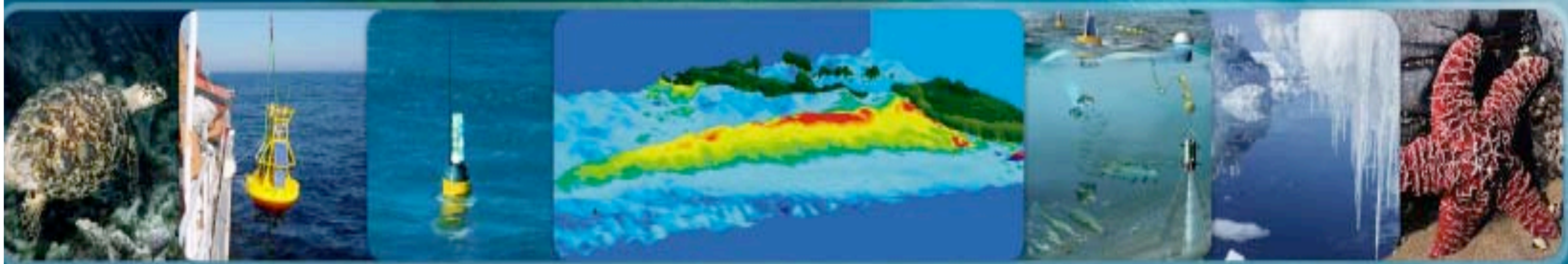
A Decade of Improvements in Data Systems and Data Sharing

S Pouliquen, S Hankin, B Keeley, J Blower,
C Donlon, A Kozyr, R Guralnick

OceanObs09
Venice 21 September 2009



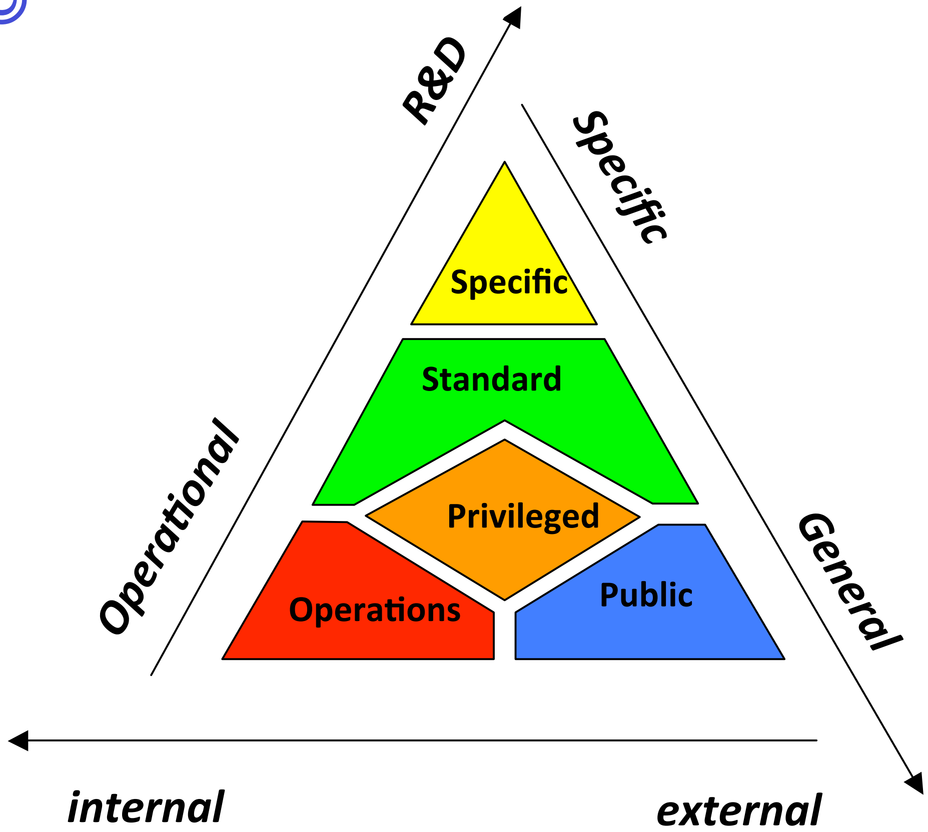
The ways of Using Data have changed

- **A lot of data are collected but not easily accessible while they could serve multiple applications**
- **The nature of science has changed : more work on context, impact synthesis rather than well individual process studies . Need for other scientists data**
- **Government agencies requirements have changed**
  - ⇒ Global scale applications requiring integrated observing system (climate change, ocean health monitoring, fisheries assessment,…)
  - ⇒ No country can pay the full bill for the data acquisition
  - ⇒ A new paradigm emerging : "Data acquired with public funds should be publically available"
- **Important demand for real time data access especially in operational oceanography and monitoring applications**
- **Information Technology and Data Management techniques are no more an obstacle to information sharing**

Ingredients for smooth Data sharing

⇒Find the compromise between user needs and providers capabilities ☺

- ⇒Categories of users from operational to research
- ⇒Categories of services from routine & robust to specific and tailored
- ⇒Categories of providers : operational agencies, research community, private sector



User classification
Courtesy to Mersea EU project

Providers used to tailor their service for their direct users .
BUT Serving other users need more work and harmonization with others

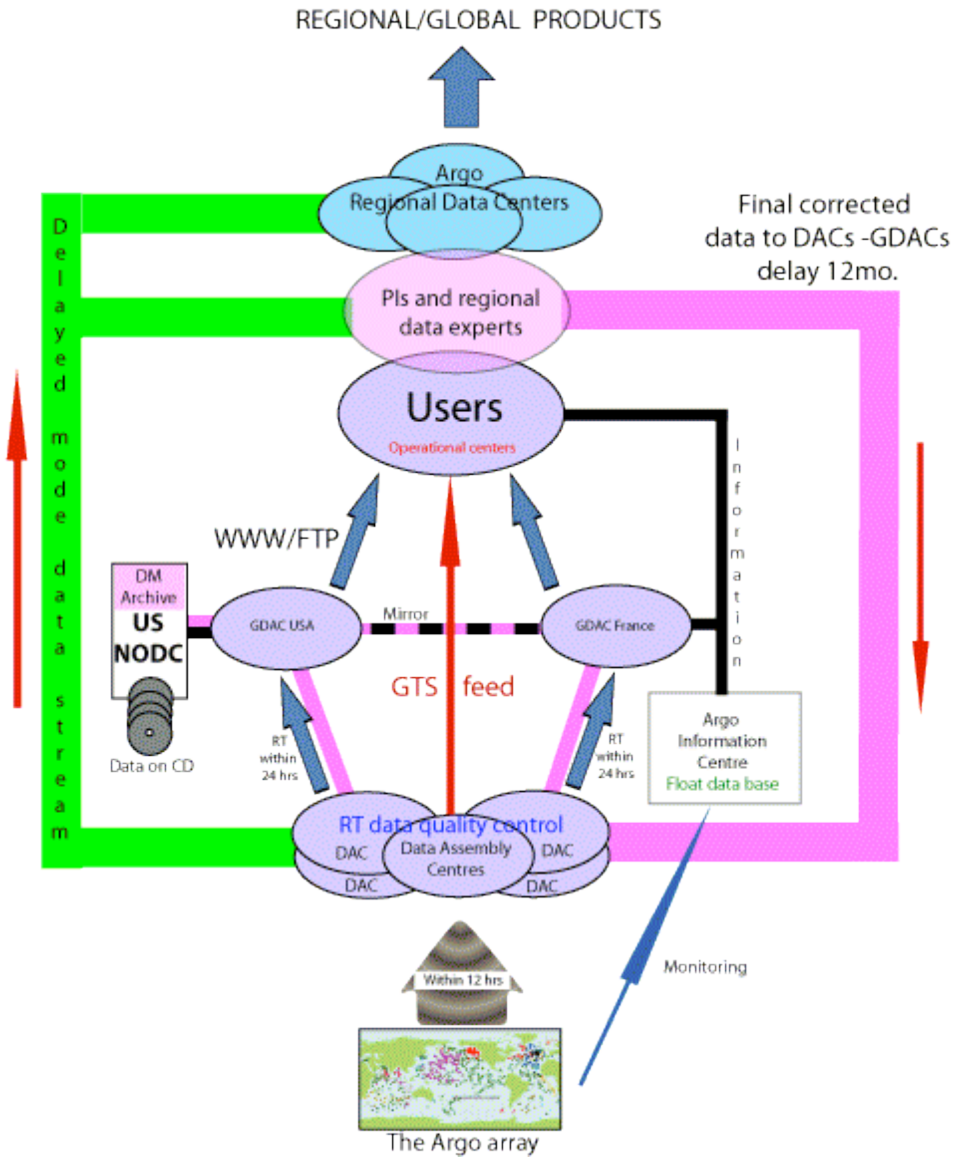
Distance between an observing system and a User

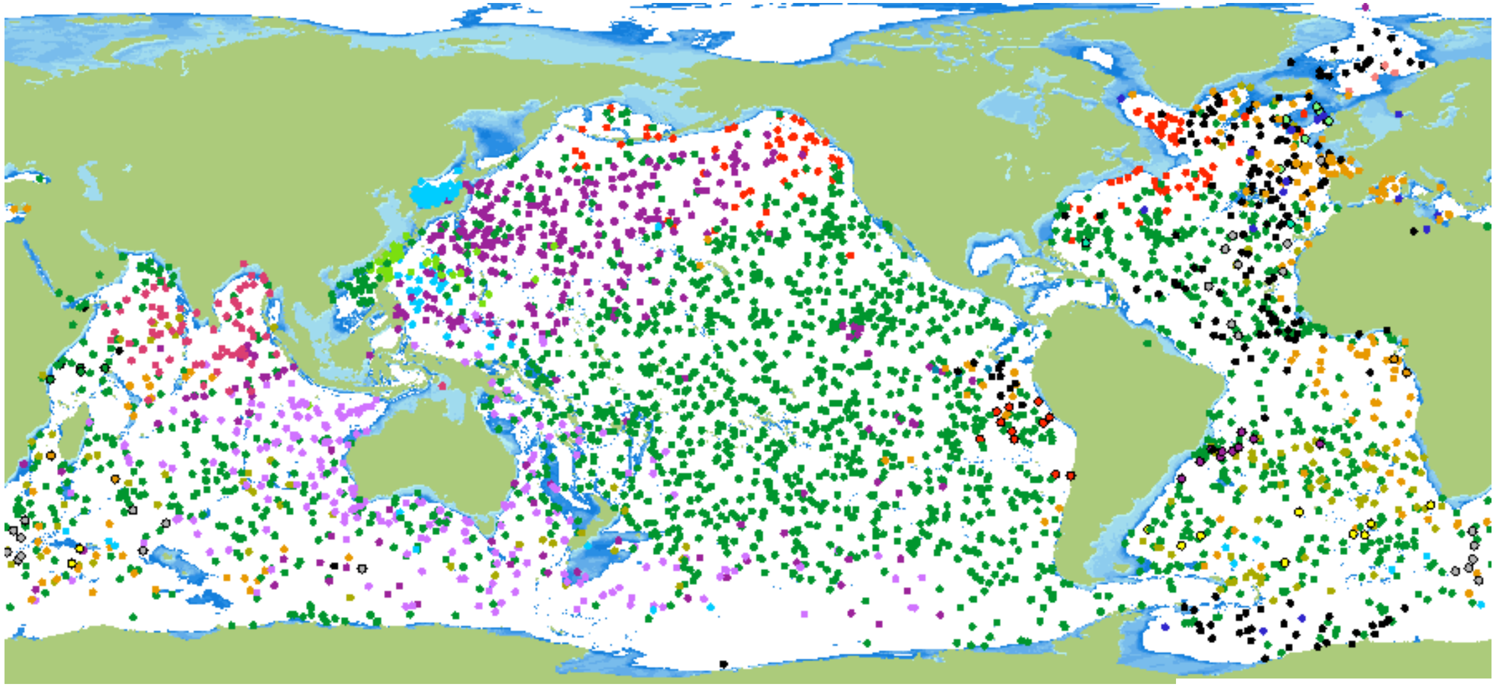
Time
???
60's
90's
Now
Soon

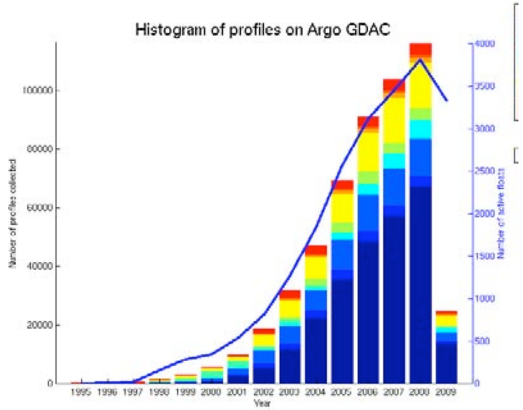
From	To	Who	What
Platform	Media	**Provider**: can be a scientist, an agency, a private company,…	• **Collects** data , transform sensor measurements into oceanographic information. • Records/**Describes** the way the data has been collected
Media	Repository	**A Data Center**	• **Archives, quality control, distributes** the data to the external users • Ensures that required **metadata**, have been filled in and **traces** the history of the processing of the data (version tracking)
Repository	Service provider	**Thematic Assembly Center**	• **Integrates** various data into a coherent product, **checks the coherency** of the data coming from various platforms, provide feedback to Data centers when anomalies are detected. • **Derives value added products** designed for a kind of application • **Ensures distribution to a wider community**
Service provider	End users	**Service providers**	• **Customizes product from specific applications** combining a wide variety of observation, models outputs and expertise.

SOME SUCCESS STORIES

- **Open data policy**
- **Multi stage data processing**
- **Centralized data holding**
- **Common Format**
- **Common Quality control procedures in Real-Time and Delayed mode**

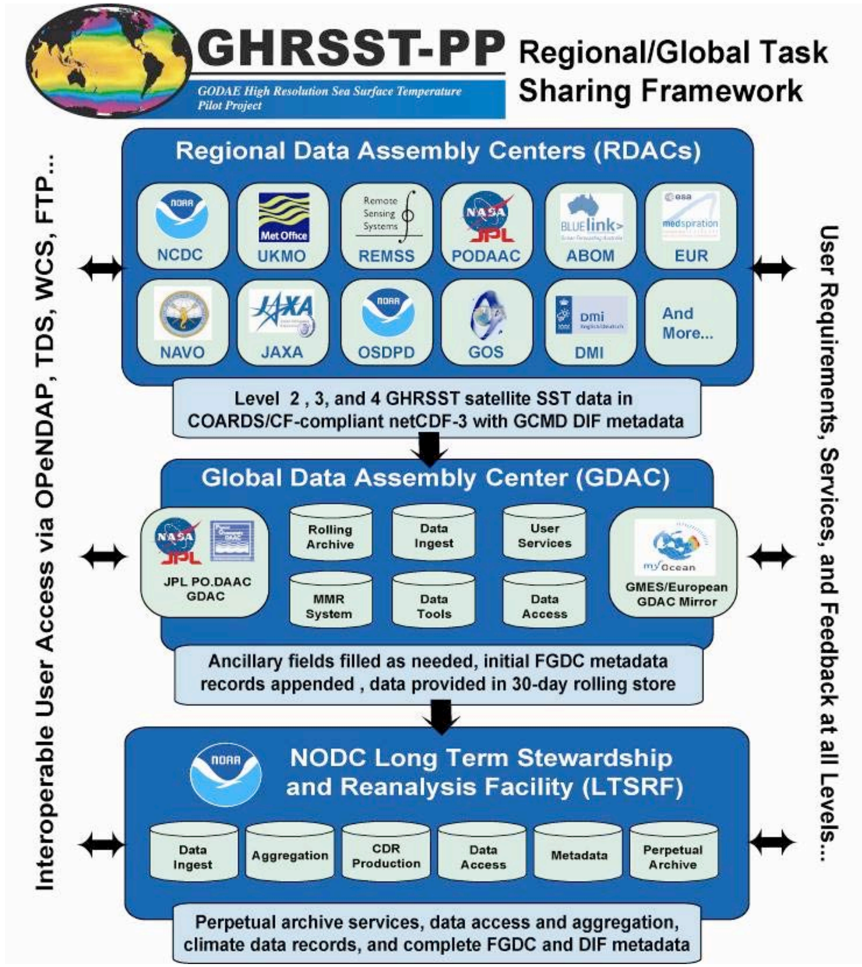


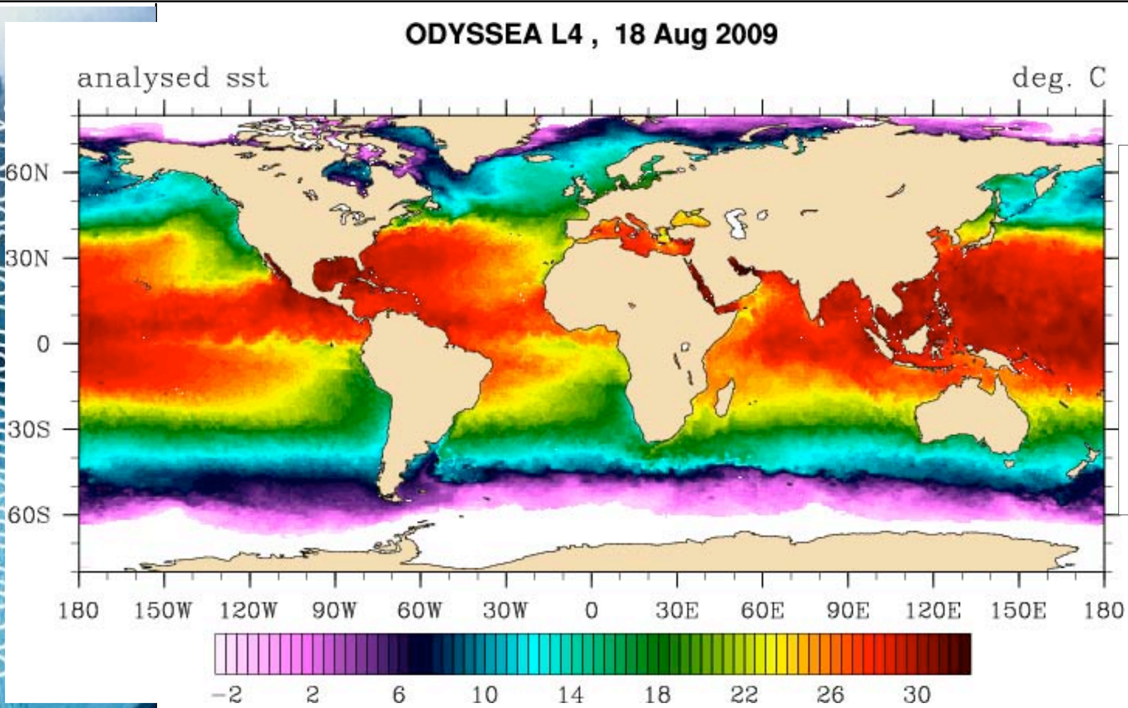


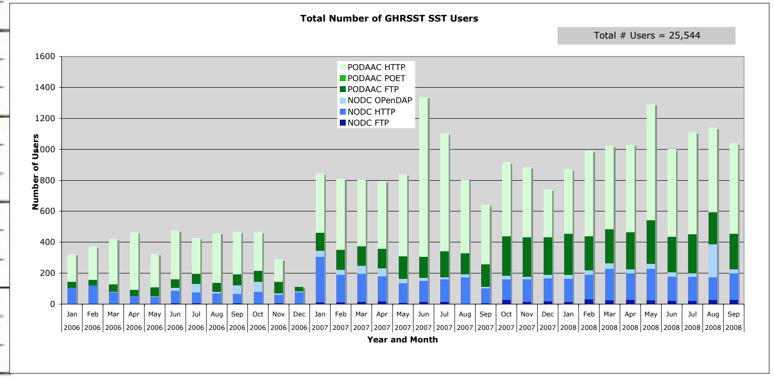


- 90% of the data are processed within 24h and available freely
- More then 320 000 profiles at GDACs. 9 000 new profiles per month
- 10 National Data Centers, 2 Gdacs, 12 Delayed mode operators, 5 Regional centers

- **Open data policy**
- **Decentralized data processing , centralized distribution at GDACs**
- **Common format data provided with uncertainty estimates and ancillary parameters**
- **Value added products generated at Regional Data Centers**



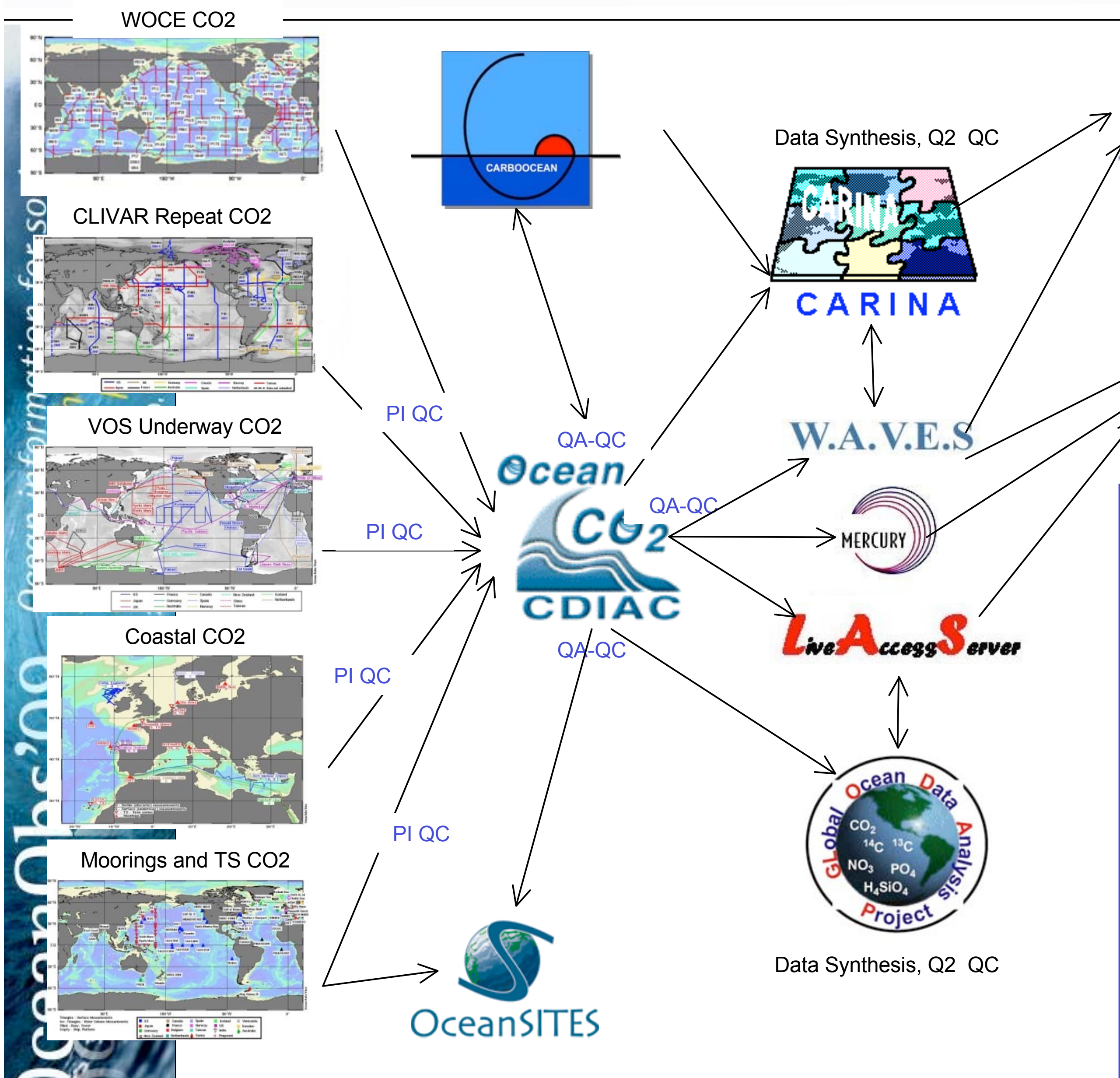




Number of users at US-GDAC since June 2006

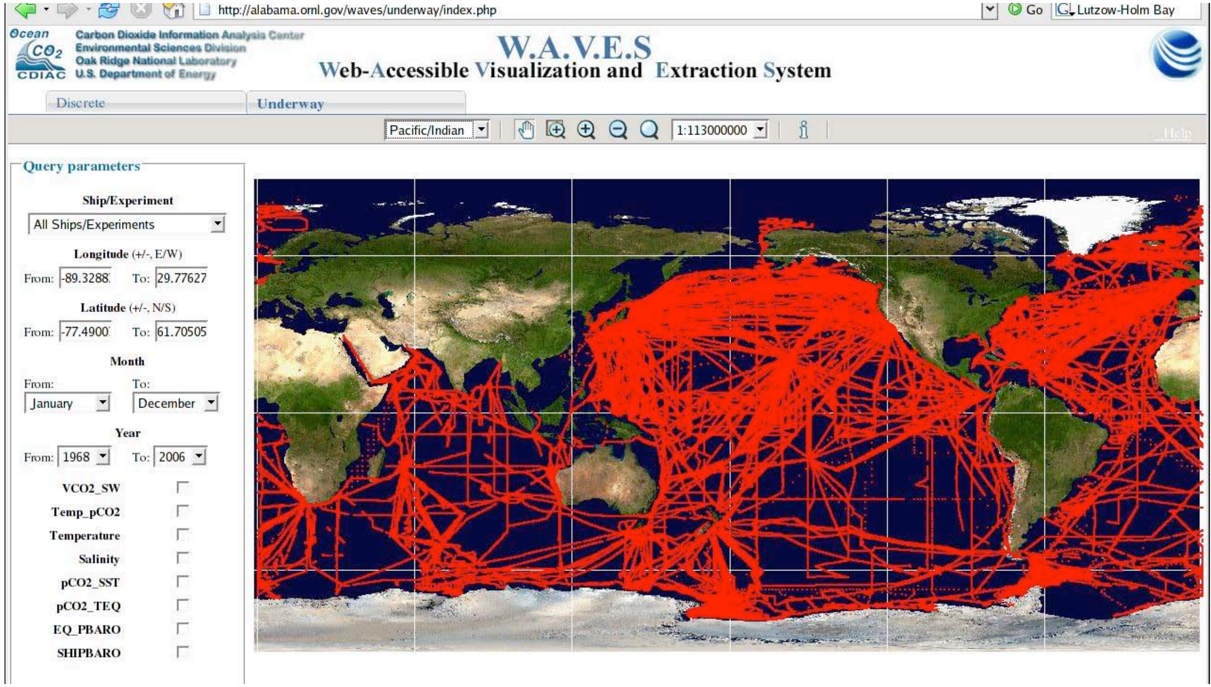
- **Total archive exceed 20TB from 1981 to present**
- **About 200 users downloading the data from GDAC every month, 1000 on the WWW**
- **2 GDACS, 11 RDACs, 1 Long term repository and reanalysis facility.**

In-Situ Carbon Data : CDIAC



- **Open data policy**
- **Processing made by Pis**
- **Centralized data holding**
- **Additional Quality control made by CDIAC to all data**
- **Discovering, viewing and downloading facilities (Mercury, WAVES, LAS, CDIAC web page)**
- **Data Synthesis, Global database**

Web interface to surface measurements

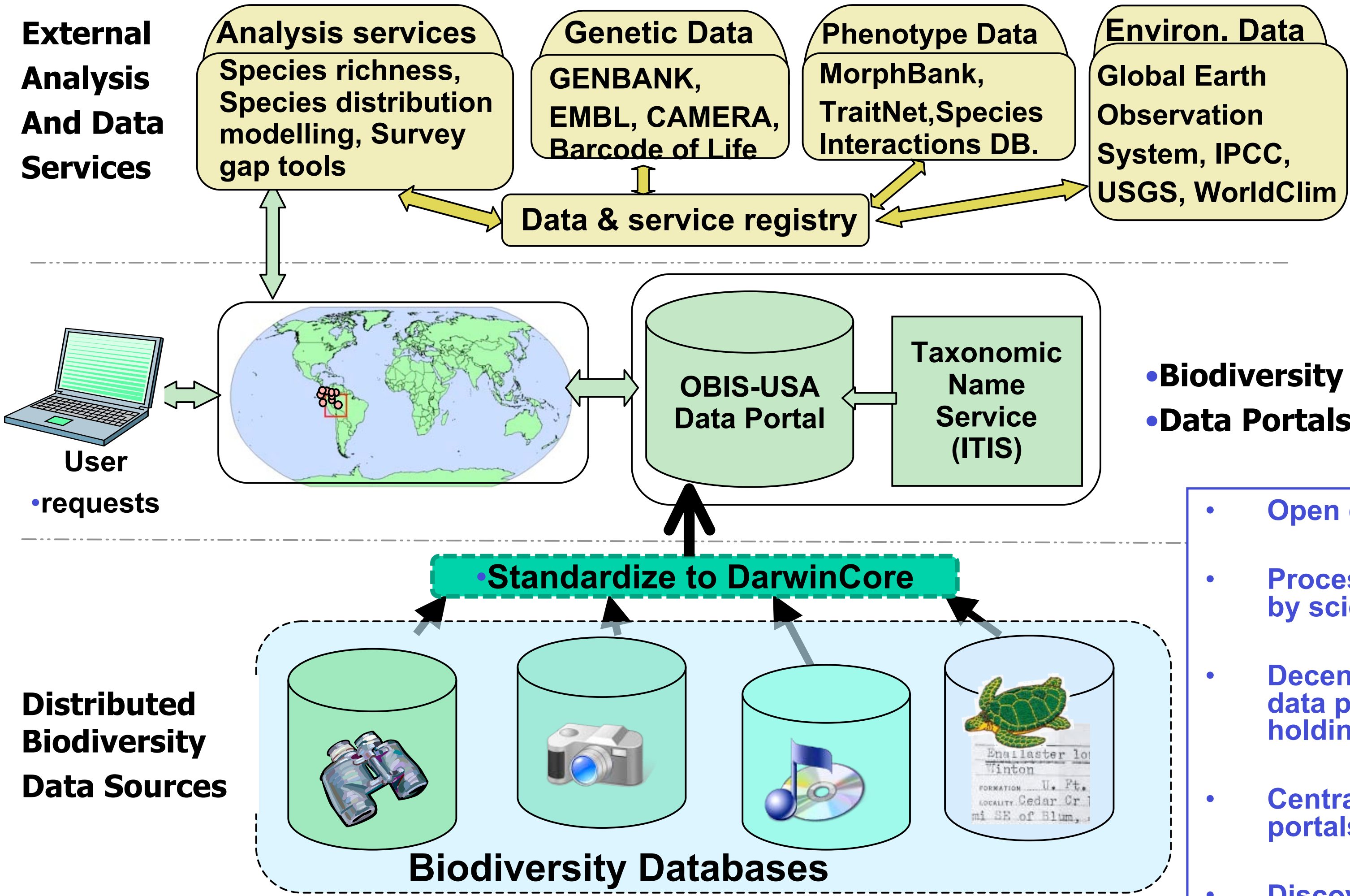


•LDEO Database V2007 consist 4.5 million surface measurements of CO_2

• Covered from 1960s to present time

• The CARBOOCEAN SOCAT database will be open through CDIAC in 2010 with 7.5 million surface measurements of CO_2

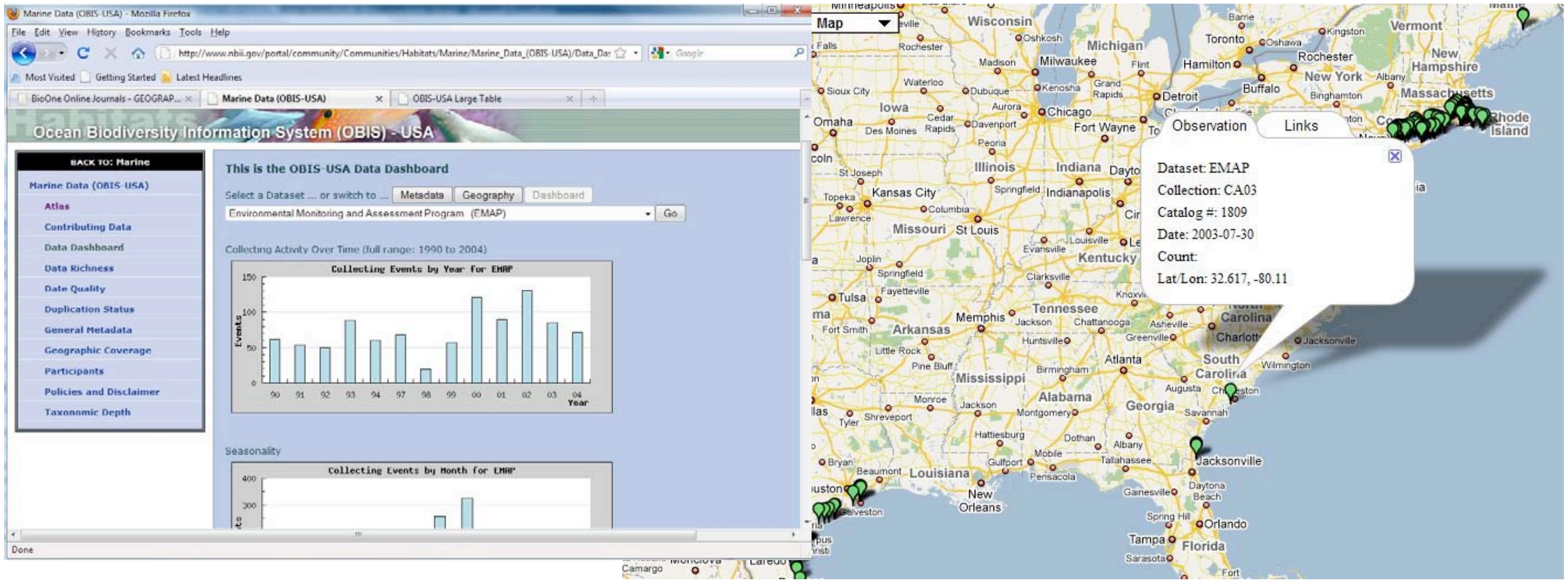
Biological In-situ Observations : IOBIS



- Open data policy
- Processing made by scientist
- Decentralized data processing & holdings
- Centralize data portals
- Discovering, viewing and downloading facilities
- Additional data and services

A Decade of improvements in Data Systems and Data Sharing

IOBIS and OBIS-USA



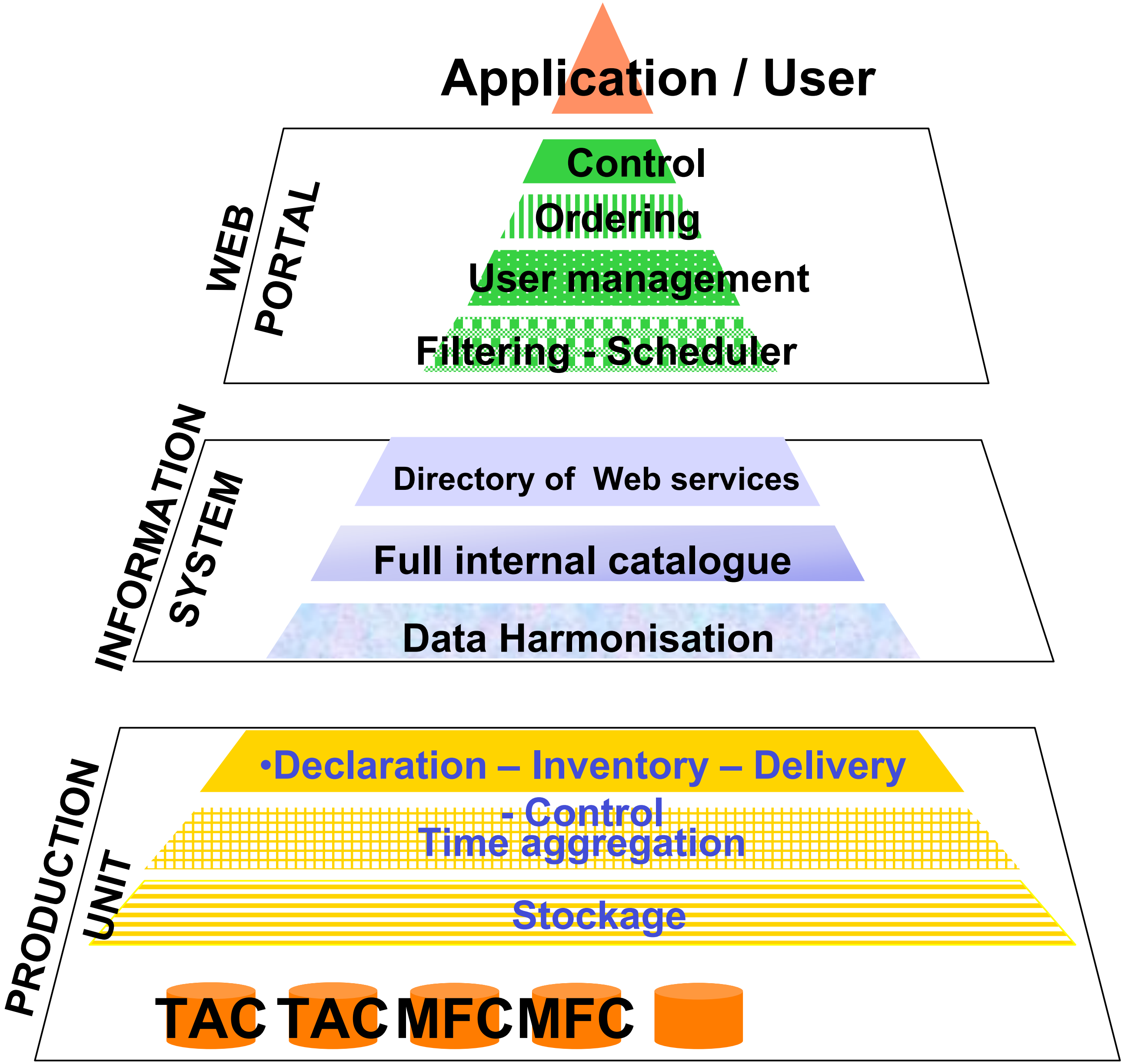
- **The Ocean Biogeographic Information System (OBIS) is composed of iOBIS and regional nodes (eg. OBIS-USA).**

- **iOBIS - 19.3 million records of 106 000 species from 660 distributed databases.**

- **OBIS-USA - 1.9 million records and 24,112 Taxa (by Scientific Name)**
  231,192 Distinct Locations, 40 Datasets
  130 Collections (representing distinct research activities)
  39 Investigators (this includes 'administrative' and 'technical' contacts)
  22 Institutions
  Date range: 1843 – 2008

- **iOBIS and OBIS-USA are both growing by leaps and bounds**

Building a Data Information System: A layered approach



InPresentation Level : Adapt to user needs (J Blower an S Hankin Talks)

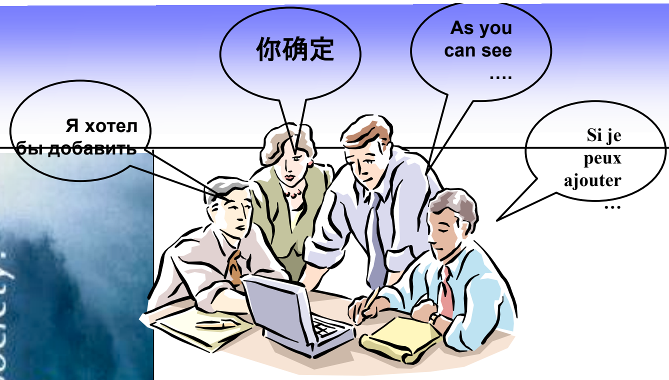
Interoperability Level : Have the appropriate plugs (J Blower Talk)

Data center level : speak the same language (B Keeley Talk)

What are the key elements of an efficient Data management network

- **Data Format and Metadata**

Data Format



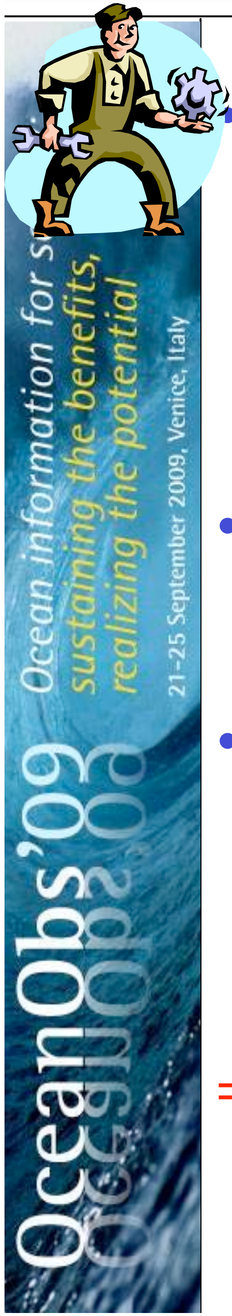
- **We need to agree on a common language**
  - ⇒ How can a user know that "subsurface temperature" is called TEMP in ARGO, Temperature in TAO, and is different from Temperature in GHRSST
  - ⇒ This is the purpose of metadata normalization,,
- **We need more than the measurements themselves:**
  - ⇒ Metadata that record the context of data acquisition (sensor, experiment, data centre, PI,…)
  - ⇒ Raw data and corrected data to enable future reprocessing
  - ⇒ Common Quality flags that characterize the data
  - ⇒ History of what has been done on the data , what are the processing steps they have gone through, Calibration information if any.

What are the key elements of an efficient Data management network

- **Data Quality: Quality depends on the time you have to "clean the data "**

Can take years !!!!

Data quality: the cleaner you want the longer it takes!!!!!



- **It's Fundamental because**
  - ⇒accepting erroneous data can cause erroneous conclusions
  - ⇒BUT rejecting extreme good data can lead to miss important events …
- **It's a challenge because no « ground truth » really exists**
- **For forecast data must be delivered with one day.. For reanalysis or climate change studies users request higher quality data set and even corrected data if possible.**

⇒**2 steps: Real time and delayed mode QC**

What did we prove in past decade

- **Open data policy** is an asset for everybody (scientists, data managers , policy makers, …)
- **Information technology and Data management** capabilities are **no more an obstacle** to data access and sharing even in real time
- It's important to **properly archive with the data all the necessary metadata information** for future generations in a common language
- **Building on common vocabularies, formats, distribution means** , even if they are not yet stamped as standards is essential to build a reliable system => **Need to set up mechanism to collaborate with standardization bodies and move to the standards** when stable
- **Feedback from users** is important and **helps improving the quality** of the data

Conclusions & Next steps for future

- **Accessible: free access to essential data**

- **Comparable agreement on quality control procedures both in real-time and delayed mode to provide datasets independent of what platform sampled it.**

- **Understandable common standards for data description and distribution. Same syntax and semantic.**

- **Recognized/cited in scientific papers that use them**

- **Moving from Observations to information, use of in-situ/ satellite/ model outputs together to provide services to users**

Yes it's possible

YES I WILL !