Exploring Persistent Identifiers for Open Time Series

Results from the joint
COOPEUS, ENVRI & EUDAT
workshop on *„persistent identifiers for open time series“*
hosted by COOPEUS
Bremen, 25-26.6.2013

Robert Huber
Universität Bremen, MARUM

COOPEUS ENVRI EUDAT

Motivation

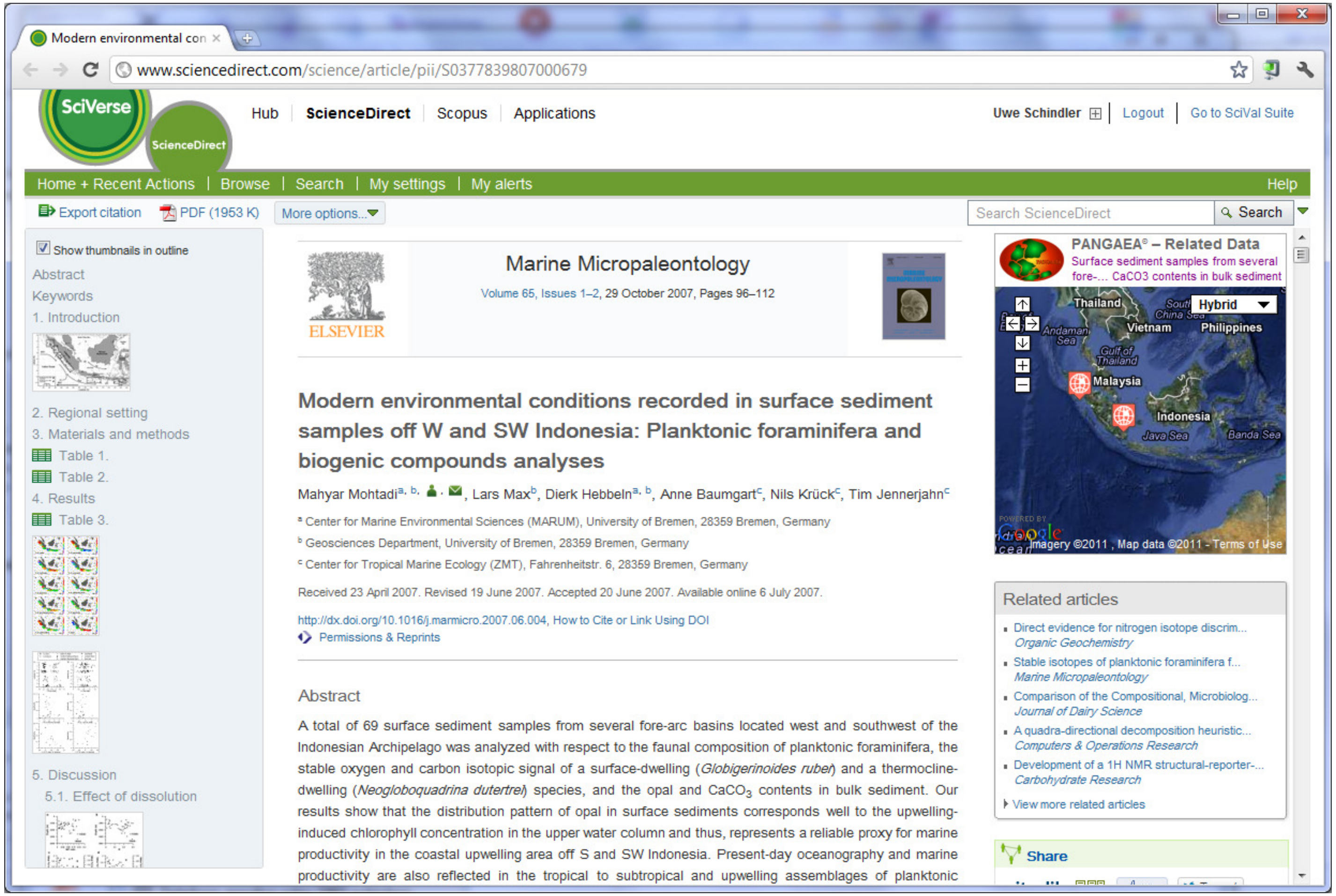
*“**A major prerequisite for the proper use of persistent identifiers (PID) e.g. within data citations is the persistence of both, identifiers as well as the integrity of the associated data set**.*

*This poses questions **when PIDs are to be used for unfinished data sets or open time series data**.*

*Such data is **typically generated within research infrastructures** (RI) during long lasting experiments such as satellite missions, environmental monitoring campaigns, or in permanent installations such as natural hazard detection and early warning systems.*

*Open time series data are **often used in research during ongoing experiments** and potentially published earlier than the underlying data set has been closed and is publicly released. “*

Data citation



Marine Micropaleontology

Volume 65, Issues 1–2, 29 October 2007, Pages 96–112

Modern environmental conditions recorded in surface sediment samples off W and SW Indonesia: Planktonic foraminifera and biogenic compounds analyses

Mahyar Mohtadi[a, b], Lars Max[b], Dierk Hebbeln[a, b], Anne Baumgart[c], Nils Krück[c], Tim Jennerjahn[c]

[a] Center for Marine Environmental Sciences (MARUM), University of Bremen, 28359 Bremen, Germany

[b] Geosciences Department, University of Bremen, 28359 Bremen, Germany

[c] Center for Tropical Marine Ecology (ZMT), Fahrenheitstr. 6, 28359 Bremen, Germany

Received 23 April 2007. Revised 19 June 2007. Accepted 20 June 2007. Available online 6 July 2007.

http://dx.doi.org/10.1016/j.marmicro.2007.06.004, How to Cite or Link Using DOI

Permissions & Reprints

Abstract

A total of 69 surface sediment samples from several fore-arc basins located west and southwest of the Indonesian Archipelago was analyzed with respect to the faunal composition of planktonic foraminifera, the stable oxygen and carbon isotopic signal of a surface-dwelling (*Globigerinoides ruber*) and a thermocline-dwelling (*Neogloboquadrina dutertrei*) species, and the opal and $CaCO_3$ contents in bulk sediment. Our results show that the distribution pattern of opal in surface sediments corresponds well to the upwelling-induced chlorophyll concentration in the upper water column and thus, represents a reliable proxy for marine productivity in the coastal upwelling area off S and SW Indonesia. Present-day oceanography and marine productivity are also reflected in the tropical to subtropical and upwelling assemblages of planktonic

PANGAEA® – Related Data

Surface sediment samples from several fore-... CaCO3 contents in bulk sediment

Related articles

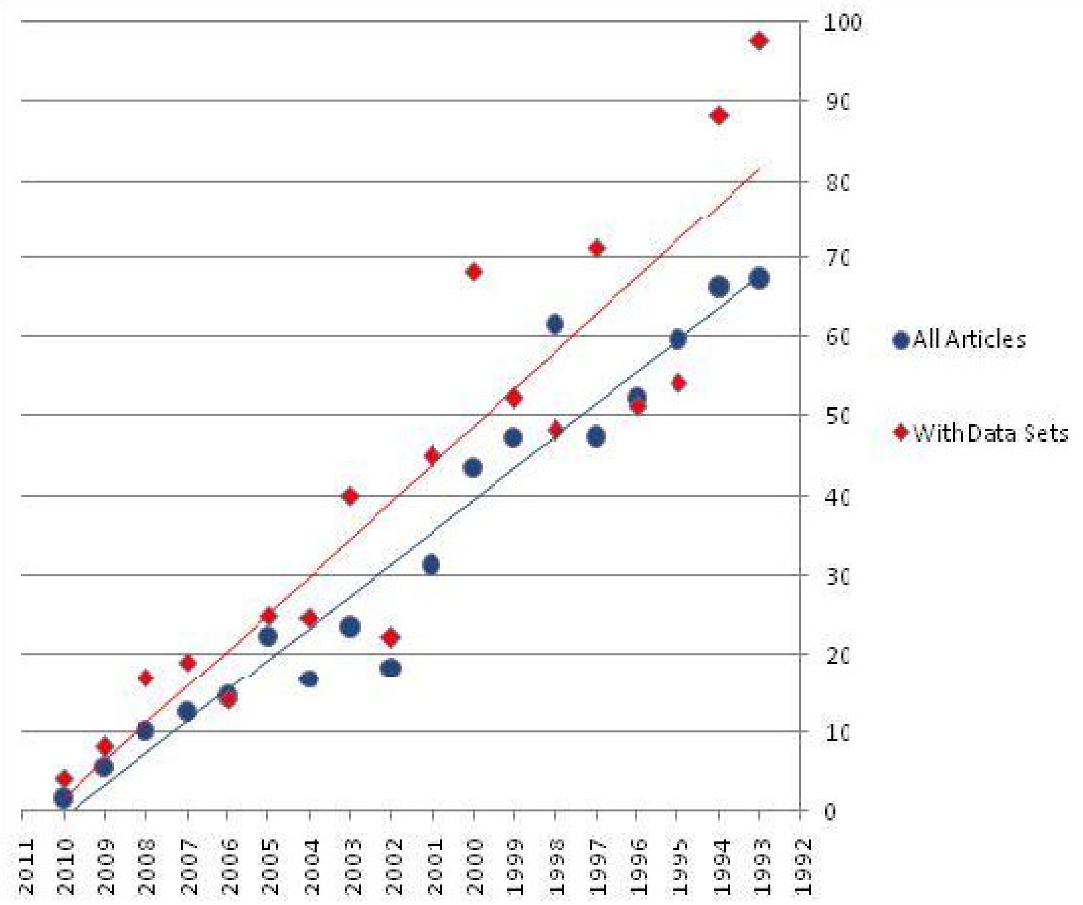
- Direct evidence for nitrogen isotope discrim... *Organic Geochemistry*
- Stable isotopes of planktonic foraminifera f... *Marine Micropaleontology*
- Comparison of the Compositional, Microbiolog... *Journal of Dairy Science*
- A quadra-directional decomposition heuristic... *Computers & Operations Research*
- Development of a 1H NMR structural-reporter-... *Carbohydrate Research*

View more related articles

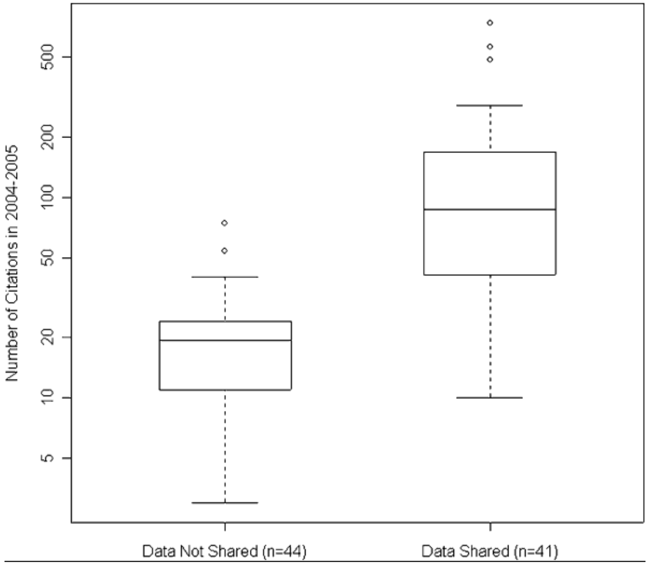
Share

COOPEUS ENVRI EUDAT

Benefits of data citation



courtesy of Jon Sears (AGU)



Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Background

- ENVRI: EMSO/ARGO standardisation meeting: the ARGO – DOI problem
- Probably many other intra project meetings (e.g. EUDAT EPOS)
- COOPEUS, ENVRI and EUDAT ***strategic workshop on future harmonization of data sharing among Research Infrastructures*** (EGU 2013)
  - Identification of PIDs as a common challenge
  - White paper draft, RI case studies and strategies
- Joint COOPEUS, ENVRI and EUDAT ***workshop on persistent digital identifiers (PID) for open time series data***

Workshop participants

Euro Argo

european multidisciplinary seafloor observatory emso

EPOS EUROPEANPLATEOBSERVINGSYSTEM

icos integrated carbon observation system

neon National Ecological Observatory Network, Inc.

LifeWatch

DataCite International Data Citation

EPIC

RDA RESEARCH DATA ALLIANCE

ICSU WORLD DATA SYSTEM

COOPEUS

ENVRI

EUDAT

Topics, goals

- Research Infrastructure case studies: PID usage status quo and strategies
- Discuss best practices for PIDs for open time series
- The 10 golden rules for the selection and use of PIDs for open time series

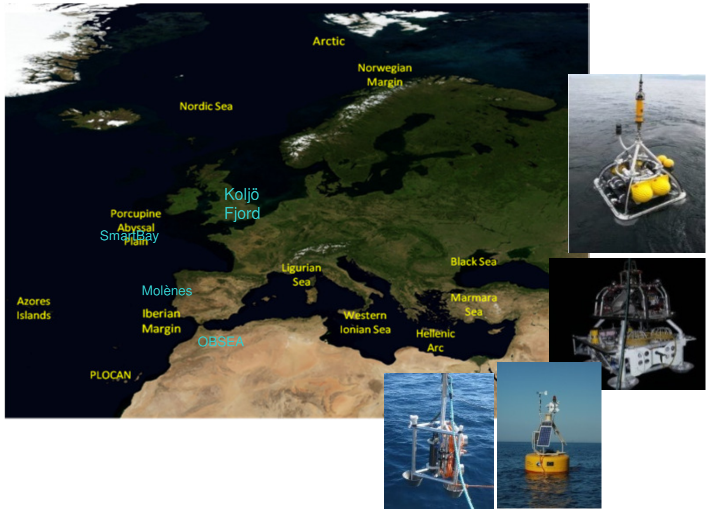
COOPEUS ENVRI EUDAT

Case study: EMSO

european multidisciplinary seafloor observatory emso

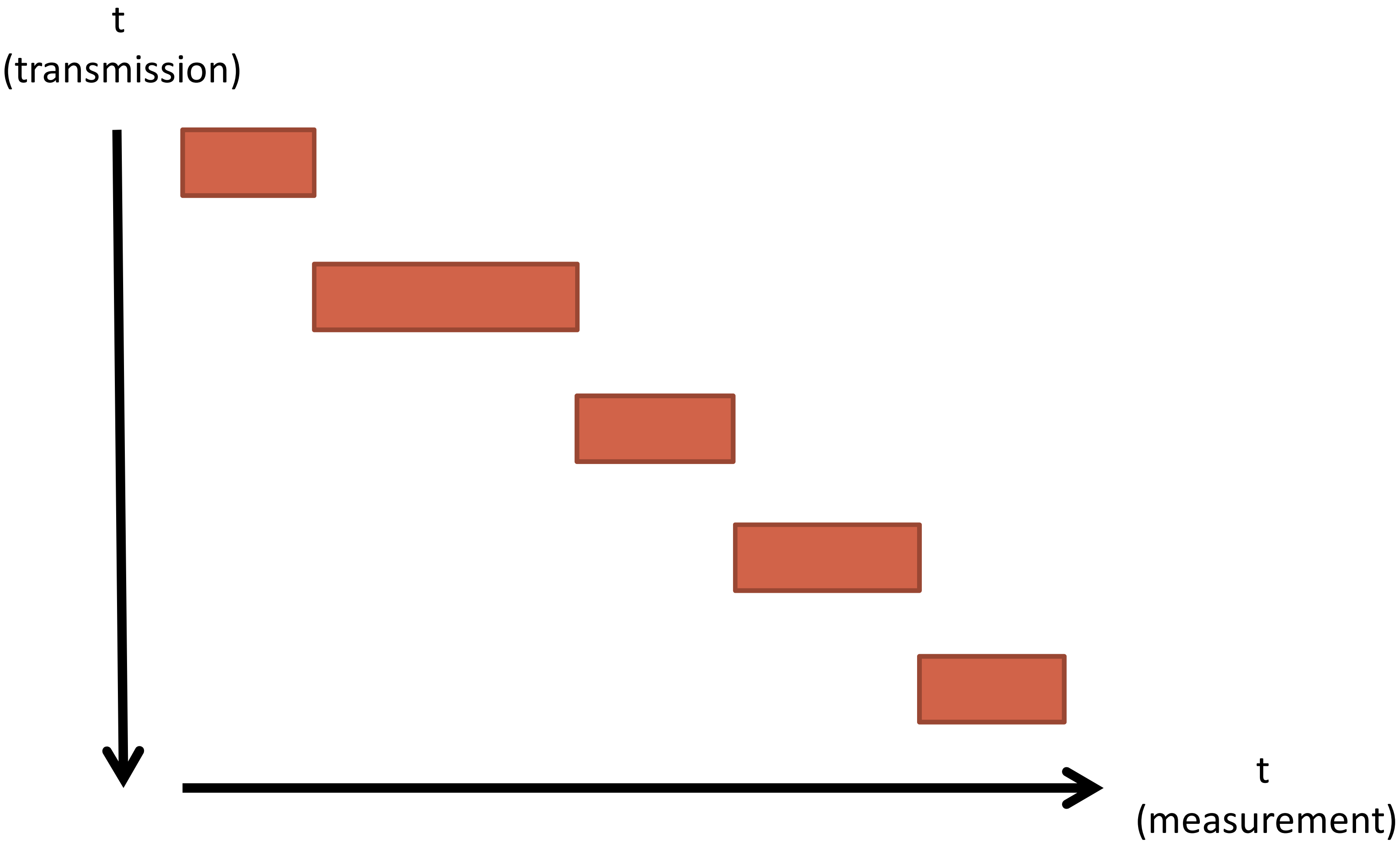
PANGAEA

- network of **fixed point, deep sea observatories**
- real-time, long-term monitoring of environmental processes
- multidisciplinary: geosphere, biosphere and hydrosphere



“The data is automatically uploaded [e.g. via OGC SOS] to the PANGAEA import queue and subsequently archived as raw data in a monthly interval.
However, there is a growing demand to accelerate both, citability and availability of open time series data. Therefore PANGAEA is also investigating additional strategies to handle such data. ”

EMSO case: growing dataset

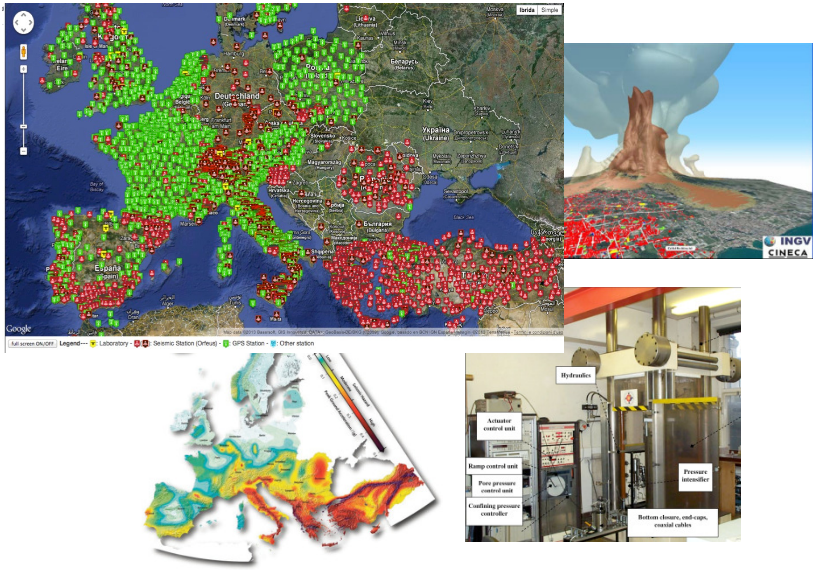


Dataset (T_5)

Case study: EPOS

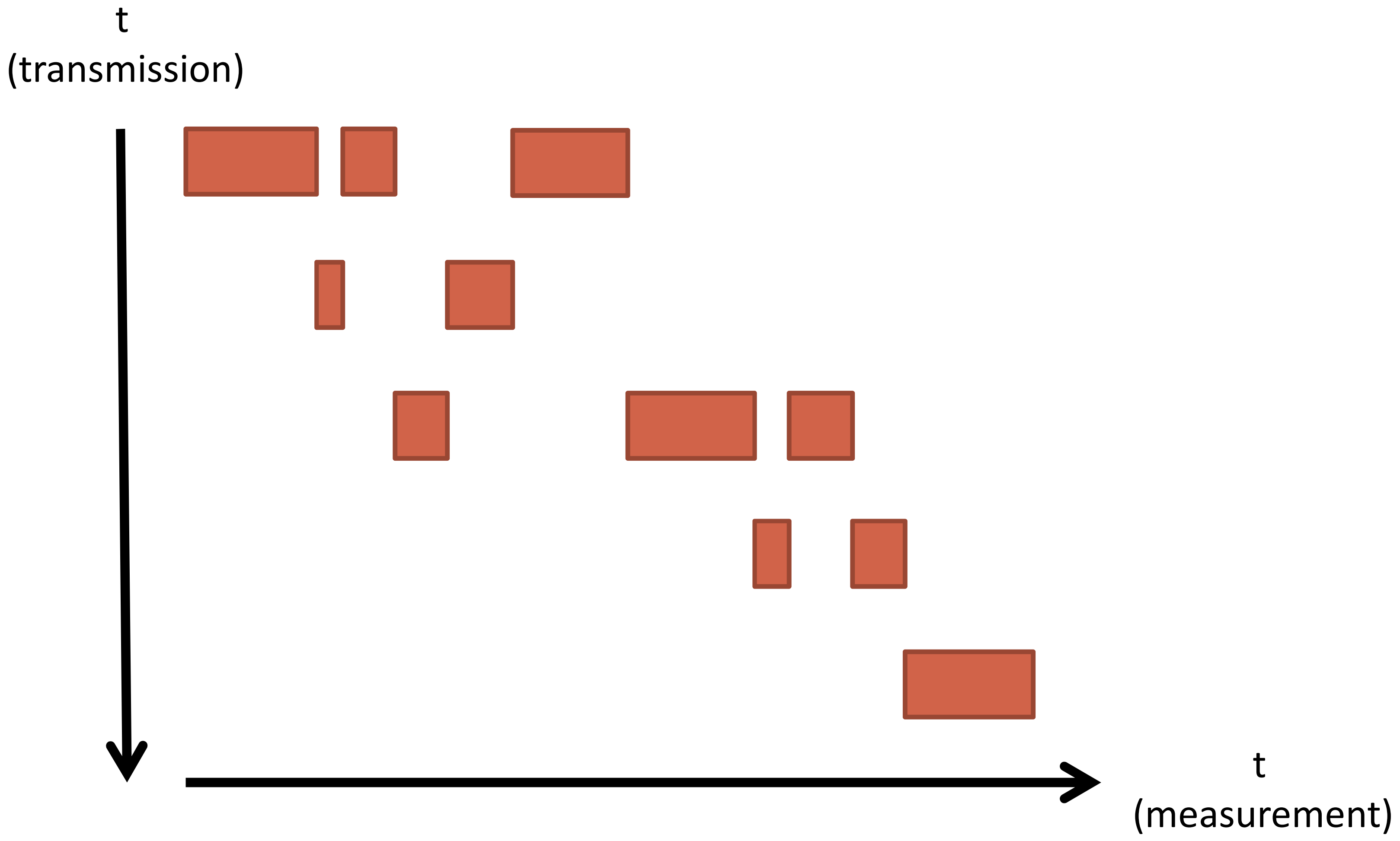
EPOS
EUROPEAN PLATE OBSERVING SYSTEM

- geophysical monitoring networks
- local observatories (including permanent
- in-situ and volcano observatories)
- experimental laboratories in Europe


INGV
CINECA
Hydraulics
Actuator control unit
Ramp control unit
Pore pressure control unit
Confining pressure controller
Pressure intensifier
Bottom closure, end-caps, coaxial cables

“One of the problems encountered by the community when seeking to uniquely identify the data digital objects is the incompleteness of the data acquired. This problem follows from data transmission from the remote station to the central data center and it consists of the presence of data gaps. These gaps are then filled in as the bandwidth of the transmission widens”

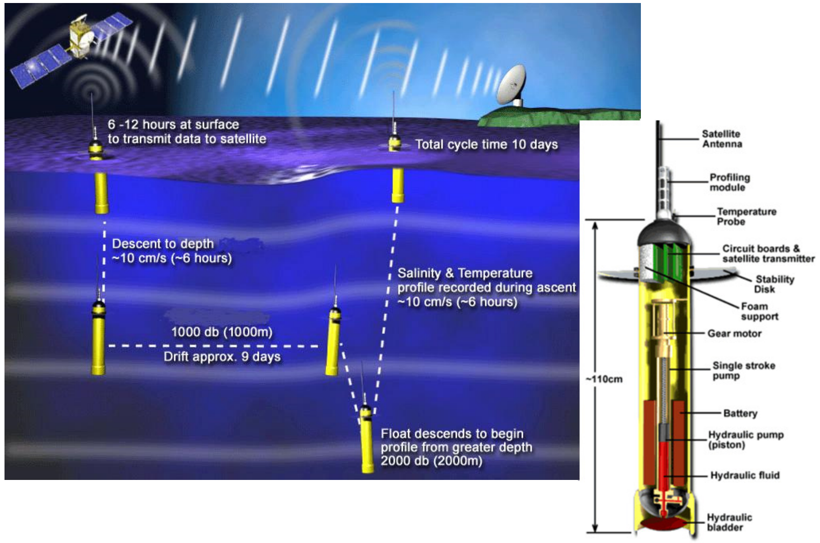
EPOS case: fragmented dataset



Dataset (T_5)

Case study: ARGO

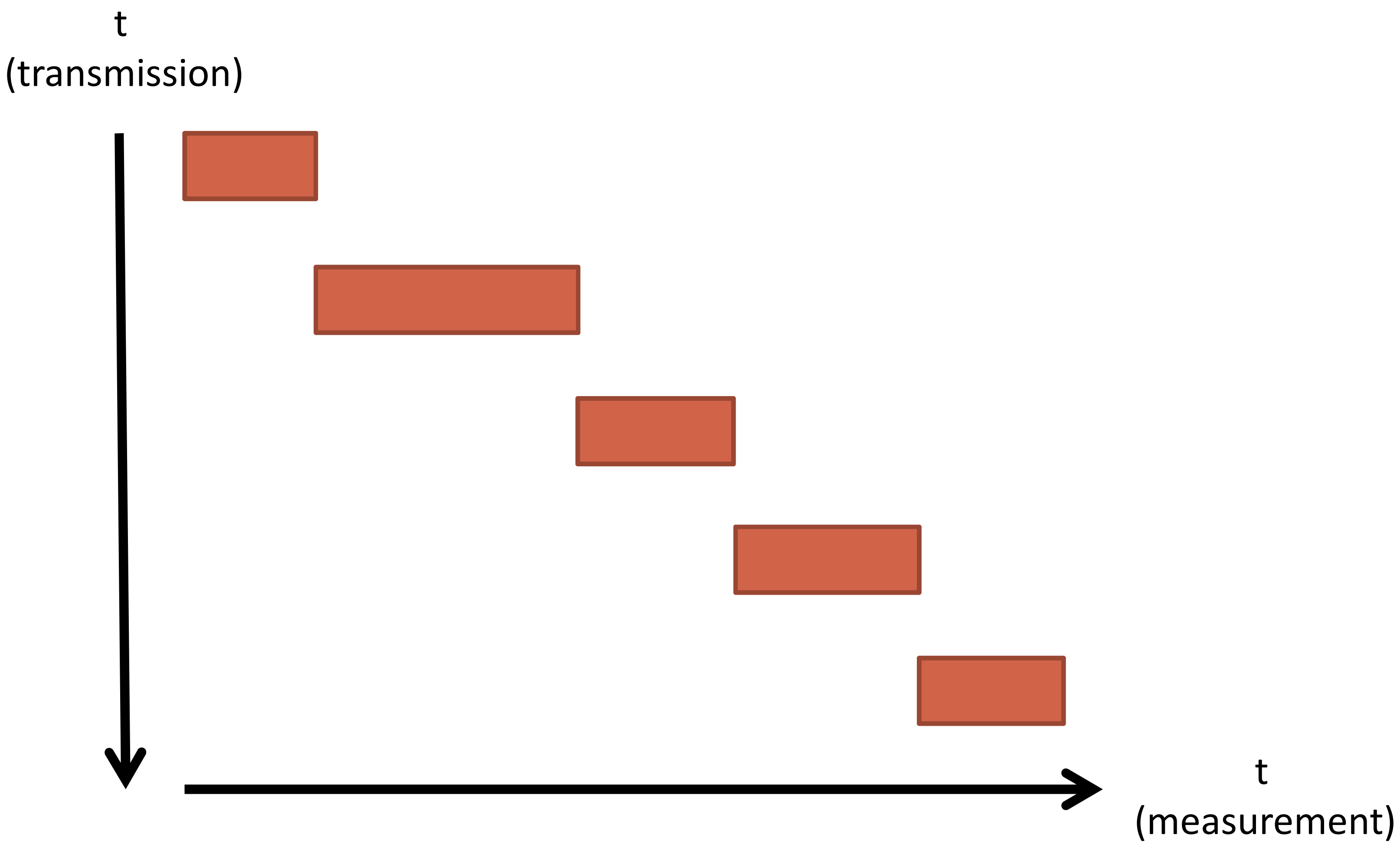
- European contribution to ARGO
- array of autonomous instruments (argo floats) deployed over the world ocean
- subsurface ocean properties: temperature and salinity over the upper 2000m of the ocean



"[…] *data from each profiler are reviewed and checked against climatological data and nearby Argo data from different profiler […]. The complication in Argo is constant mutation of the data on GDACs*. This is both through the temporal extension of the data when new profiles are collected and updates to existing data when delayed mode quality control is done."*

*Global Data Assembly Center

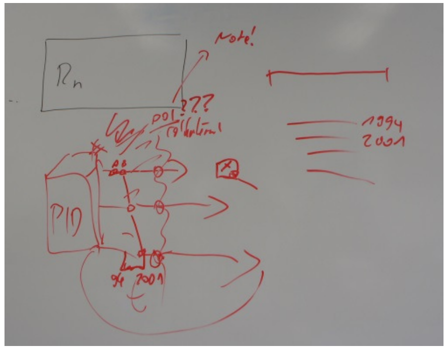
ARGO case: mutating dataset



Dataset (T_5)

Workshop results: Common requirements

- Existing technologies sufficient (handle, DOI, EPIC..)
- … given that some (additional) requirements are fulfilled
  - Fragmentation support
  - Integrity (e.g hash tag, but community specific)
  - Versioning support
  - Aggregation / Relation support
  - Notion of time as attribute



Workshop results:
The golden rules…

1. Persistence: Each datacenter must define a versioning and preservation strategy
2. PIDs must be persistent, even when datasets are deleted or changed
3. PIDs must be organized according to its use ... Publication vs. Data management
4. Time-fragmentation support (resolution).
5. Transparency: level of dynamicity in the data-set must be defined in PID.(e.g. **growing** ,evolving, **fragmented)**
6. Procedure for PID generation must be consistent, transparent , documented and financial affordable
7. PIDs should be assigned early as possible ...
8. Levels of granularities must be standardized within each scientific field
9. Data center must provide a citation template

Workshop results:
Metadata requirements

1. Level of dynamicity in the data-set.
2. Include timestamp to identify „*version*“ (identify time relative to changes to the dataset)
3. Fragment identification, relation
4. Content of request selection used (the query..)
5. Creation date of the whole timeseries (in addition to publication date)

Workshop results: Citation rules:

<author> . **(<release date range>)**: <*dataset title*>. [version: <version>|subset: <temporal range>]. <publisher>.[[**<resource type (growing dataset , evolving dataset , fragmented dataset)>]].** <PID>@<fragment identifier>. [accessed: <access date>]

Examples:

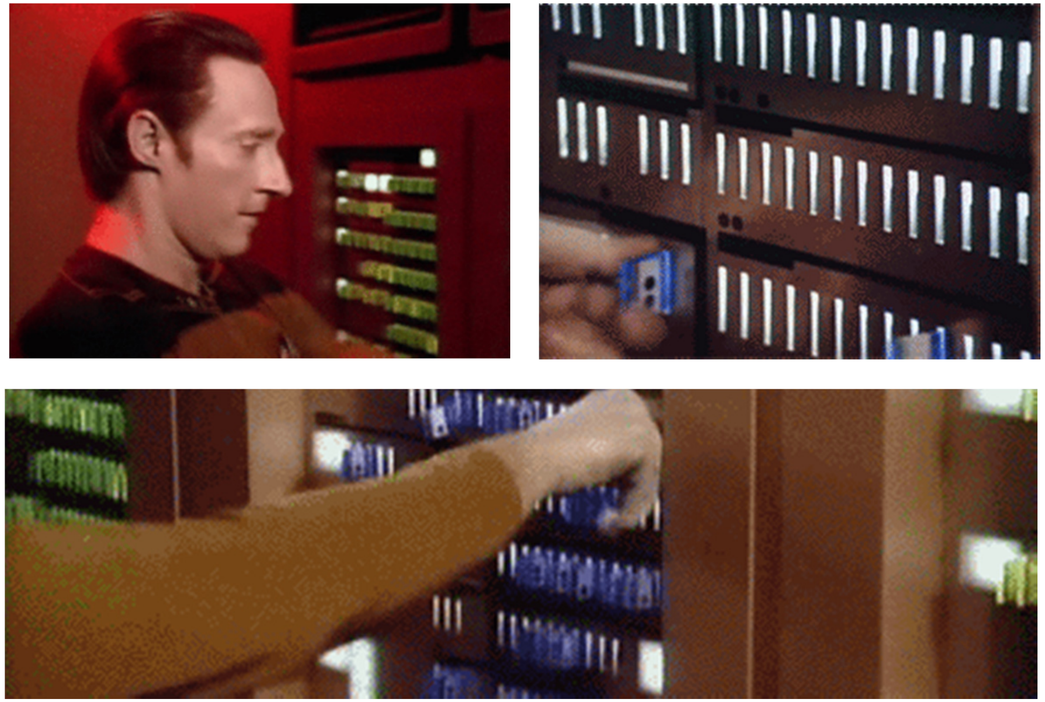
- Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. version: 1.2. Responsible Data Archive. [**evolving dataset]. ** PID:123456789@version=1.2
- Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. subset: 2010-01-01 - 2010-12-13. Responsible Data Archive. **[growing dataset].** PID:123456789@range=2010-01-01-2010-12-13
- Doe, J. **(2009-2011)**: *Dynamic Data Set Title*. version: 1.2. Responsible Data Archive. [**fragmented dataset].** PID:123456789. accessed: 2012-12-01@version=1.2

COOPEUS ENVRI EUDAT

Future work

- White paper on PID s for open time series
- Contribution to RDA working group on data citation (Ari)

Contact: rhuber@uni-bremen.de

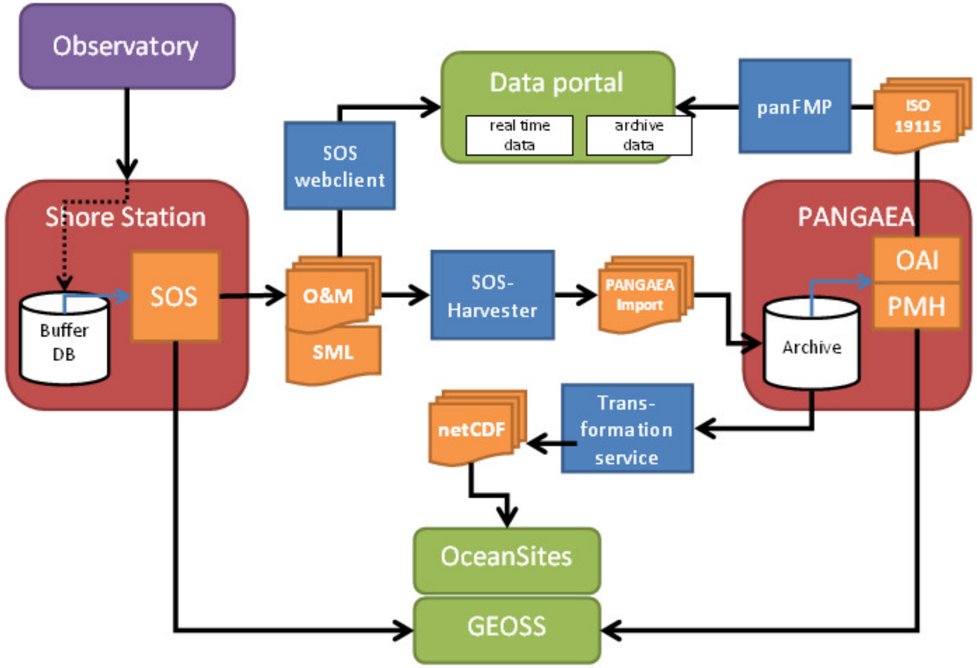


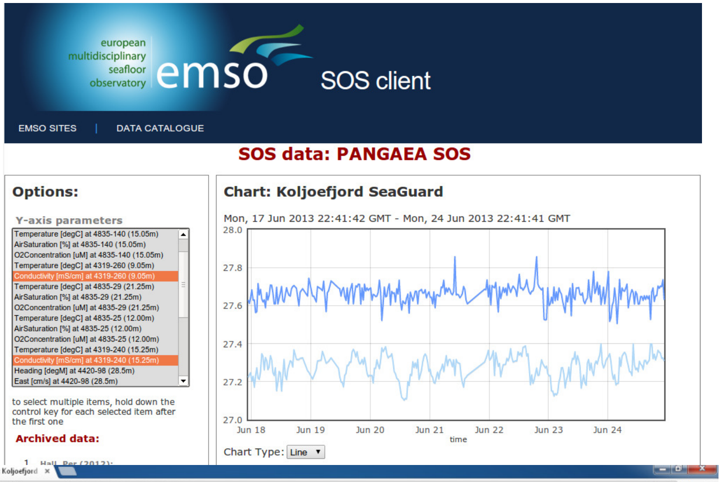
Thank you..

Koljoefjord observatory

- operated by the University of Gothenburg (Per Hall et al.) & MARUM
- has been operating for about 2 years
- EMSO test site
- Cabled, multi-sensor underwater observation systemMain node and land station connected to the Internet via 3G
- Real-time access and remote control
- Data archived in PANGAEA

Koljoefjord II




european multidisciplinary seafloor observatory
emso
SOS client
EMSO SITES | DATA CATALOGUE

SOS data: PANGAEA SOS

Options:

Y-axis parameters

- Temperature [degC] at 4835-140 (15.05m)
- AirSaturation [%] at 4835-140 (15.05m)
- O2Concentration [uM] at 4835-140 (15.05m)
- Temperature [degC] at 4319-260 (9.05m)
- Conductivity [mS/cm] at 4319-260 (9.05m)
- Temperature [degC] at 4835-29 (21.25m)
- AirSaturation [%] at 4835-29 (21.25m)
- O2Concentration [uM] at 4835-29 (21.25m)
- Temperature [degC] at 4835-25 (12.00m)
- AirSaturation [%] at 4835-25 (12.00m)
- O2Concentration [uM] at 4835-25 (12.00m)
- Temperature [degC] at 4319-240 (15.25m)
- Conductivity [mS/cm] at 4319-240 (15.25m)
- Heading [degM] at 4420-98 (28.5m)
- East [cm/s] at 4420-98 (28.5m)

to select multiple items, hold down the control key for each selected item after the first one

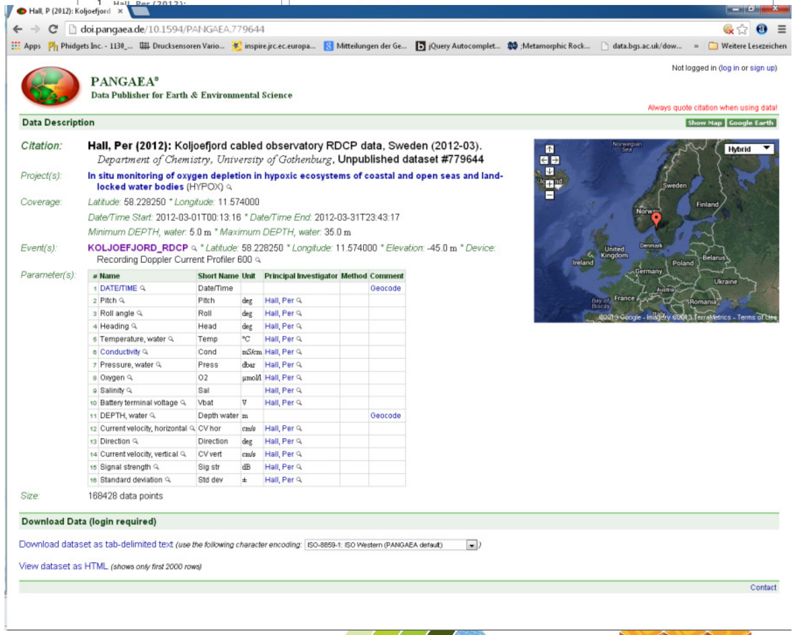
Archived data:

1 Hall, Per (2012):

Chart: Koljoefjord SeaGuard

Mon, 17 Jun 2013 22:41:42 GMT - Mon, 24 Jun 2013 22:41:41 GMT

Chart Type: Line ▾



Not logged in (log in or sign up)

PANGAEA®
Data Publisher for Earth & Environmental Science

Always quote citation when using data!

Data Description

Citation: **Hall, Per (2012):** Koljoefjord cabled observatory RDCP data, Sweden (2012-03). *Department of Chemistry, University of Gothenburg*, Unpublished dataset #779644

Project(s): **In situ monitoring of oxygen depletion in hypoxic ecosystems of coastal and open seas and land-locked water bodies** (HYPOX)

Coverage: *Latitude:* 58.228250 * *Longitude:* 11.574000
Date/Time Start: 2012-03-01T00:13:16 * *Date/Time End:* 2012-03-31T23:43:17
Minimum DEPTH, water: 5.0 m * *Maximum DEPTH, water:* 35.0 m

Event(s): KOLJOEFJORD_RDCP * *Latitude:* 58.228250 * *Longitude:* 11.574000 * *Elevation:* -45.0 m * *Device:* Recording Doppler Current Profiler 600

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	DATE/TIME	Date/Time				Geocode
2	Pitch	Pitch	deg	Hall, Per		
3	Roll angle	Roll	deg	Hall, Per		
4	Heading	Head	deg	Hall, Per		
5	Temperature, water	Temp	°C	Hall, Per		
6	Conductivity	Cond	mS/cm	Hall, Per		
7	Pressure, water	Press	dbar	Hall, Per		
8	Oxygen	O2	µmol/l	Hall, Per		
9	Salinity	Sal		Hall, Per		
10	Battery terminal voltage	Vbat	V	Hall, Per		
11	DEPTH, water	Depth water	m			Geocode
12	Current velocity, horizontal	CV hor	cm/s	Hall, Per		
13	Direction	Direction	deg	Hall, Per		
14	Current velocity, vertical	CV vert	cm/s	Hall, Per		
15	Signal strength	Sig str	dB	Hall, Per		
16	Standard deviation	Std dev	±	Hall, Per		

Size: 168428 data points

Download Data (login required)

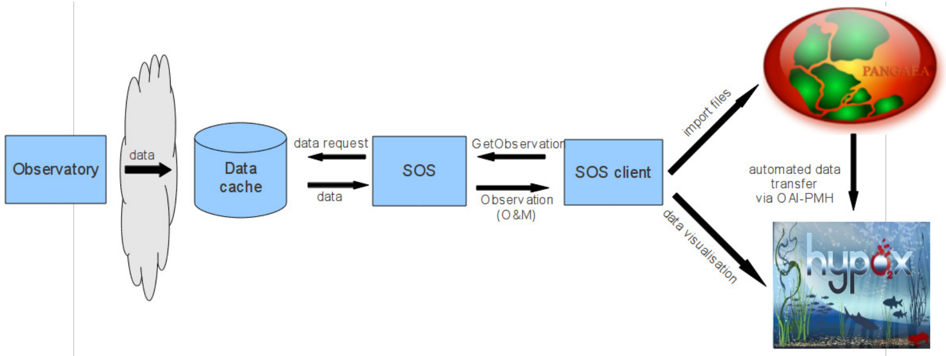
Download dataset as tab-delimited text *(use the following character encoding:* ISO-8859-1: ISO Western (PANGAEA default) ▾ *)*

View dataset as HTML *(shows only first 2000 rows)*

Contact

COOPEUS ENVRI EUDAT

Koljoefjord III



SOS = Sensor Observation Service
O&M = Observations and Measurements
OAI-PMH = Open Archives Initiative – Protocol for Metadata Harvesting

Automatic workflow:

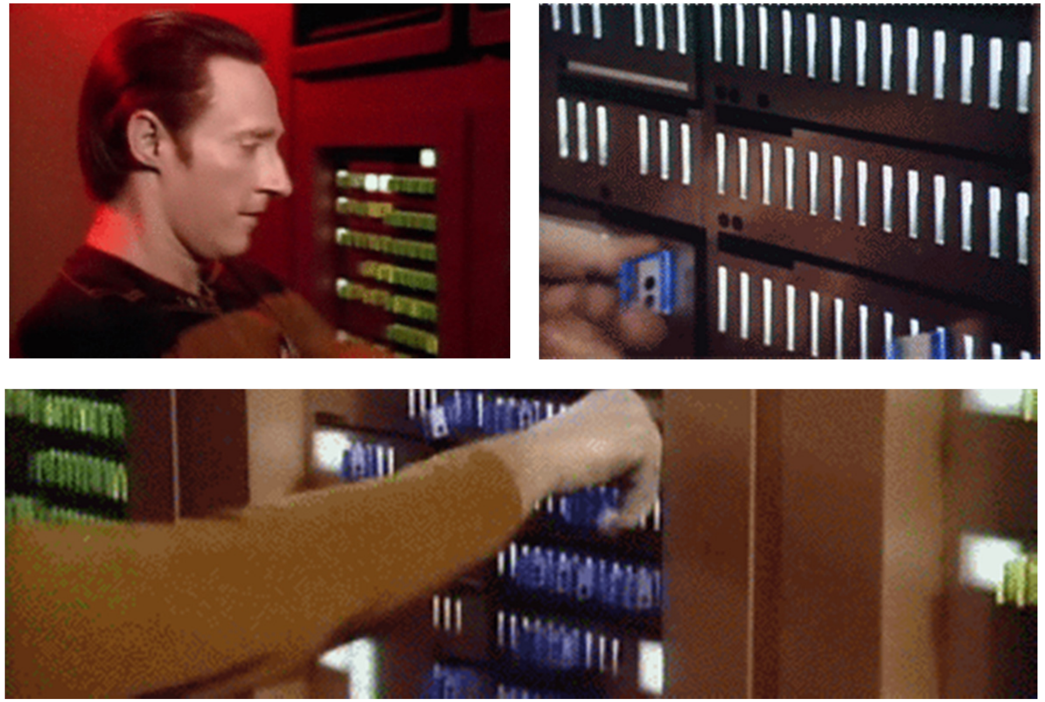
- Data parsers are waiting for data to be pushed by the instruments (cache)
- SOS client will connect to SOS and request data periodically (GetObservation)
- SOS client will generate import files from retrieved data
- Import files are used to persistently store data in PANGAEA
- ---> Standardised workflow
- ---> Interoperability achieved by implementing OGC standards
- ---> Data providers do not have to worry about submitting their data

COOPEUS ENVRI EUDAT

Koljoefjord PIDs

Task: decide how to archive open time series data and make it citable by assigning unique PID

- Possible strategies:
  - one open dataset that gets constantly filled up, identifiable by one DOI
  - or: split up data into parcels of defined temporal granularity, assign DOI to each parcel
- Decision: monthly datasets, because..
  - It suited the data owners
  - Possibility to compare observatory data to data collected monthly in the same area by the Swedish Hydrological Institute (quality checks!)
  - monthly datasets easier to handle than very large monolithic datasets
- But what about would a monthly opendata set?
  - See PID workshop requirements
  - Add parameters to DOIs?



Thank you..

COOPEUS ENVRI EUDAT

Case study results: PID assignment strategies

- Placeholder strategies:
  - PID on abstract or initial data set (e.g. initially empty)
  - PID on delegate document (e.g. data QC handbooks, readmes)
  - PID on data product (e.g. images)
- Versioning strategies:
  - New version after reprocessing
  - New version after update
- Fragmenting strategies:
  - Define appropriate subsets (e.g. monthly)

COOPEUS ENVRI EUDAT