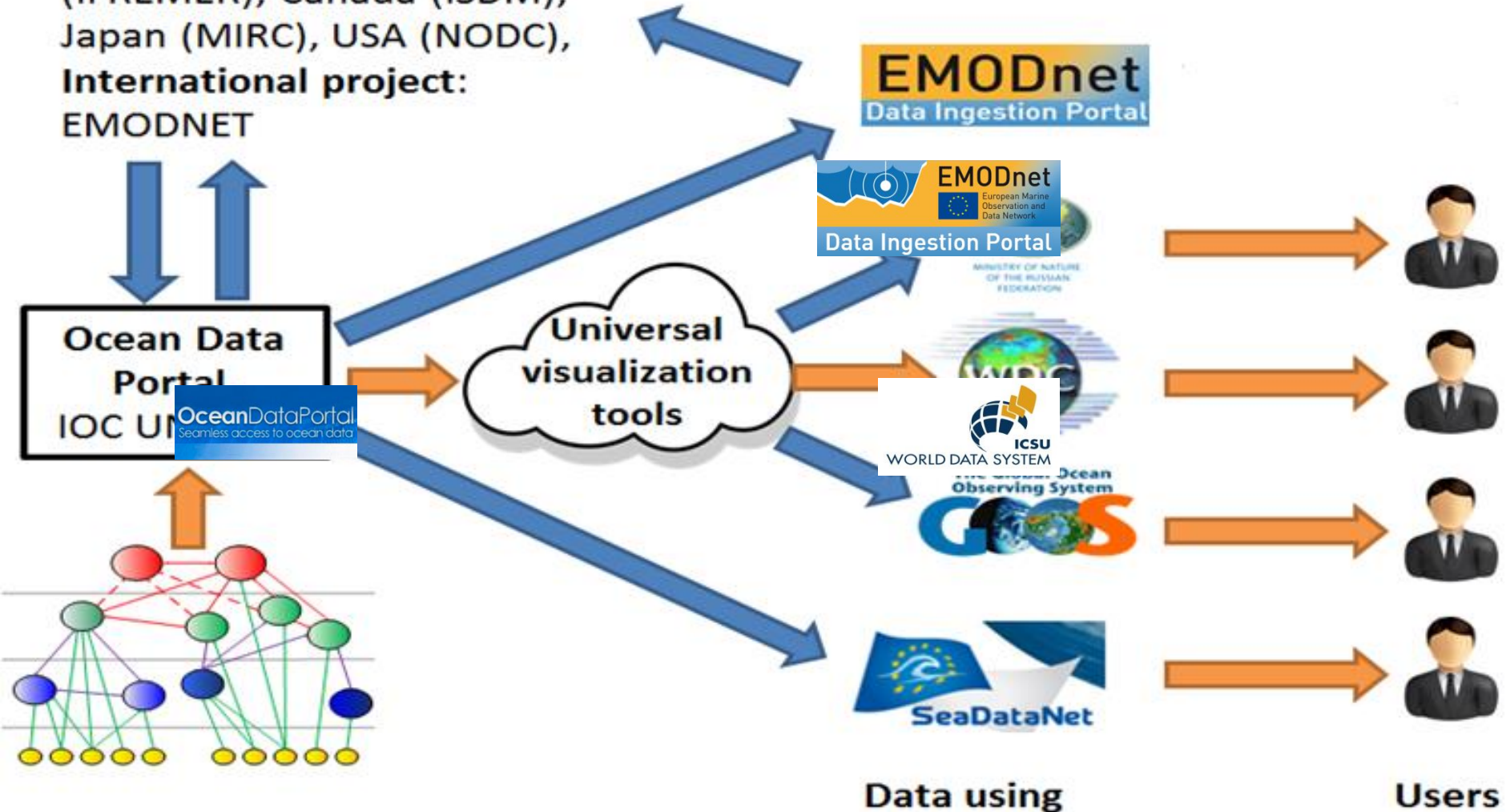
FAIR data requirements (go-fair.org)

Findable:
1. MD and data are assigned a globally unique and permanent ID
2. Data is described using rich metadata
3. The MDs clearly and explicitly include the IDs of the data they describe
4. The MDs and data are recorded in a searchable resource

Accessible:
1. MD and data can be obtained by their ID using a standardized communication protocol, which is open, free and universal, allows for authentication and authorization
2. MD is available even if the data is no longer available

Interoperable:
1. MD and data use vocabularies that also follow FAIR
2. MD and data include qualified references to other MD
3. MD and data are in accordance with existing standards

Reuse:
1. MD and data are richly described by a variety of attributes
2. MD and data are licensed to use
3. MD and data are associated with a detailed origin of the data (Lifecycle)

TRUST – principles (WDS)

1) **Transparency**. Be transparent about specific data repository services.

2) **Responsibility**. Be responsible for ensuring the authenticity and integrity of data stores, and for the reliability and consistency of data maintenance.

3) **User Focus**. Ensure compliance with data governance and users expectations.

4) **Sustainability**. Maintaining services and preserving data in the long term.

5) **Technology**. To provide the infrastructure and capabilities to support secure, consistent, and reliable services.

Data sovereignty means that data holders (researchers or organizations) should have the right to decide what data they share, with whom, and on what terms.

Development of FAIR requirements and TRUST principles

- **Distributed heterogeneous data should**:
Accompanied by persistent IDs, metadata and documentations for discovery, citation, and reuse documentation, and an indicator of their maturity (an extension of the **Data cite** metadata schema).

- Include data services, metadata specifications, guidelines, recommendations, data management policies and plans.

- Accompanied by means of interaction for data exchange - formats, metadata, services and infrastructure.

- Be integrated for analysis.

- Stored in trust accredited data centers.

- Certified for FAIR requirements.

- Be evaluated by metrics and indicators of data state (completeness, relevance, reliability, response time, etc.).

Formats for collection, storage, exchange

- **Formats** for International (WDC), interdepartmental (Gosuslugi, ESIMO) exchange and interaction between applications

- **Collection formats** for presenting measurement results - completeness of reflection of object properties. It should provide all observable and related information about methods of obtaining, data sources, measuring systems.

- **Data storage formats** - the main requirement is maximum data packing in order to reduce the number of stored media, the presence of metadata attributes in the format.

- **Inverted data formats** - storing data that has passed the stage of structural transformations. The requirement is to bring it closer to data processing, simple structure, high access speed.

- **Formats for storing data processing results**. The formats such as "Time series", "Grid data" are distinguished. The structure should be such that the data from these arrays can be used by most DBMS.

Requirements for data exchange formats

- The format should provide the ability to transfer data using formats based on XML, JSON, NetCdf;
- Any system that recorded the file must subsequently read it and restore all the information;
- Storage media must have technological compatibility at the read-write level.
- The exchanged datasets should be documenting;
- All records must have key attributes;
- The list of parameters must correspond to the composition of observations defined by the current manuals and guidelines;
- The main parameters must have signs of quality;
- When exchanging data, common classifiers and codes are use
- All parameters names must be standardized

Attribute	Name	Code
U_Id	Identification	M4000
O_Name	Organization: name	M4107
Lat	Latitude	M4311
Ta	Temperature of air: measured	P0001_00

Data types

- Structured (point, profile, grid)

- Spatial data has latitude, longitude and type of spatial representation (point, line, polygon)

- Object - this is text-graphic unstructured information - documents, drawings, images, photos, video, sound, and others.

- The images catalog should contain information about the date, time, type and name of the observation platform, instruments, etc.
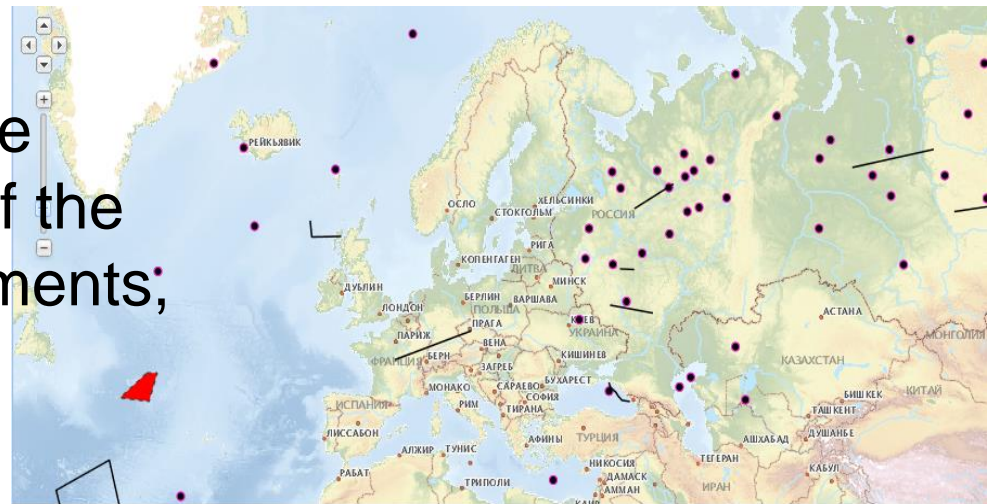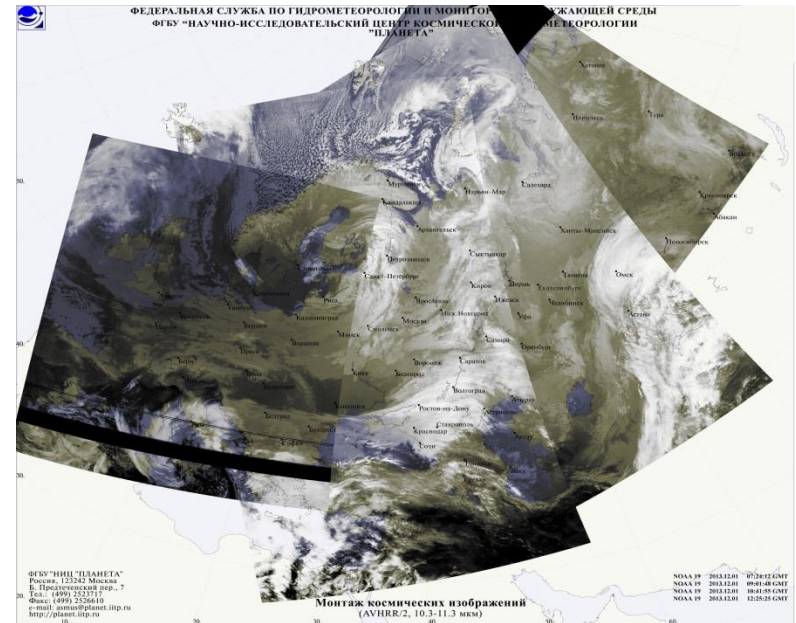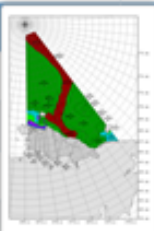
Information on object files

- Catalogue with links to files with documents, maps and figures.
- For all objects, common fields such as ID, date / time of creation, date / time of editing, author, etc. will be highlighted, including fields for organizing links between objects, etc.
- When working with attributes, it is necessary to create generalized formats of fields (number, string, list).
- Descriptions of entities and parameters should be stored separately.

Formats of data

- YaOGMD - Language of description of hydrometeorological data
- CDF, NetCdf - international data storage and exchange formats
- ODV is an international format for storing and processing ocean data
- JSON - a format for exchanging data between applications
- PDS (Planetary Data System) - format for transmitting and storing satellite observations
- GRIB (GRid In Binary), BUFR - WMO standards for data transmission
- GeoTIFF is an open format for representing raster data along with georeferenced metadata.
- GeoPDF is an advanced PDF format for viewing spatial information in Adobe Reader.
- XML is a means of exchanging data with XML schemas for:
  - RSS - news feeds,
  - GeoRSS - Geographic Data Exchange
  - vCard, iCalendar, rel-directory, hReview, etc. – micro formats
  - Dublin Core - internet resources
  - EML (Ecological Metadata Language)
- THREDDS (UCAR, UNIDATA) - server provides a catalogue, metadata and services for access to scientific data, the NetCdf exchange format.
  ERDDAP - converter for major formats


```xml
<?xml version="1.0"?>
<quiz>
 <qanda seq="1">
  <question>
  Who was the forty-second
  president of the U.S.A.?
  </question>
  <answer>
  William Jefferson Clinton
  </answer>
 </qanda>
 <!-- Note: We need to add
 more questions later.-->
</quiz>
```

Multidimensional structure for time series

There are many time series with different space-time resolution scales:

Station name: AAA **Geo-region**: Ru **Spatial**: Fix point **Time resolution**: hour **Date**: GGGG, MM, DD, HH **Parameter**: Air temp	**Station name**: BBB **Geo-region**: Ru **Spatial**: Fix point **Time resolution**: day **Date**: GGGG, MM, DD, HH **Parameter :** Pressure	**Enterprise**: CCC **Geo-region**: Chita **Spatial**: Fix point **Time resolution**: month **Date**: GGGG, MM, DD, HH **Parameter:.** Production release	**City**: DDD **Geo-region**: Ru **Spatial**: Fix point **Time resolution**: Year **Date**: GGGG, MM, DD, HH **Parameter:** birth-rate

Information on time series

ID_time series	Lat	Long	Begin date	End date	Time resolution	Param eter	Author

Time series

ID_time series	Date	Value

Product types

- **An application** (stand-alone) is a type of product that has a permanent link on the web and involves interaction between the user and the visual interface
- **Services** (web, API, rest, portlets) - a type of product characterized by the presence of a permanent link in the network and assuming interaction between two software components.

- OGS standard **mapping services** (WMS, WFS, WCS) - interactive maps
- **Combining** observed and predicted parameters values on one graph
- **Indicating** parameter values.
- Information product **matrices**
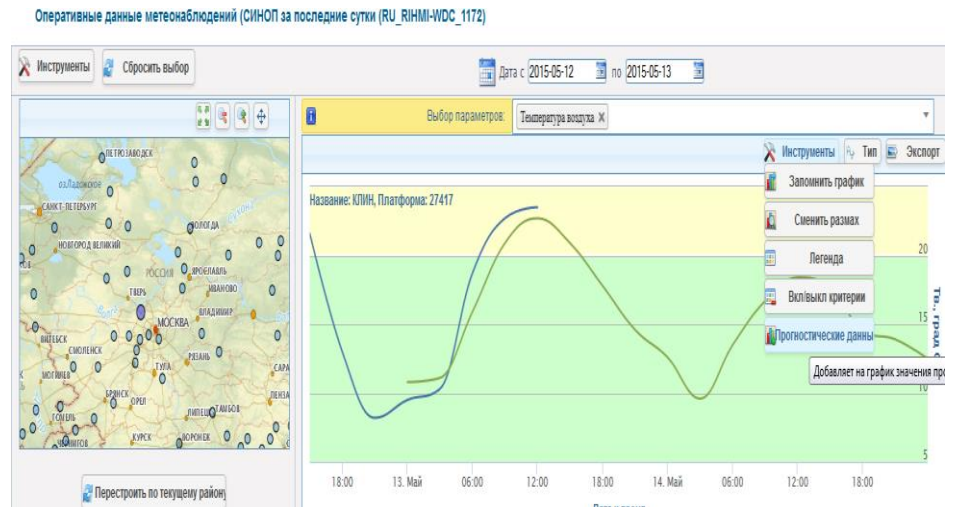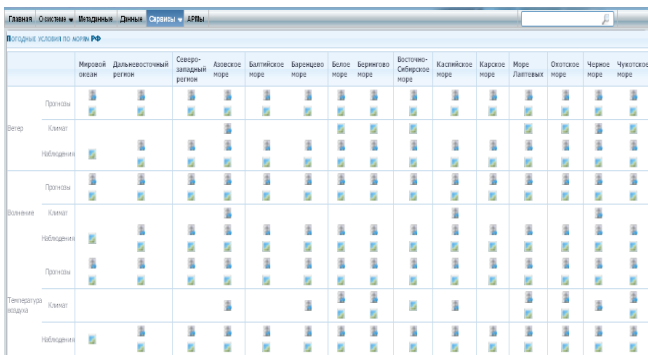
Data access services

- A **web service** is a software system identified by a web address with standardized interfaces. Web services can communicate with each other and with third-party applications through protocol-based messages (UDDI, WSDL, SOAP, XML, XML thematic schemas). Receive data only.
Application programming interface (API) - a set of ready-made classes, procedures, functions, structures and constants provided by an application (library, service) for use in external software products. Using the API, you can organize access to the database.
**REST services** are not a standardized application.
Portlet is a pluggable, plug-in user interface component of a web page based on the JSR 168: Portlet Specification or WSRP standards for remote portlets.
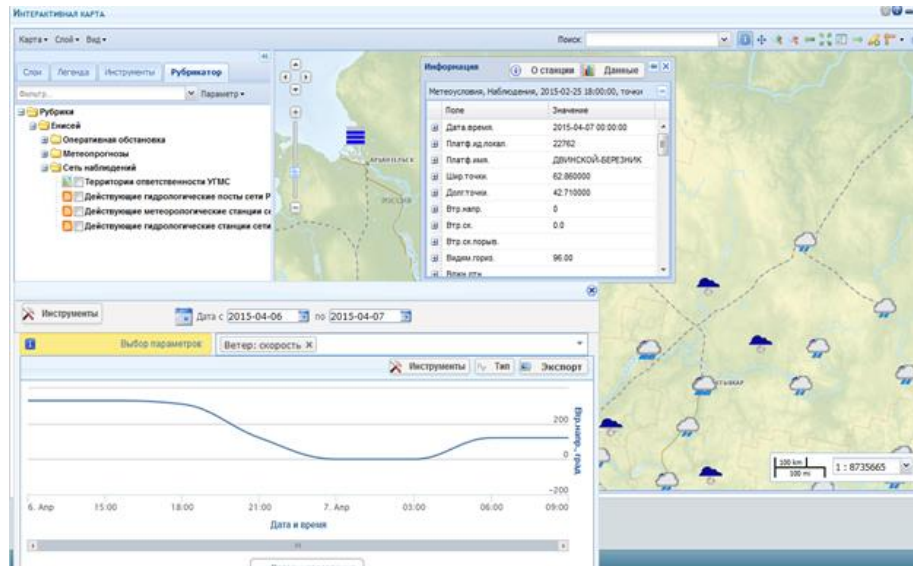
Data presentation (export) formats

- HTML files

- Loading ASCII files with separator

- Export CSV, XML, PDF / A

- Loading data in a zip archive

- Saving charts as files (gif, tif).

- Based on the saved graphs, screens, it can create presentations in ppt or pdf format.
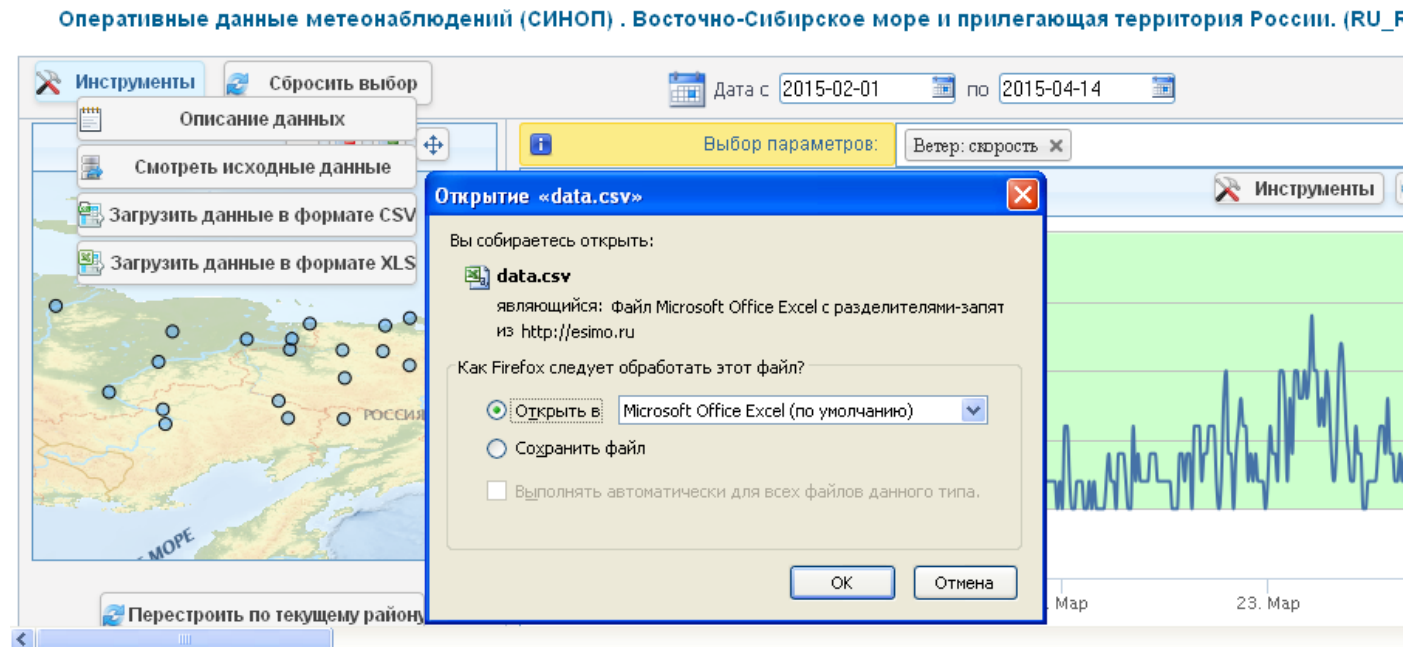
Conclusion

- FAIR requirements for data, TRUST- principles for data warehouses presented
- For a one-time exchange of hydrometeorological data, the formats HDF, NetCdf are used, and for a permanent regular exchange - language XML, JSON.
- For the first time in hydrometeorology, the idea of standardizing formats for storing and exchanging data in the form of time series, grid data, and object file directories is proposed.
- In addition to standardization of formats, it is necessary to standardize common codes and classifiers used in data exchange, a unified parameter vocabulary.
- The prospects are the development of tools of machine-to-machine interaction with existing international and national systems, conveyor data processing for all processes, starting from data collection and their use in business processes

THANK YOU FOR YOUR ATTENTION !

INTegrated ARctic Observing System (INTAROS)
EC HORIZON 2020-BG-09-2016

Agreement between Ministry of Education and Science of Russia and RIHMI-WDC № RFMEFI61618X0103