Institute of Oceanology, Polish Academy of Sciences

Development of data management processes

Marcin Wichorowski

IO PAN

SeaDataNet Final Meeting, IFREMER, Brest, September 15th

Oceanographic data at IO PAN

Extensive archives of information

Institute of Oceanology (IO PAN) maintains data gathered during research activities performed for more than 50 years. These data, originating from different sources are being processed according to common practice elaborated through generations of oceanographers, however there appears to be a very strong demand for standardization of procedures, especially while facing problems regarding cooperation in data exchange

Oceanographic data at IO PAN

Up to late ’90...

Data were managed at unit level, just for publication purposes. No policy for storage, sharing, long time preservation of data collections. High risk of data degradation and lose.

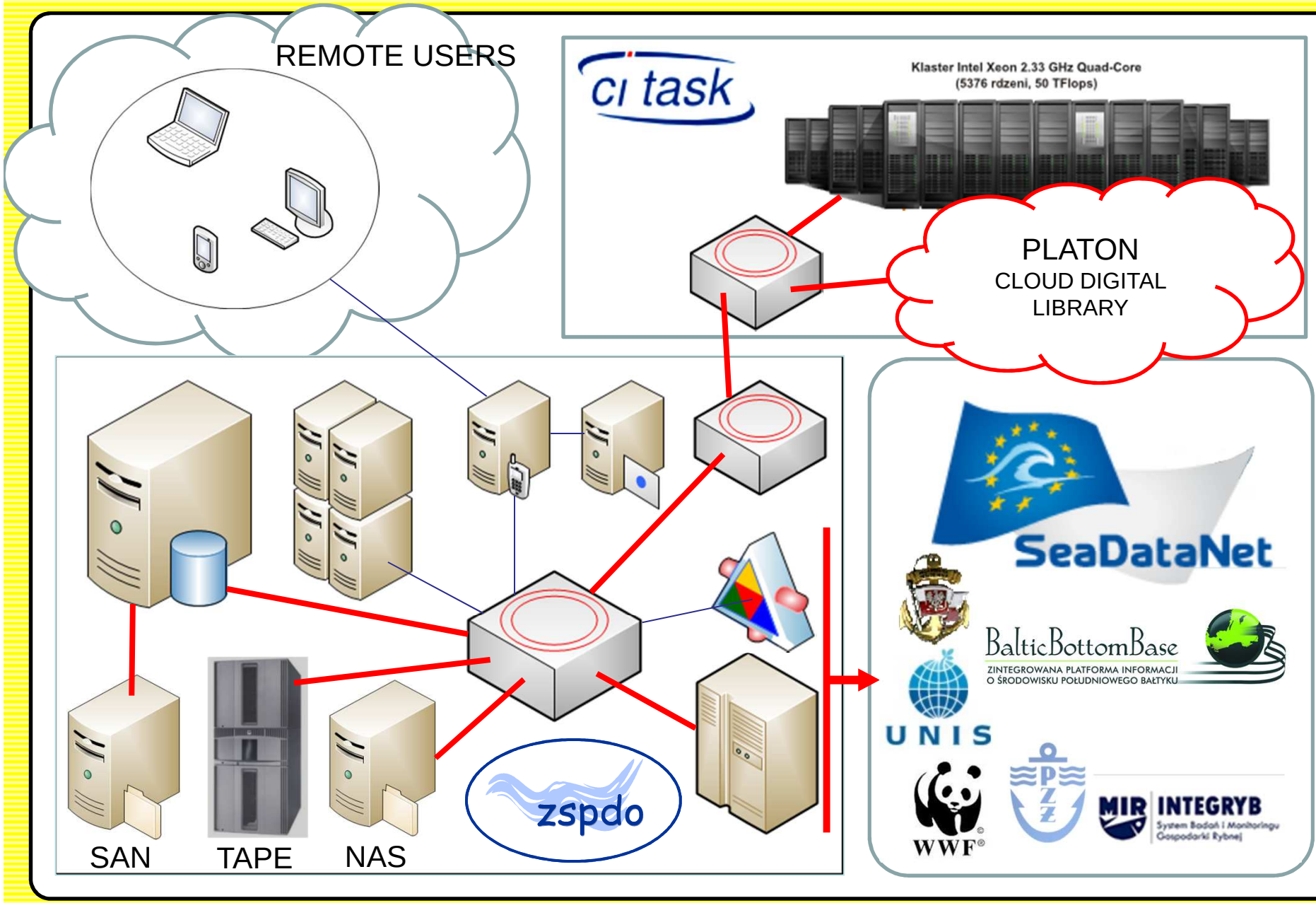
Data management performed on the basis of good practice (good will?) of data originators, in some cases only scientific papers and reports remains.

Before SDN...

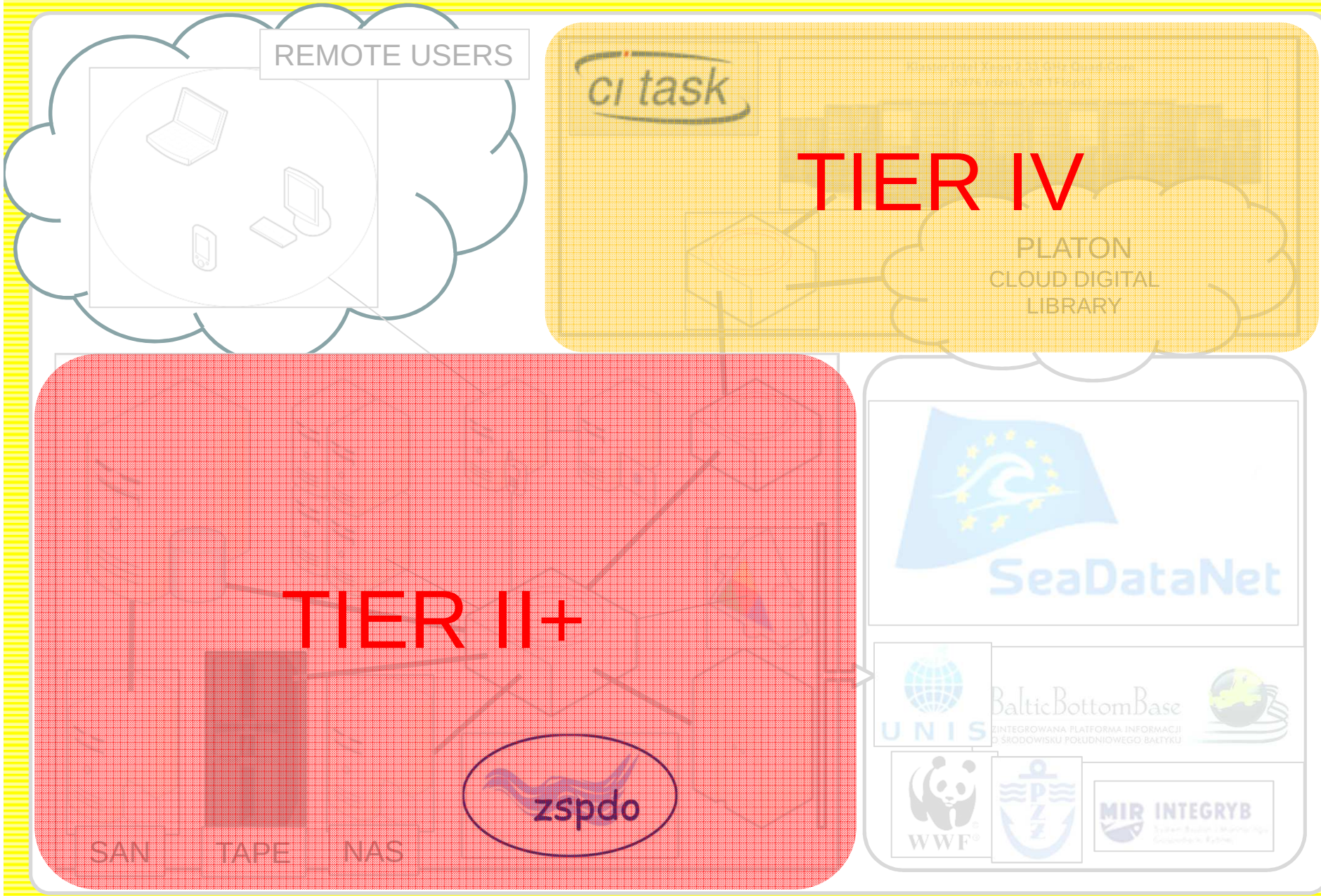
National projects focusing on data management improvement. Transfer of data to digital form, development of policy and deployment of databases reducing risk of degradation. Aggregation of data collections at organisation level. Lack of standards.

2008 – letter of intent, SDN2 associated partner

Data Center Infrastructure

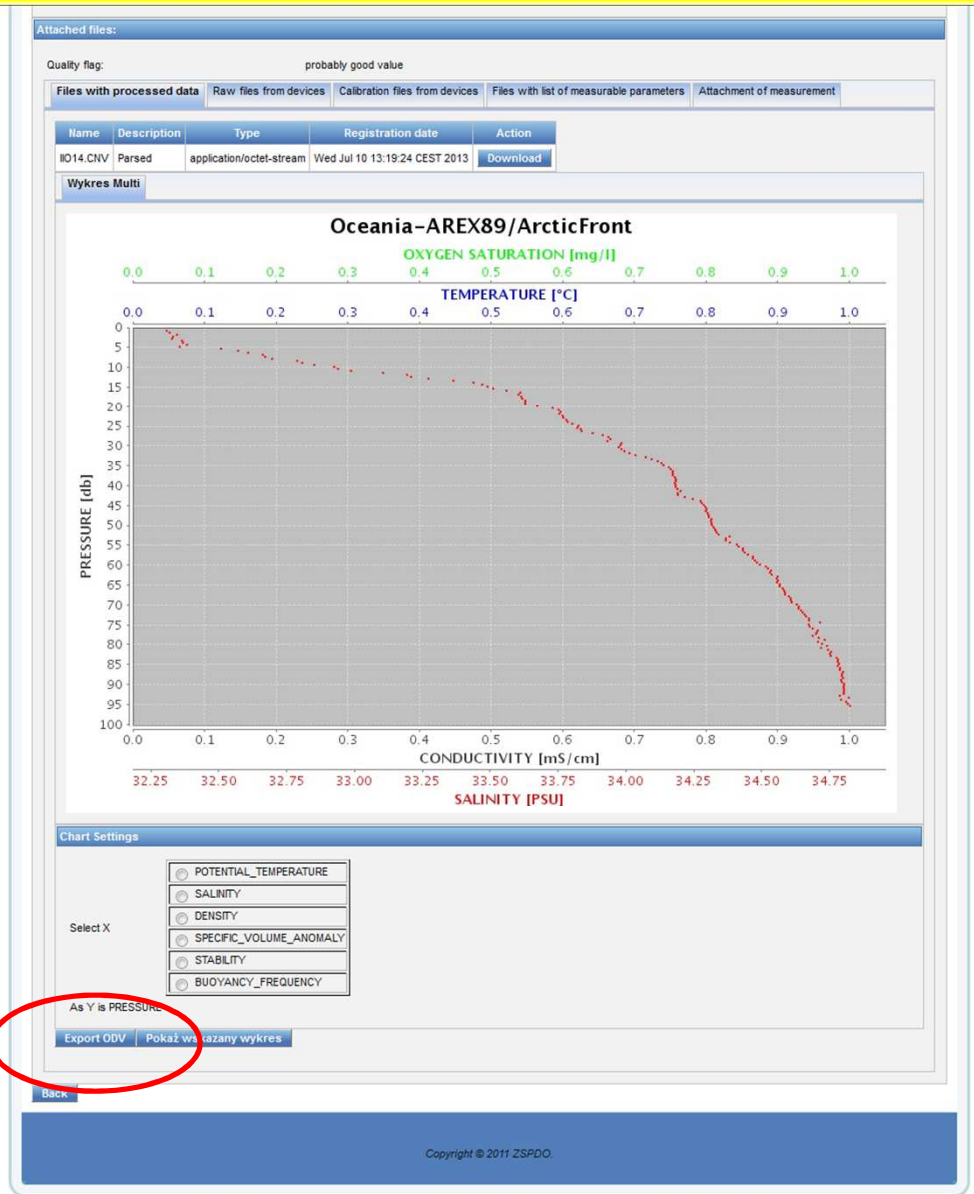


Data Center Infrastructure : TIA 942 standard



Development of ODC: Data sets collection

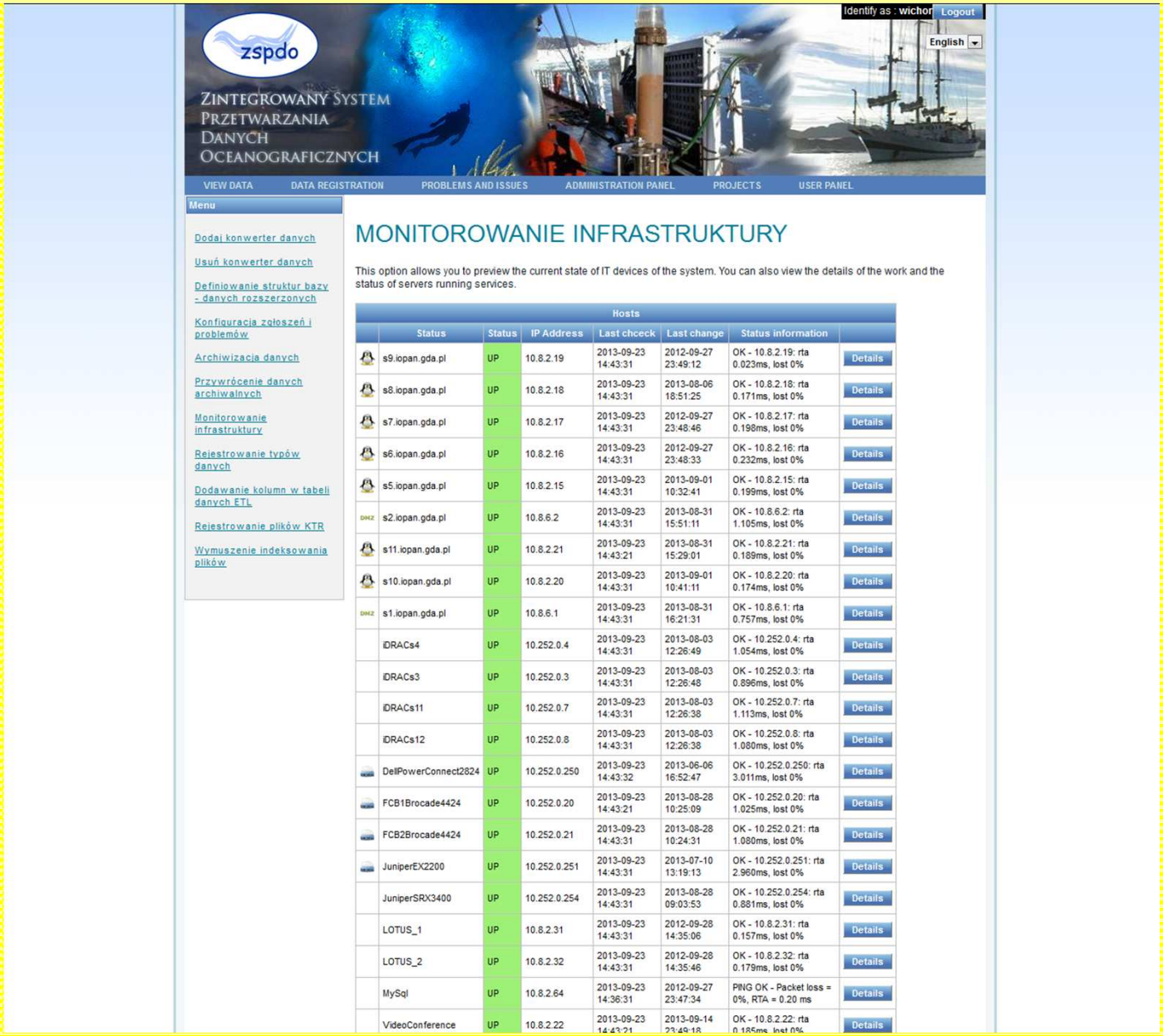
Development of ODC: Data set details



Development of ODC: Vocabulary browsing

Nazwa	Druga nazwa	Url	Definicja
NVS mappings	NERC Vocabulary Server mappings index	http://vocab.ndg.nerc.ac.uk /list/C970/401	A catalogue of the mappings between vocabularies held in the NERC Vocabulary Server.
Geo-Seas skewness descriptors	Geo-Seas sediment grain-size skewness descriptors	http://vocab.ndg.nerc.ac.uk /list/GS50/1	Terms used to describe sediment grain-size skewness coefficients in the Geo-Seas project
Biological entity names	BODC parameter semantic model biological entity names	http://vocab.ndg.nerc.ac.uk /list/S250/29	Terms used to describe biological entities (organisms or parts thereof) in the BODC Parameter Usage Vocabulary
MEDATLAS Parameter Usage Vocabulary	MEDATLAS Parameter Usage Vocabulary	http://vocab.ndg.nerc.ac.uk /list/P09/40	MEDATLAS Parameter Usage Vocabulary
SeaDataNet data access	SeaDataNet data access mechanisms	http://vocab.ndg.nerc.ac.uk /list/L071/2	Terms for mechanisms by which data objects described by SeaDataNet Central Data Index (CDI) records may be obtained before SeaDataNet is fully implemented.
SeaDataNet APG	SeaDataNet Agreed Parameter Groups	http://vocab.ndg.nerc.ac.uk /list/P031/21	Terms agreed within the EU SeaDataNet community to describe coarse-grained groupings of related measurement phenomena.
CSR units	SeaDataNet Cruise Summary Report quantification units	http://vocab.ndg.nerc.ac.uk /list/L181/1	Terms used as units in the quantification of what was collected or measured in a Cruise Summary Report (ROSCOP) record.
SeaDataNet device categories	SeaDataNet device categories	http://vocab.ndg.nerc.ac.uk /list/L05/33	SeaDataNet device categories
SeaDataNet Geospatial Feature Types	SeaDataNet Geospatial Feature Types	http://vocab.ndg.nerc.ac.uk /list/L021/1	SeaDataNet profile of ISO MD_GeometricObjectTypeCodeTerms code list. Known in SEA-SEARCH as 'Library 2' or cdi_measurement_codes.
SeaDataNet Contact Roles	SeaDataNet Contact Roles	http://vocab.ndg.nerc.ac.uk /list/C865/4	Terms used to define the responsibilities for a contact (person or organisation) either within the SeaDataNet project or for the datasets described by SeaDataNet metadata.
MD_DatatypeCode	MD_DatatypeCode	http://vocab.ndg.nerc.ac.uk /list/G120/1	Datatype of element or entity
MD_DimensionNameTypeCode	MD_DimensionNameTypeCode	http://vocab.ndg.nerc.ac.uk /list/G130/1	Name of the dimension
SeaVoX Platform Categories	SeaVoX Platform Categories	http://vocab.ndg.nerc.ac.uk /list/L06/9	SeaVoX Platform Categories
MD_CharacterSetCode	MD_CharacterSetCode	http://vocab.ndg.nerc.ac.uk /list/G090/1	Name of the character coding standard used in the resource
SeaDataNet PDV deprecates	SeaDataNet Parameter Discovery Vocabulary deprecates	http://vocab.ndg.nerc.ac.uk /list/P022/75	Deprecated terms describing fine-grained related groups of measured phenomena designed to be used in dataset discovery interfaces.
SeaDataNet Disciplines	SeaDataNet Parameter Disciplines	http://vocab.ndg.nerc.ac.uk /list/P081/3	Terms used to classify SeaDataNet Agreed Parameter Groups to provide topic/theme level terms in a hierarchical parameter discovery interface
IHB Sea Areas	International Hydrographic Bureau (1953) sea areas	http://vocab.ndg.nerc.ac.uk /list/C161/8	Terms used for sea areas from International Hydrographic Bureau, Limits of Oceans and Seas (Special Publication No. 23), 3rd edition 1953.
POGO ships	Partnership for Observation of the Global Ocean ships of interest	http://vocab.ndg.nerc.ac.uk /list/P173/1	Research vessels deemed to be of interest to POGO. 'Of interest' is defined as active ocean-going research vessels greater than 60m in length

Data Center Infrastructure : Infrastructure monitoring



Data Center Infrastructure : Help desk management

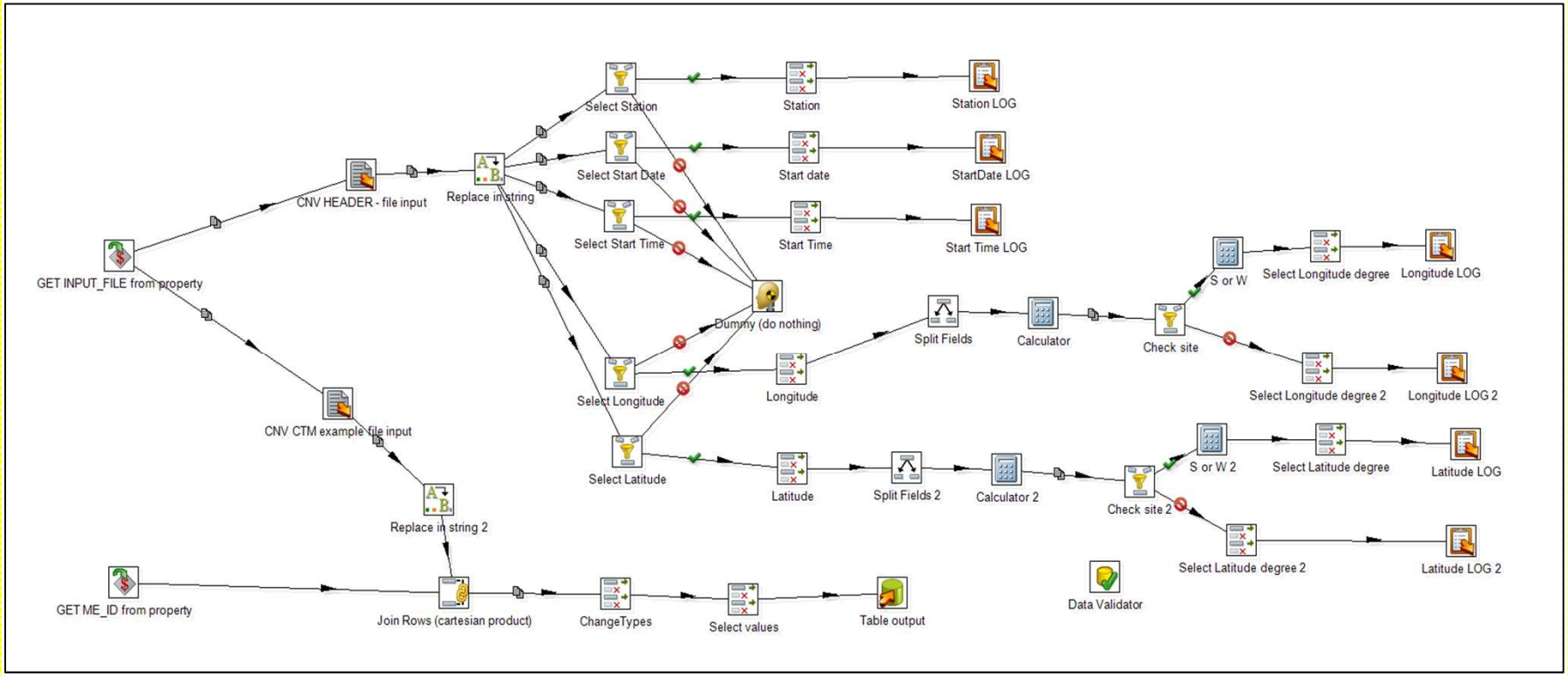
Data Center Infrastructure : Data transformation

```
* Seabird SBE 49 Raw Data File:
* Number of Bytes Per Scan = 6
* System UpLoad Time = Wed Mar 0
* Ship:       Oceania
* Cruise:     Stolpe Channel 2003
* Station:    1rs
* Latitude:   N54 50.707
* Longitude:  E19 18.278
* Bottom:     110
# nquan = 7
# nvalues = 210
# units = metric
# name 0 = pr: pressure [decibar
# name 1 = t0: temperature [deg
# name 2 = sal00: salinity, PSS-
# name 3 = sigma-t00: density, s
# name 4 = sva: specific volume
# name 5 = flag:  0.000e+00
# name 6 = nbin: number of scans
# span 0 = 1.000, 105.500
# span 1 = 2.4134, 8.3162
# span 2 = 7.3739, 13.4070
# span 3 = 5.8815, 10.3359
# span 4 = 1709.57, 2148.93
# span 5 = 0.000e+00, 0.000e+00
# span 6 = 13.0000, 201.0000
# interval = decibars: 0.5
# start_time = Wed Mar 07 2007 1
........
# derive_date = Mar 20 2007 09:1
# derive_in = 1RS000.CNV GUILDL1
# strip_date = Mar 20 2007 09:10
# strip_in = 1RS000.CNV
# file_type = ascii
*END*
      1.000      2.5273      7.375
      1.500      2.5464      7.376
      2.000      2.5096      7.375
      2.500      2.4939      7.375
```

```
<?xml version="1.0"?>
- <transformation>
  - <info>
      <name>Transformation IOPAN test</name>
      <description/>
      <extended_description/>
      <trans_version/>
      <trans_type>Normal</trans_type>
      <trans_status>0</trans_status>
      <directory>/</directory>
    - <parameters>
      - <parameter>
          <name>INPUT_FILE</name>
          <default_value>default</default_value>
          <description>DATA IN</description>
        </parameter>
      - <parameter>
          <name>MEASUREMENTEVENT_ID_PROPERTY</name>
          <default_value>default</default_value>
          <description>Measurement_ID</description>
        </parameter>
      </parameters>
    + <log>
    + <maxdate>
      <size_rowset>10000</size_rowset>
      <sleep_time_empty>50</sleep_time_empty>
      <sleep_time_full>50</sleep_time_full>
      <unique_connections>N</unique_connections>
      <feedback_shown>Y</feedback_shown>
      <feedback_size>50000</feedback_size>
      <using_thread_priorities>Y</using_thread_priorities>
      <shared_objects_file/>
      <capture_step_performance>N</capture_step_performance>
      <step_performance_capturing_delay>1000</step_performance_capturing_delay>
      <step_performance_capturing_size_limit>100</step_performance_capturing_size_limit>
      <dependencies> </dependencies>
      <partitionschemas> </partitionschemas>
      <slaveservers> </slaveservers>
      <clusterschemas> </clusterschemas>
      <created_user>-</created_user>
      <created_date>2013/07/31 09:20:29.637</created_date>
      <modified_user>-</modified_user>
      <modified_date>2013/09/05 10:53:42.204</modified_date>
  </info>
  <notepads> </notepads>
+ <connection>
+ <connection>
+ <connection>
+ <connection>
- <order>
  - <hop>
      <from>GET INPUT_FILE from property</from>
      <to>CNV BODY file input</to>
      <enabled>Y</enabled>
    </hop>
  - <hop>
      <from>GET INPUT_FILE from property</from>
      <to>CNV HEADER - file input</to>
      <enabled>Y</enabled>
    </hop>
  - <hop>
      <from>CNV HEADER - file input</from>
      <to>Replace in string</to>
```

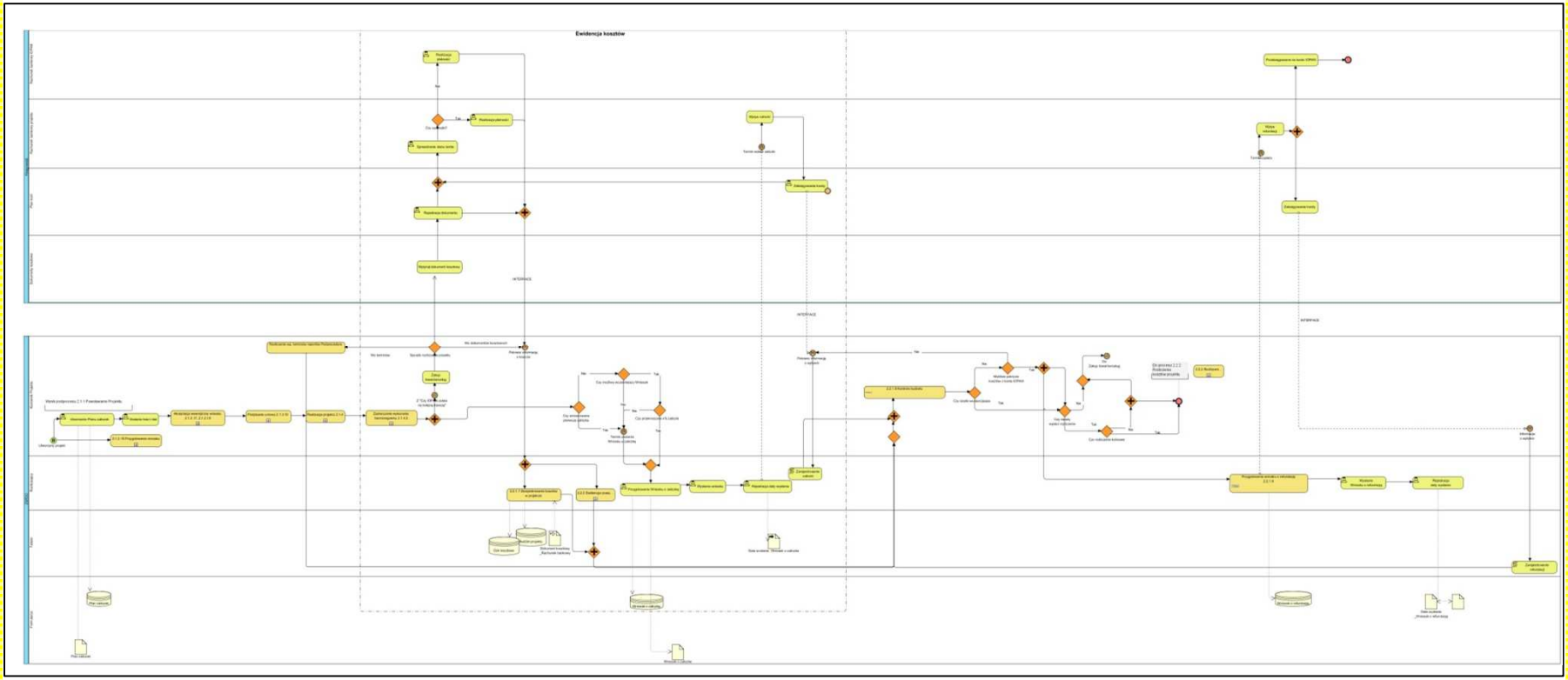
Data Center Infrastructure : Pentaho Business Analytics

- Data access and ETL
- Data discovery and analysis (OLAP engine)
- Data mining

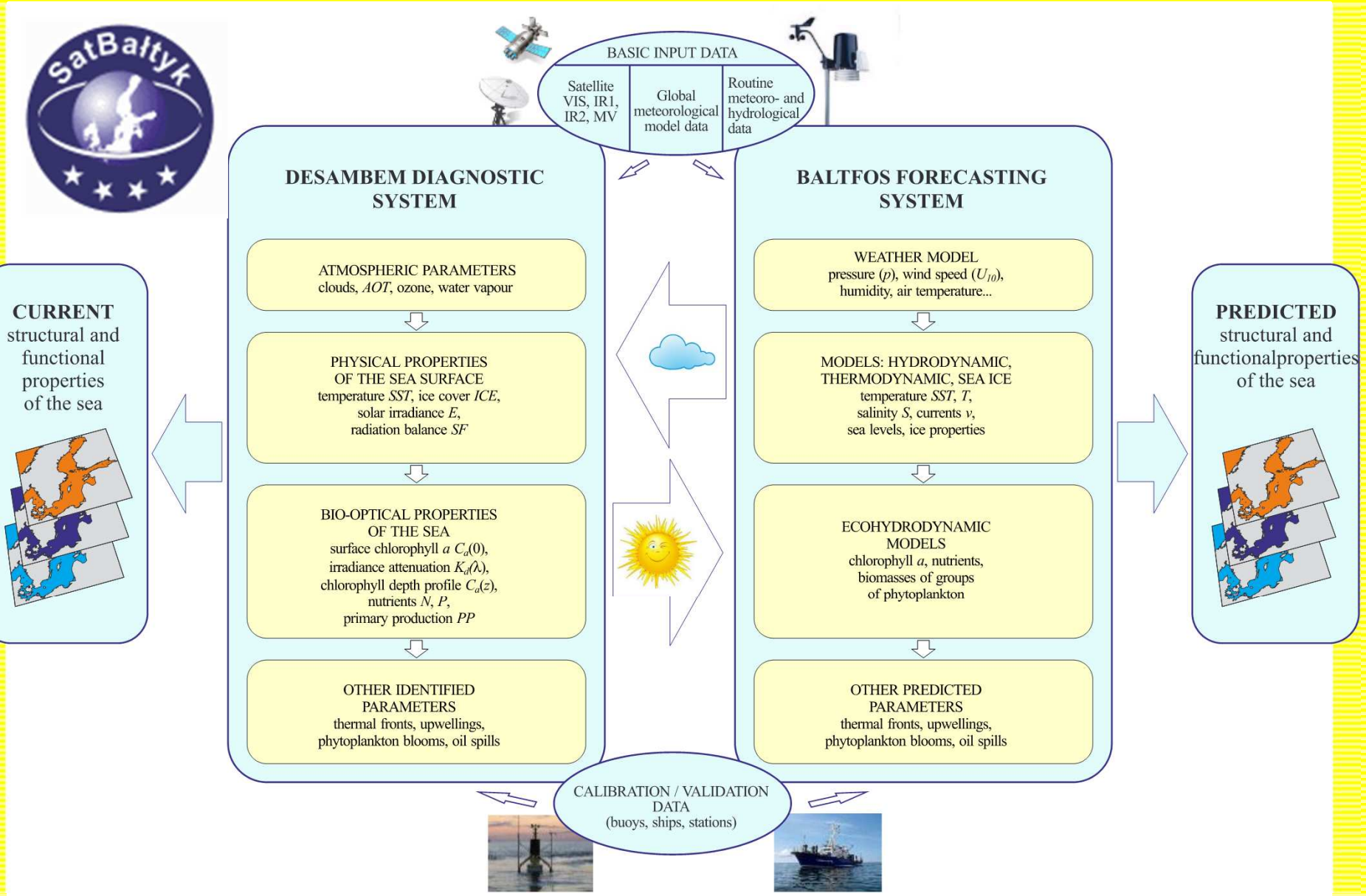


Data Center Infrastructure: Process management

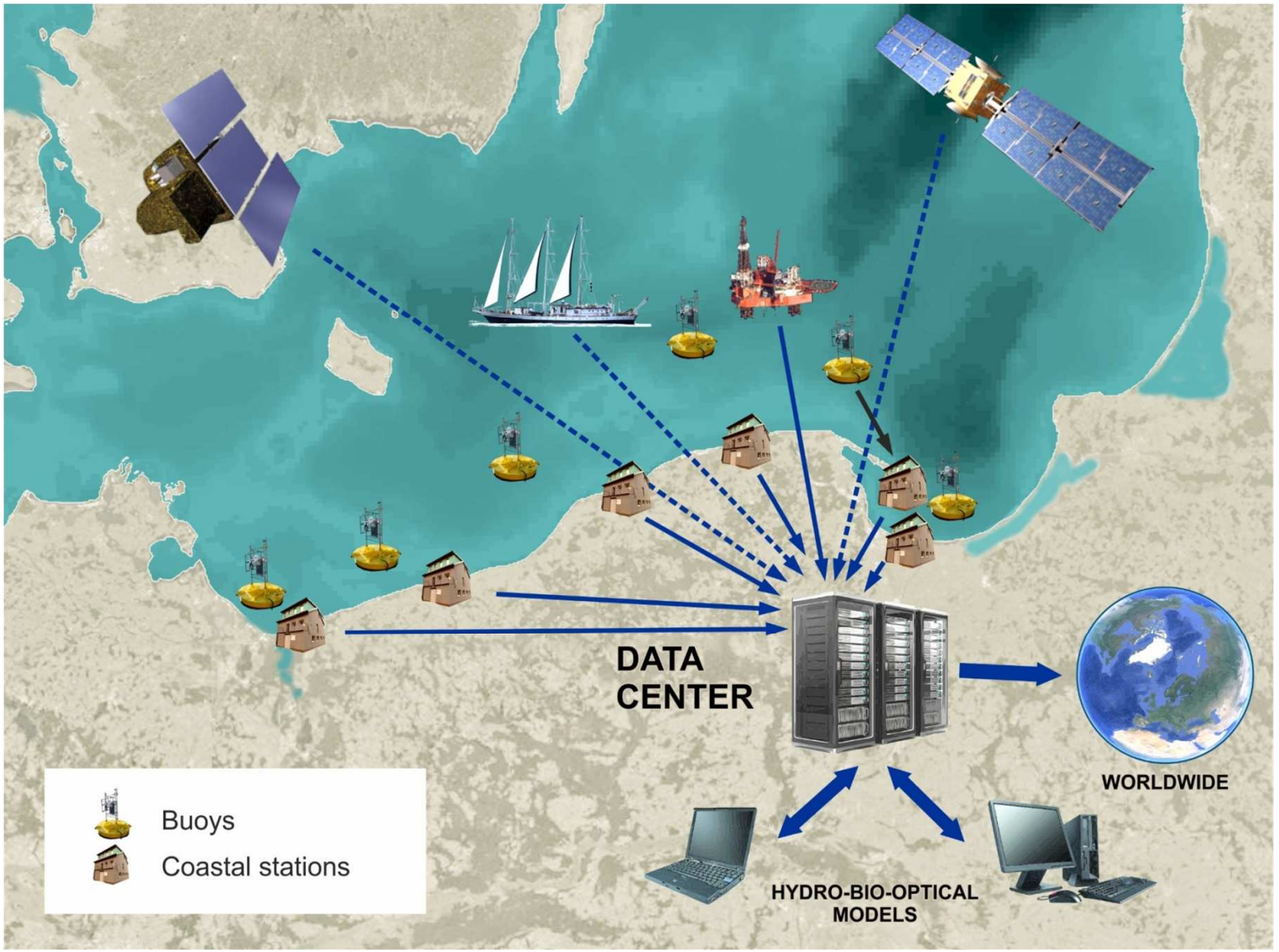
- Processes are defined using BPMN v.2
- Visual Paradigm/Dysant Framework
- Workflow definition by graphical GUI



SatBałtyk Operating System

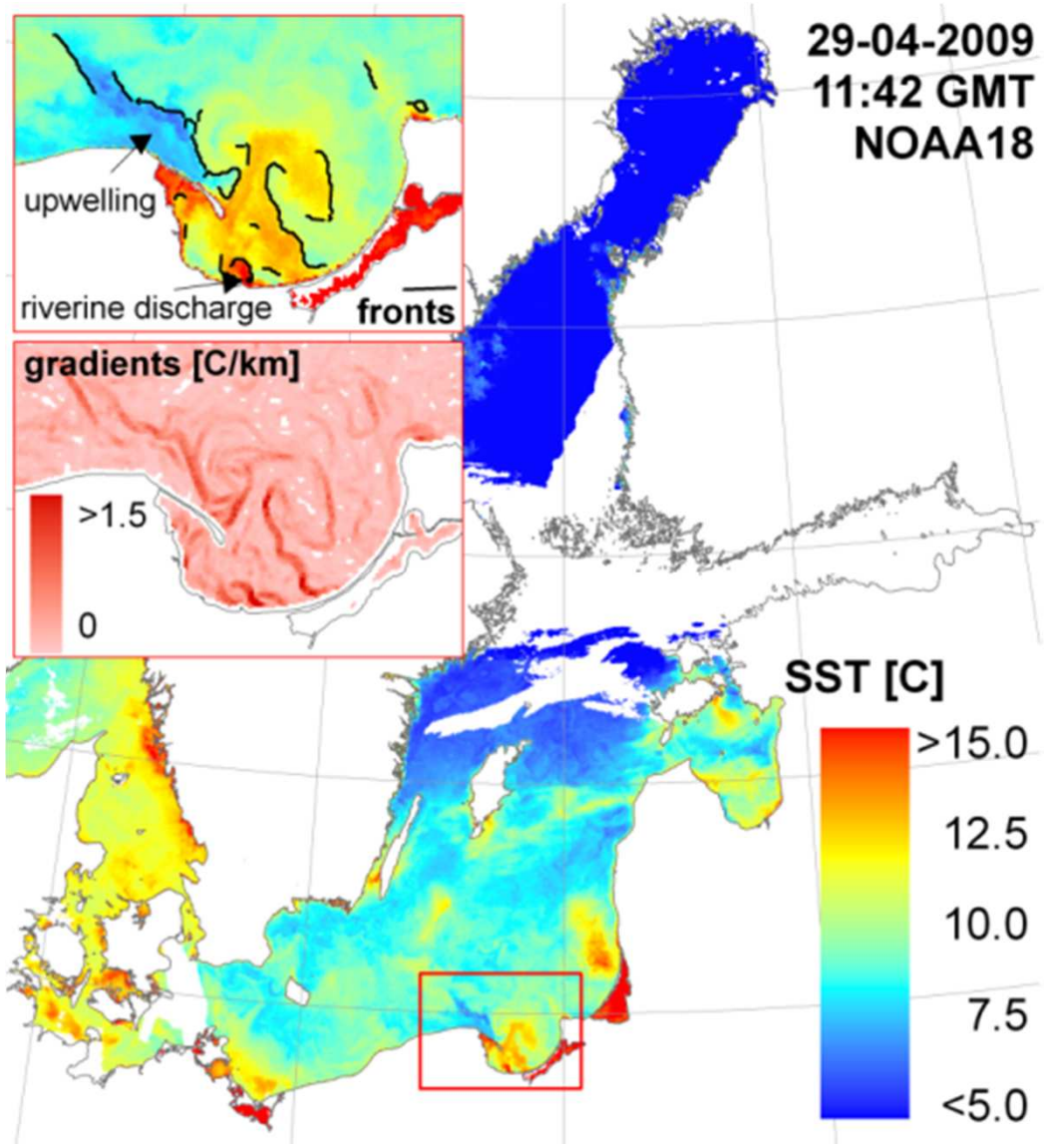


SatBałtyk Operating System



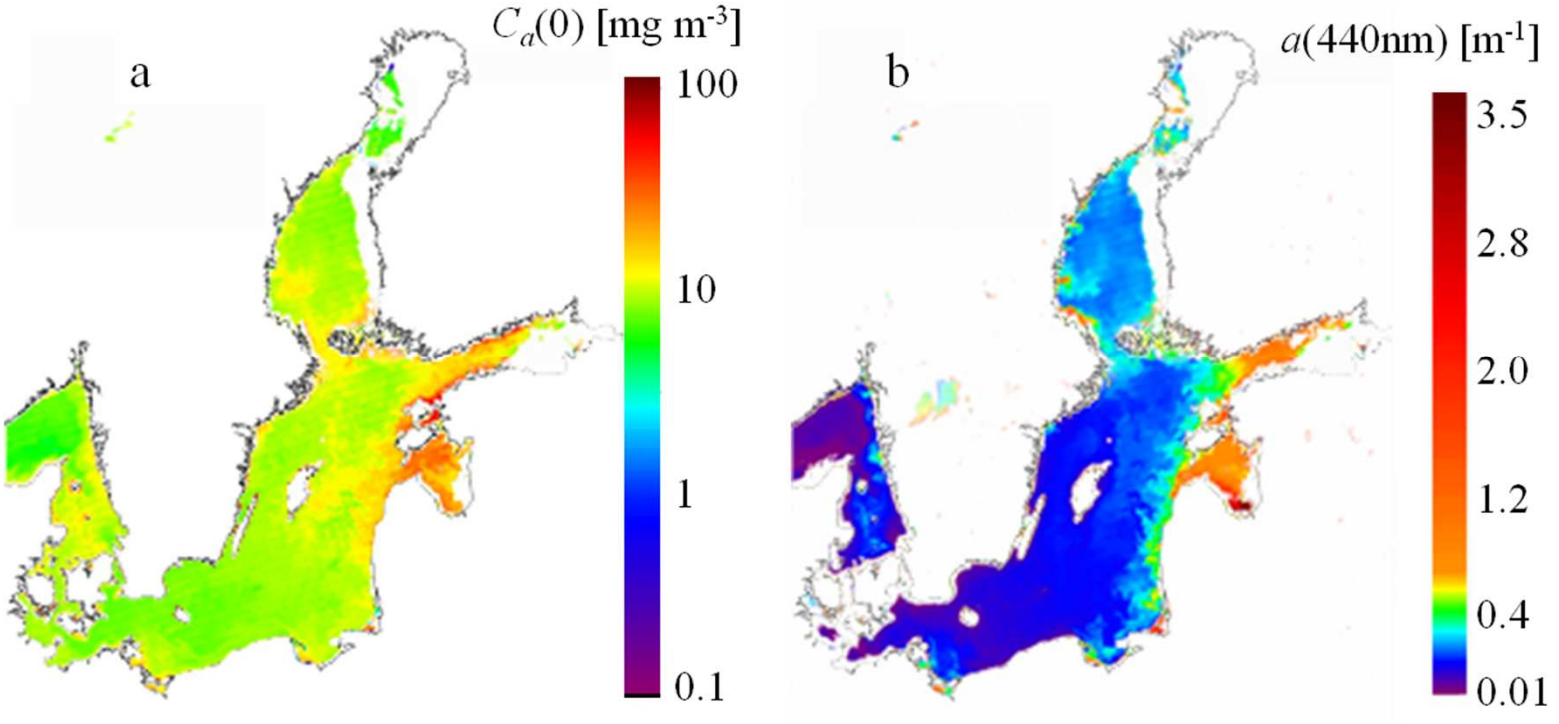
Sea Surface Temperature

Sea surface temperature determined on the basis of satellite data for a cloudless atmosphere and derivatives (map of temperature gradients, location of thermal fronts)



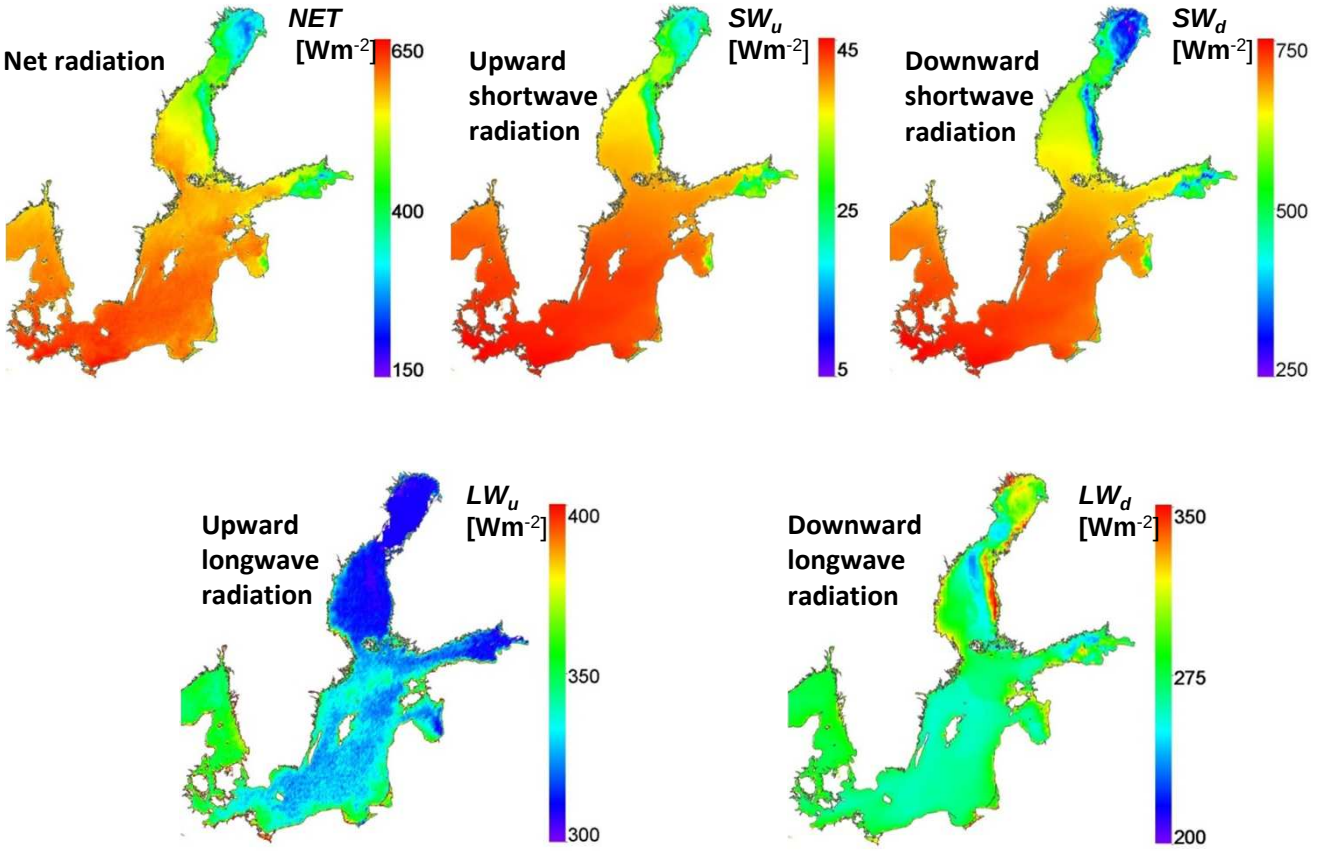
Ocean colour

Surface chlorophyll *a* $C_a(0)$ concentrations (a), and the coefficient of total absorption of light of wavelength 440 nm by dissolved and suspended matter in the sea water, *a*(440nm) (b), on 24th April 2011 *a*(440nm)= $10^{(0.096-0.965*\log(x))}$, where *x* is the ratio of the sea's reflectance for light wavelengths 490 and 665 nm, that is $x=R_{rs}(490)/R_{rs}(665)$.



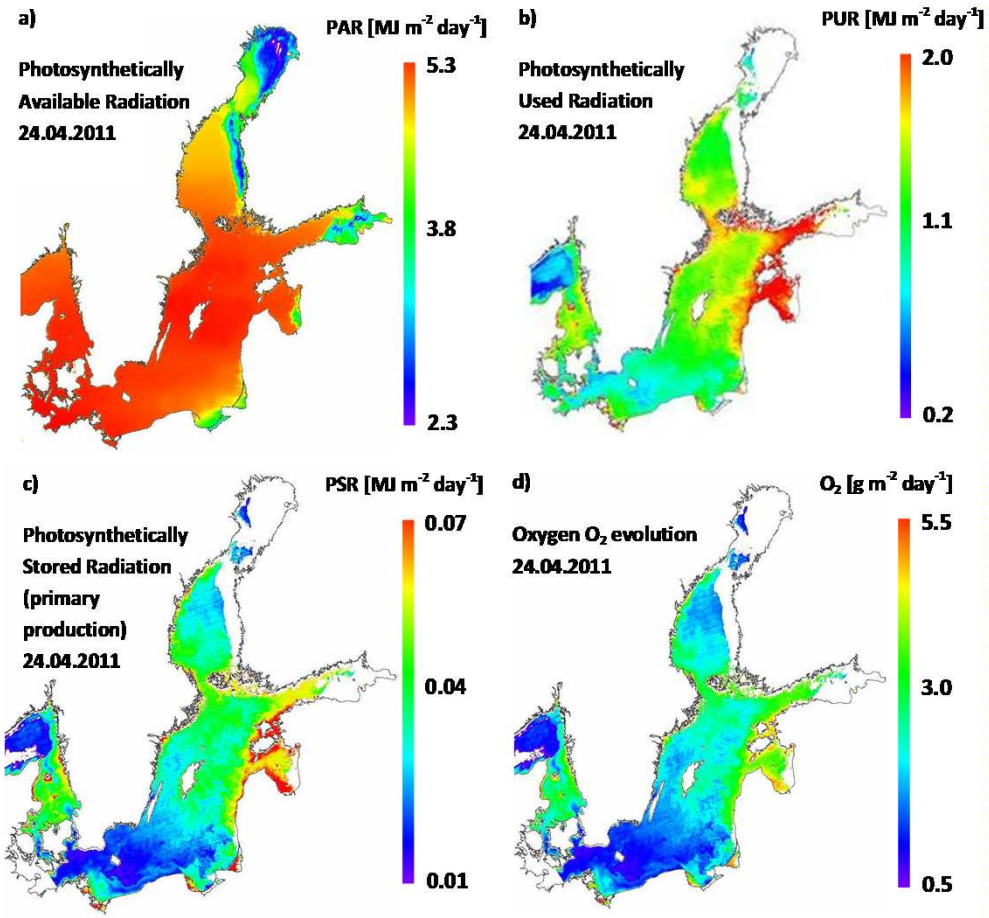
Radiation balance

Instantaneous *NET* radiation flux and its corresponding downward and upward components (short-wave – SW_d, SW_u; long-wave – LW_d, LW_u) at 11:00 UTC on 24th April 2011 at the Baltic Sea surface



Photosynthesis

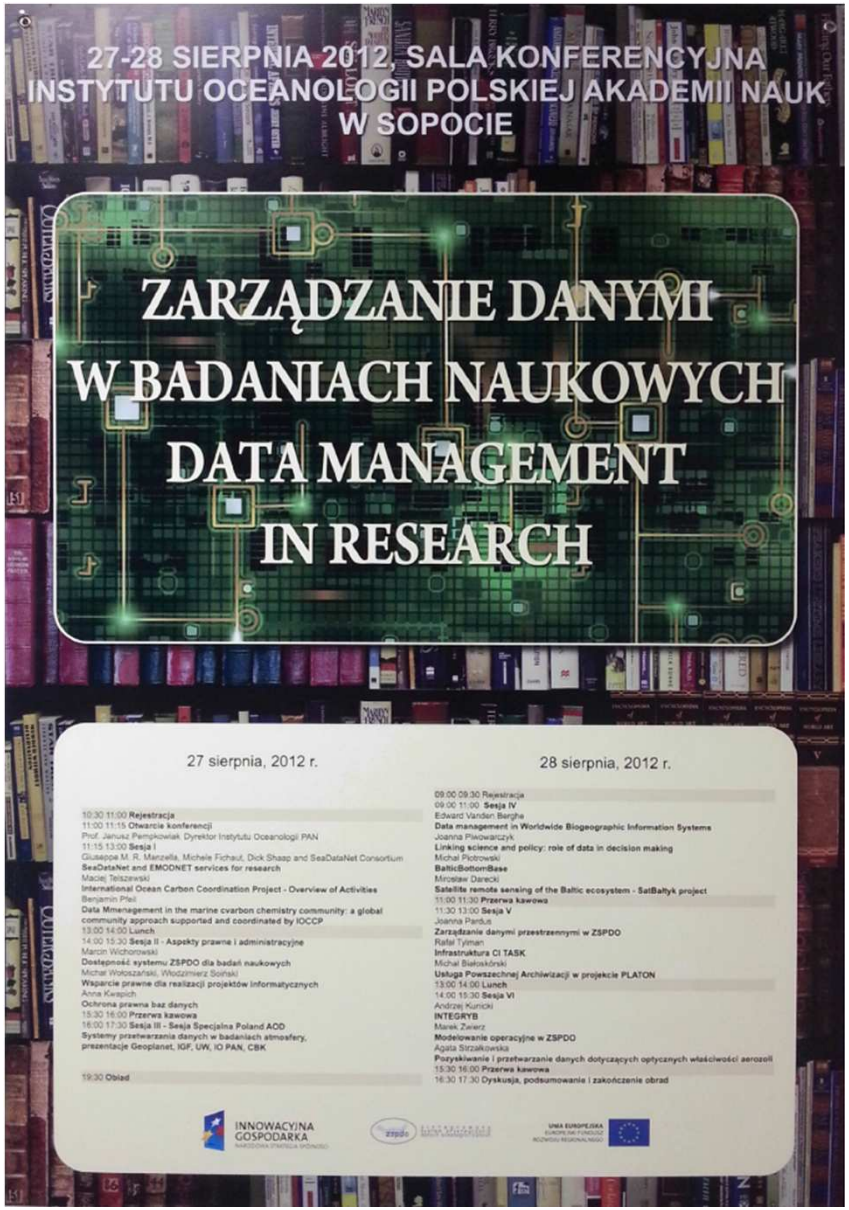
- **Photosynthetically Available Radiation** (400-700 nm), *PAR* (a)
- **Photosynthetically Used Radiation** (the excitation energy of marine phytoplankton pigments equal to the energy of the radiation absorbed by these pigments), *PUR* (b)
- **Photosynthetically Stored Radiation** (the energy incorporated into the ecosystem, that is, primary production in energy units), *PSR* (c);
- **Quantity of oxygen** O_2 released during photosynthesis in the Baltic (d)



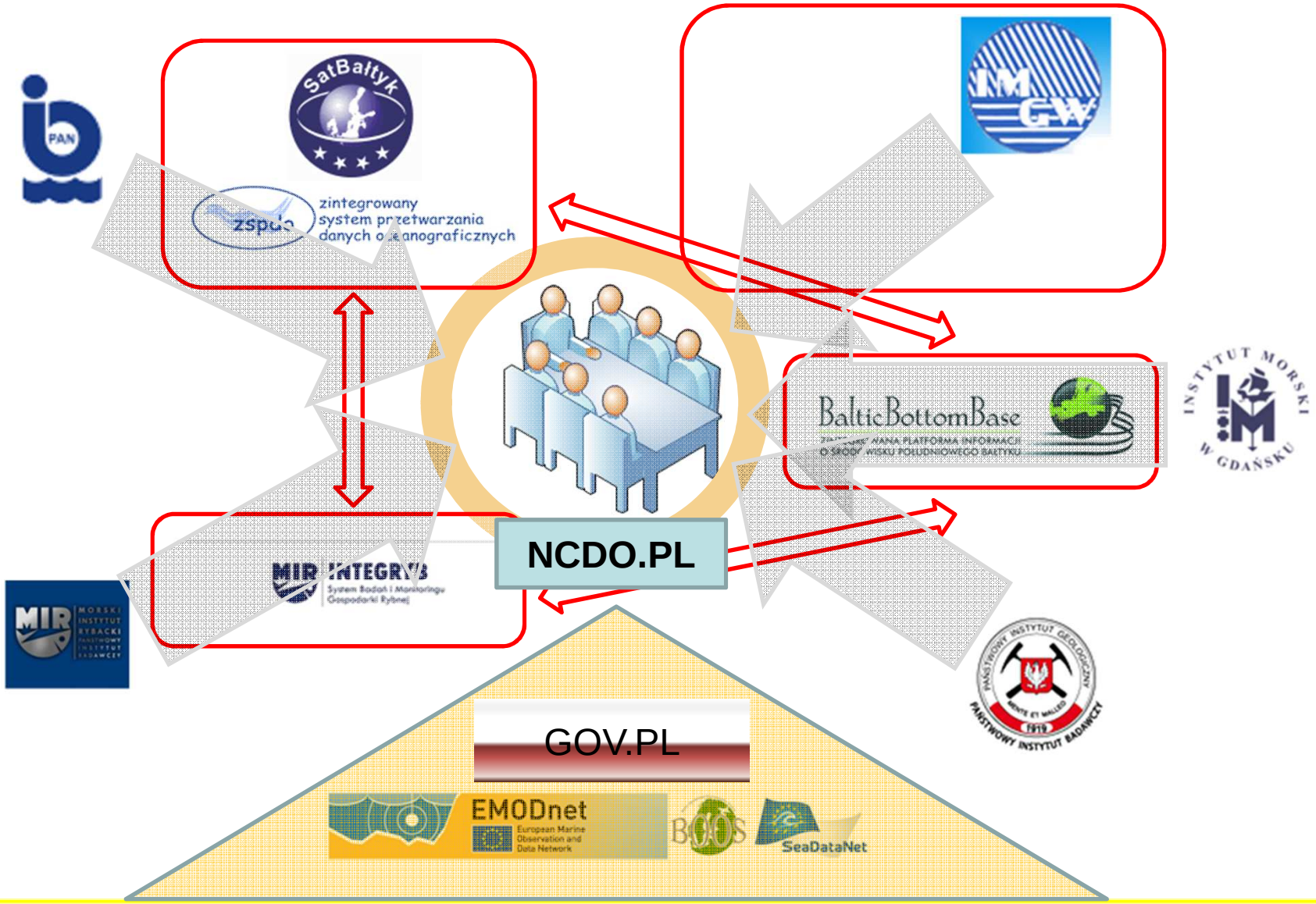
Benefits for the community

- normalization of data and information exchange procedures deploying widely used standards (eg. those developed within SeaDataNet)
- improvement of data exchange with international organisations (OpENDAP, NetCDF, WFS)
- automatisation of metainforamation discovery and publication
- improved management of research projects with developed tools for budget and resources administration
- monitoring of the administration processes and quality management within projects
- improved efficiency of research works providing fast and easy access to data repository
- cross-functioning integration with helpdesk, groupware, eAccess and eLearning modules

Promotion of data management



Aliance for NODC.PL



Conclusions

- Cooperation with SeaDataNet has changed policy, procedures, practices of data management at IO PAN and other organisations .

- Deployment of standards elaborated within SDN frame enforced development of data centre infrastructure providing services for scientific and oceanographic community

- New challenges were trigger of development of new initiatives both inside and outside the organisation


THANK YOU !