Enhancing ARGO Floats Data Re-usability

Gianpaolo Coro, Paolo Scarponi, Pasquale Pagano
Istituto di Scienza e Tecnologie dell'Informazione A. Faedo - CNR, Pisa, Italy

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the ENVRI PLUS project (grant agreement No 654182).

Context

Argo data (www.argo.ucsd.edu) record environmental parameters (e.g. Ph, salinity, pressure, chlorophyll) since Jan. 1999, through a large network of floats. Data are assembled by Global Data Assembly Centers. They are largely used in environmental monitoring systems.

Issues:

- Argo data are in **NetCDF-Point** and **CSV** formats; detailed metadata are described externally to the files;
- Parsing and usage is **hardware demanding** and **requires specific coding**;
- **Not directly usable** in common processing and visualization tools;
- Repositories **not always** easily **accessible**;
- **Poor compliance with Open Science** requirements of re-usability, repeatability, reproducibility.

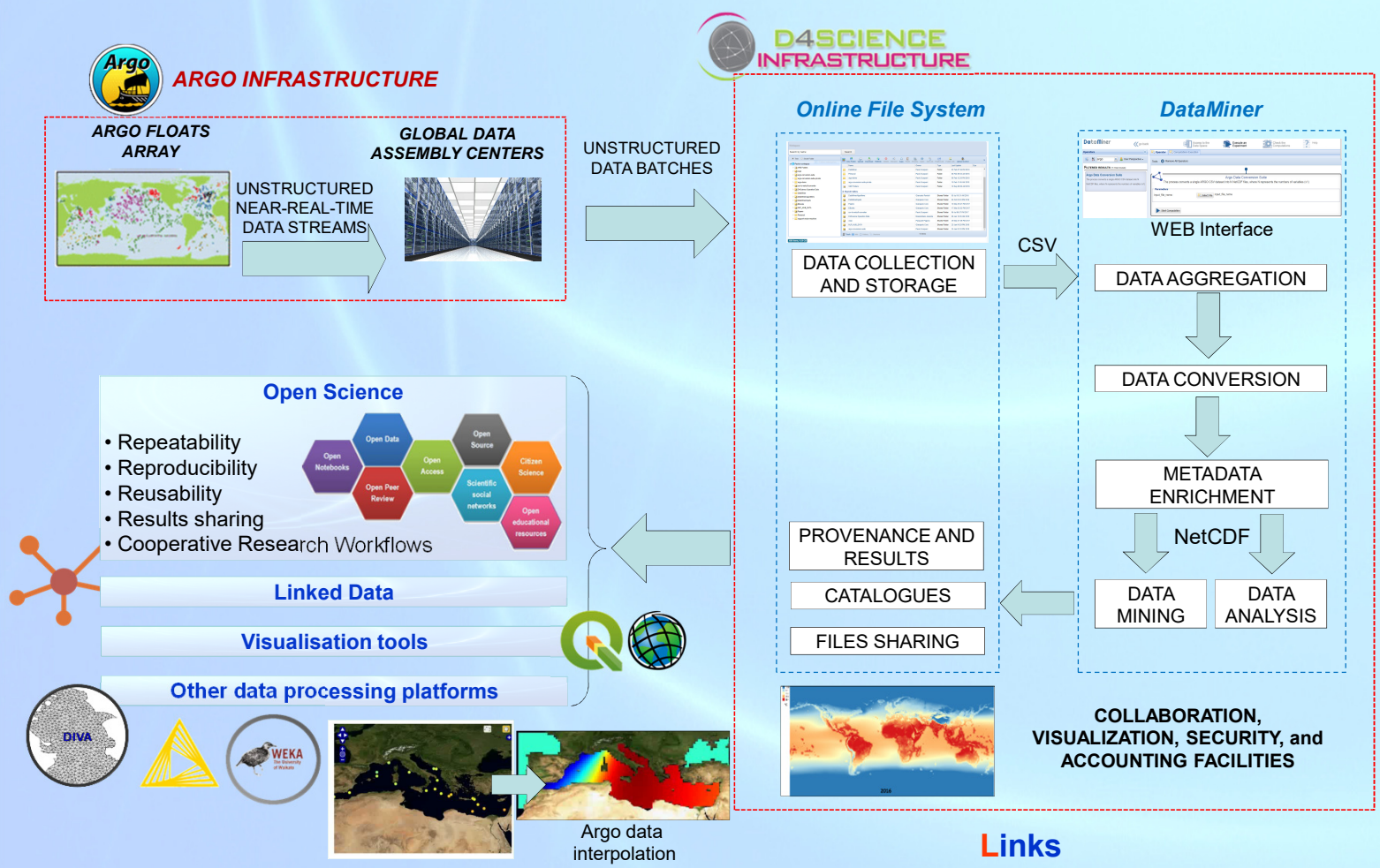
Solution

We implemented an **Open Science oriented workflow** to convert (**5584) Argo datasets** into standardized **NetCDF-CF Grid** files. We used the **DataMiner** cloud computing system of the **D4Science e-Infrastructure** (www.d4science.org):

Workflow:
1.Retrieve monthly-observation datasets from Argo;
2.Represent metadata (i.e. variable names and descriptions, units of measure, etc.), in compliance with the Climate and Forecast (CF) standard vocabulary;
3.Uniform Z dimension by porting all depths to meters;
4.Generate 3D grids with 10 logarithmic-divided depth intervals and 0.5° longitude-latitude resolution;
5.Clamp observation values to this 3D grid and associate average-aggregated values to each cell;
6.Create NetCDF-CF files for every monthly ARGO dataset and variable.

Advantages:

- Argo files are processed by **20 machines** in DataMiner (Ubuntu, 32GB RAM, 16 vCores);
- The transformation process is **reusable** – published under the OGC **Web Processing Service** standard;
- **Provenance** (link to the original input, processing metadata, etc.) is attached to each produced dataset;
- **NetCDF-CF Grid** files are directly **usable** by many modelling and visualization tools;
- Access via **OPeNDAP** protocol is enabled and metadata are described in **ISO-19139** format;
- Availability within a free-to-use **Virtual Research Environment** and via **high availability services** in D4Science.



Conclusions

- Our approach is **compliant with Open Science advices** and allows **adding specific terms** as variables and global attributes in order to **connect Argo data to domain-specific ontologies**.
- Metadata in ISO-19139 and NetCDF-CF Grid files allow for
  - **Easy retrieval** of ARGO data;
  - **Connecting** data **to other catalogues;**
  - **Using** data **in other processes**, e.g. the SeaDataNet DIVA service, ecological niche models, etc.;
  - **Visualising data** in common tools (e.g. ArcGIS and QGIS).
- Publication in a free-to-use Virtual Research Environment enables **security, access monitoring,** and **files/information sharing**.

Links

D4Science **BiodiversityLab Virtual Research Environment**:
https://services.d4science.org/group/biodiversitylab/

VRE Data Catalogue:
https://services.d4science.org/group/biodiversitylab/data-catalogue

VRE Integrated Visualization Tool:
https://services.d4science.org/group/biodiversitylab/geo-visualisation

Argo Data Conversion WPS Process:
https://services.d4science.org/group/scalabledatamining/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ARGO_DATA_CONVERSION_SUITE