~~Earth~~ Science in the Data Era

How data is changing the way we do science

Summary

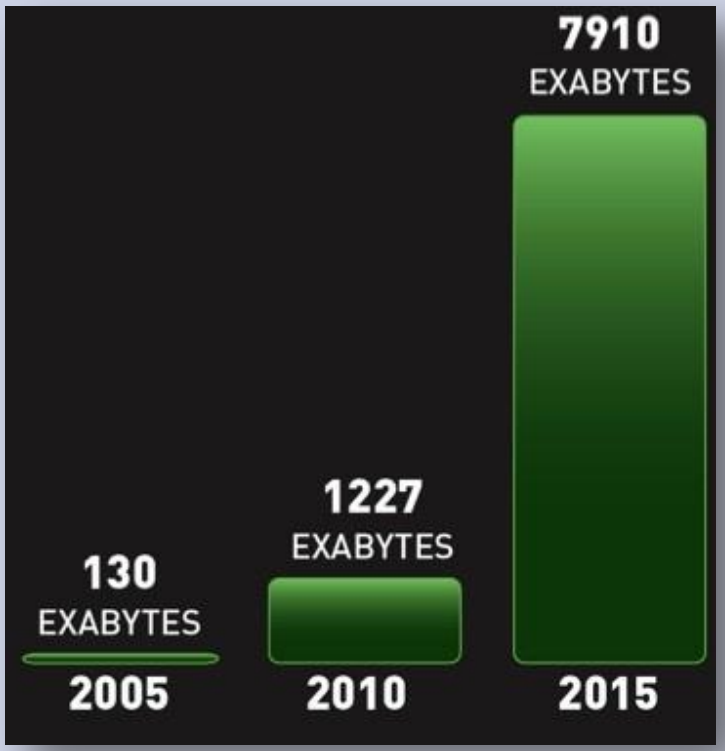
- Data & Science
- The Earth Science Case

Data & Science

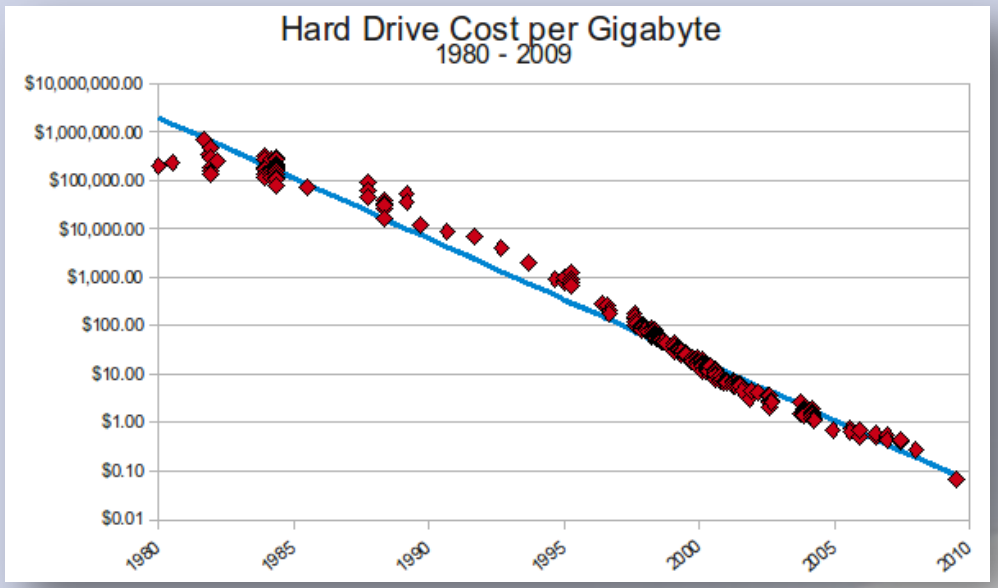
Where are we?

Where to go next?

Data Grow & Storage Cost

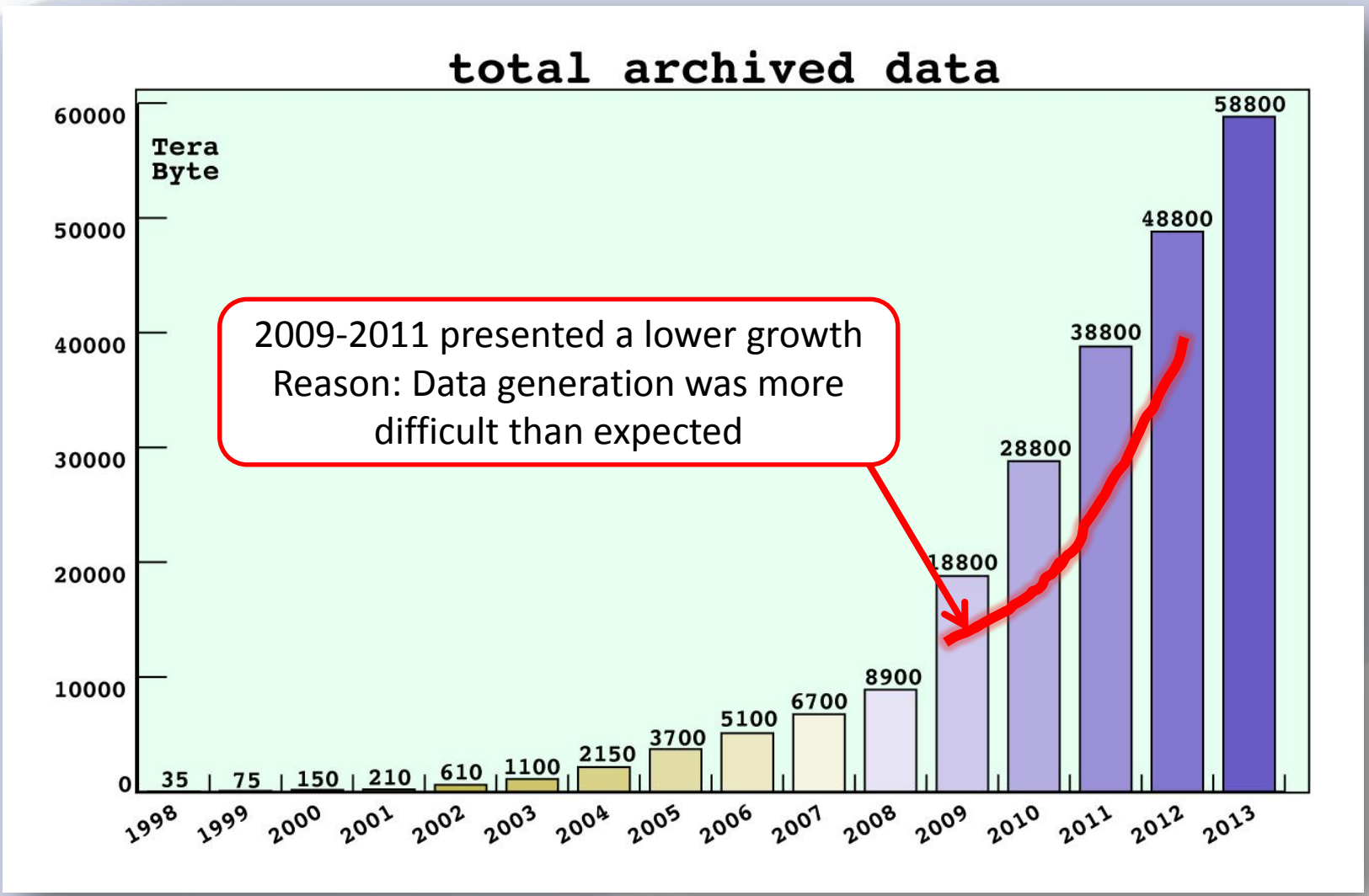


IDC's Digital Universe Study, sponsored by EMC, June 2011



http://www.mkomo.com/cost-per-gigabyte

Expected growth rate for data archive @ DKRZ (real case)



Data Management Challenge

- "[…] the number of files the datacenter will have to deal with will grow by a factor of **75**, at least. Meanwhile, the number of IT professionals in the world will grow by less than a factor of **1.5**"

IDC's Digital Universe Study, sponsored by EMC - http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf - 29.05.2012

And Security, of course…

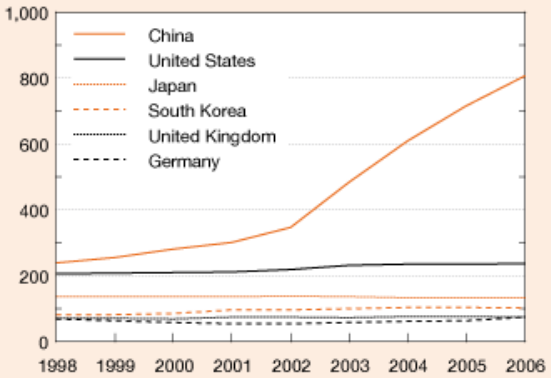
- “[…] only about half the information that should be protected is protected.”

IDC’s Digital Universe Study, sponsored by EMC - http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf - 29.05.2012

Scientific Community Growth

Figure O-8
First university degrees in natural sciences and engineering, selected countries: 1998–2006

Thousands



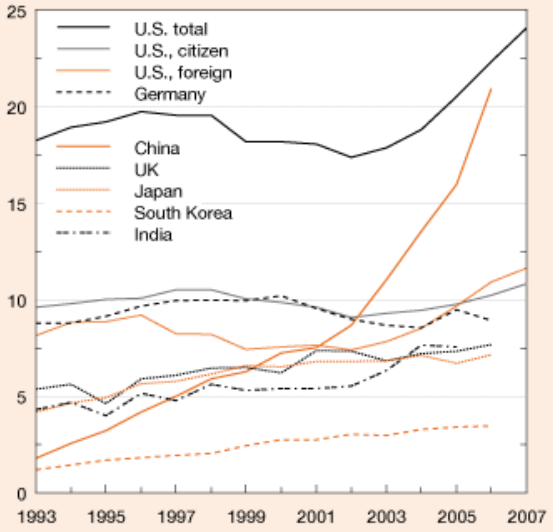
NOTE: Natural sciences include physical, biological, earth, atmospheric, ocean, agricultural, and computer sciences and mathematics.

SOURCES: China—National Bureau of Statistics of China, China Statistical Yearbook, annual series (Beijing), various years; Japan—Government of Japan, Ministry of Education, Culture, Sports, Science and Technology, Higher Education Bureau, Monbusho Survey of Education; South Korea and Germany—Organisation for Economic Co-operation and Development, Online Education Database, http://www.oecd.org/education/database/; United Kingdom—Higher Education Statistics Agency; and United States—National Center for Education Statistics, Integrated Postsecondary Education Data System, Completions Survey; and National Science Foundation, Division of Science Resources Statistics, Integrated Science and Engineering Resources Data System (WebCASPAR), http://webcaspar.nsf.gov.

Science and Engineering Indicators 2010

Figure O-9
Doctoral degrees in natural sciences and engineering, selected countries: 1993–2007

Thousands



UK = United Kingdom

NOTE: Natural sciences include physical, biological, earth, atmospheric, ocean, agricultural, and computer sciences and mathematics.

SOURCES: China—National Bureau of Statistics of China, China Statistical Yearbook, annual series (Beijing), various years; Japan—Government of Japan, Ministry of Education, Culture, Sports, Science and Technology, Higher Education Bureau, Monbusho Survey of Education; South Korea—Organisation for Economic Co-operation and Development (OECD), Online Education Database, http://www.oecd.org/education/ database/; United Kingdom—Higher Education Statistics Agency; Germany—Federal Statistical Agency, Prüfungen an Hochschulen, and OECD, Online Education Database, http://www.oecd.org/education/database/; and United States—National Center for Education Statistics, Integrated Postsecondary Education Data System, Completions Survey; and National Science Foundation, Division of Science Resources Statistics, Integrated Science and Engineering Resources Data System (WebCASPAR), http://webcaspar.nsf.gov.

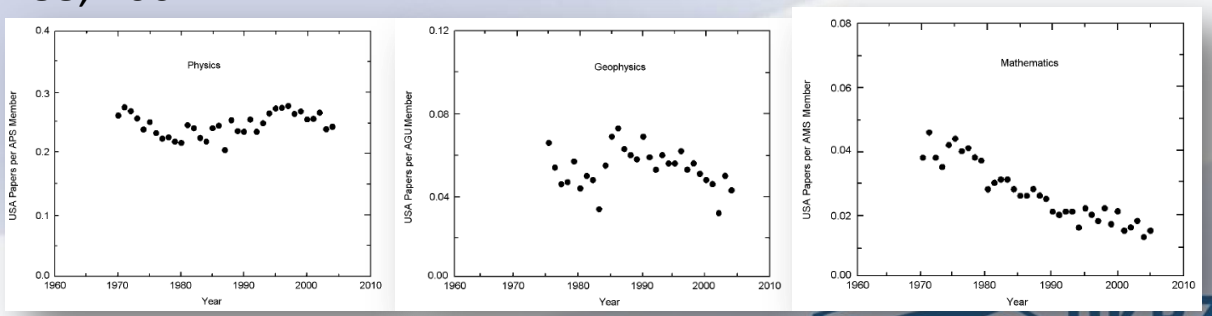
Science and Engineering Indicators 2010

DKRZ

Scientists Productivity

- “In the past 30–35 years there have been no increases in the average annual number of published papers per scientist [in America in physics, astronomy, geophysics, mathematics, and chemistry]”

H. Abt, “The publication rate of scientific papers depends only on the number of scientists,” *Scientometrics*, vol. 73, no. 3, pp. 281–288, 2007.



Seen so far…

- Data storage costs are decaying exponentially
- Data generation is growing exponentially
- Data is undermanaged (e.g. security)
- Scientists / IT personnel is barely growing
- Scientists “productivity” remains constant

To think about

- Are current methods for processing the increasing amount of data good enough?
- Can we work as we do now with 100x more data?
- Can we do 100x more “science”?
- (Can we remember 100x more Acronyms?)

What is holding data growth?

- Expenditure
- Technology
- Or perhaps we are?

The Human factor

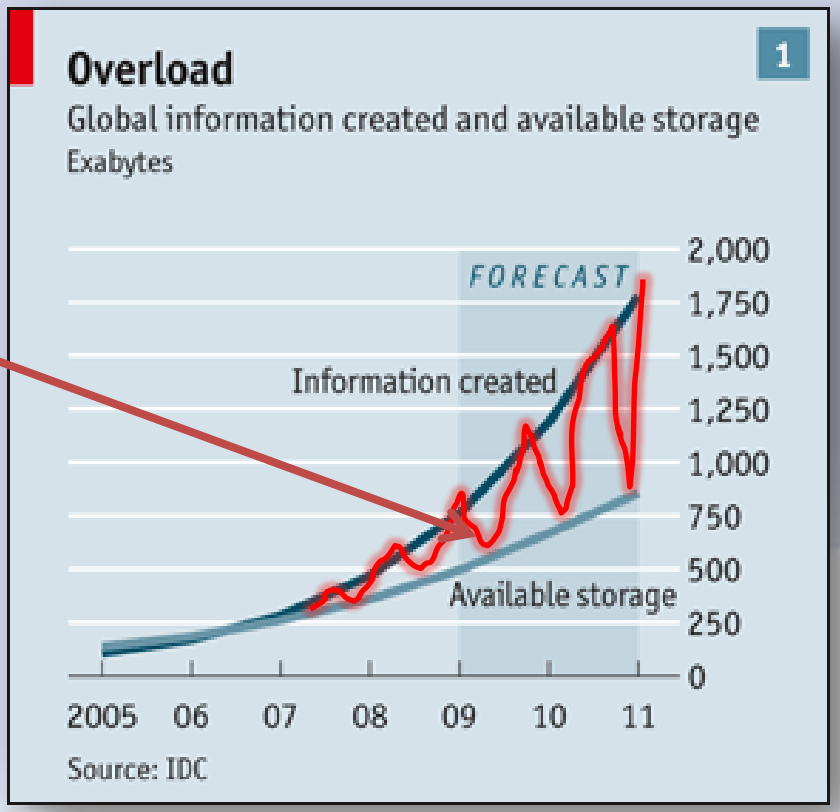
- “The generation of new information is limited by two key factors, by the incurring **economic costs** and by the capacity of the **human brain** to process and store data and information; the controlling agent needs to retain an overall understanding even when data is generated by semi-automatic processes.”

C. Gros, G. Kaczor, and D. Marković, “Neuropsychological constraints to human data production on a global scale,” *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 85, no. 1, pp. 1–5, 2012.

Furthermore

- We have no “time” to process data at all…

Someone has to decide what isn’t important…



What now?

- Data generation will not slow down
- *We* can't keep the pace much longer
- … so science will have to go through a major "change" (again)

The three paradigms of science

- Thousand years ago – empirical science
- Last few hundred years – analytical science
- Last few decades – computational science (simulations)

- Now the next paradigm shift is due

  – cf. Thomas S. Kuhn: The Structure of Scientific Revolutions

The Fourth Paradigm: Data-Intensive science

- T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009
  - http://research.microsoft.com/en-us/collaboration/fourthparadigm/
- *"The architecture for data-intensive computing should be based on storage, computing, and presentation services at every node of an interconnected network."* (p. 116)
- *"We should aim to provide scientists with a cyberinfrastructure on top of which it should be easy to build a large-scale application capable of exploiting the world's computer-represented scientific knowledge."* (p. 171)
- *"Scientific publication will become a 24/7, worldwide, real-time, interactive experience."* (p. 225)

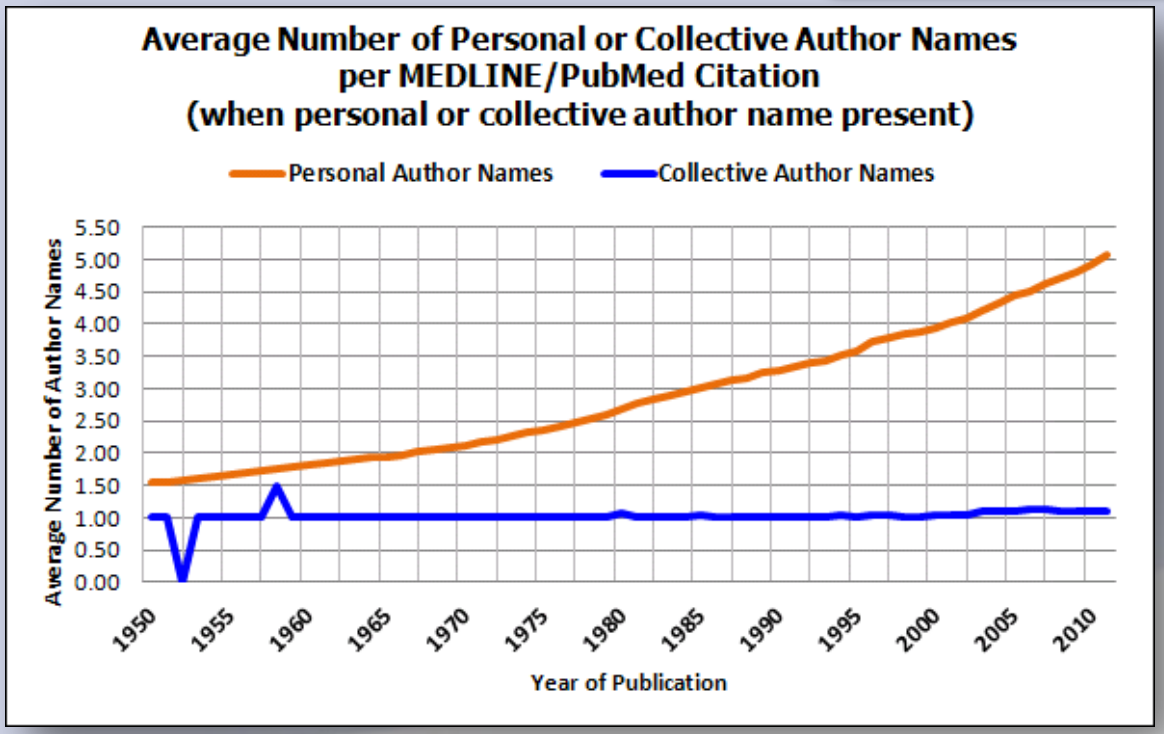
The Fourth Paradigm: Data-Intensive science

development

- "*Scientific ~~publication~~ will become a 24/7, worldwide, real-time, interactive experience.*" (p. 225)

"The ATLAS Experiment at the CERN Large Hadron Collider," Journal of Instrumentation, vol. 3, no. 08, p. S08003–S08003, Aug. 2008.

- 2926 Authors, 169 Institutions



The Earth Science Case

Data management in the CMIP5 project

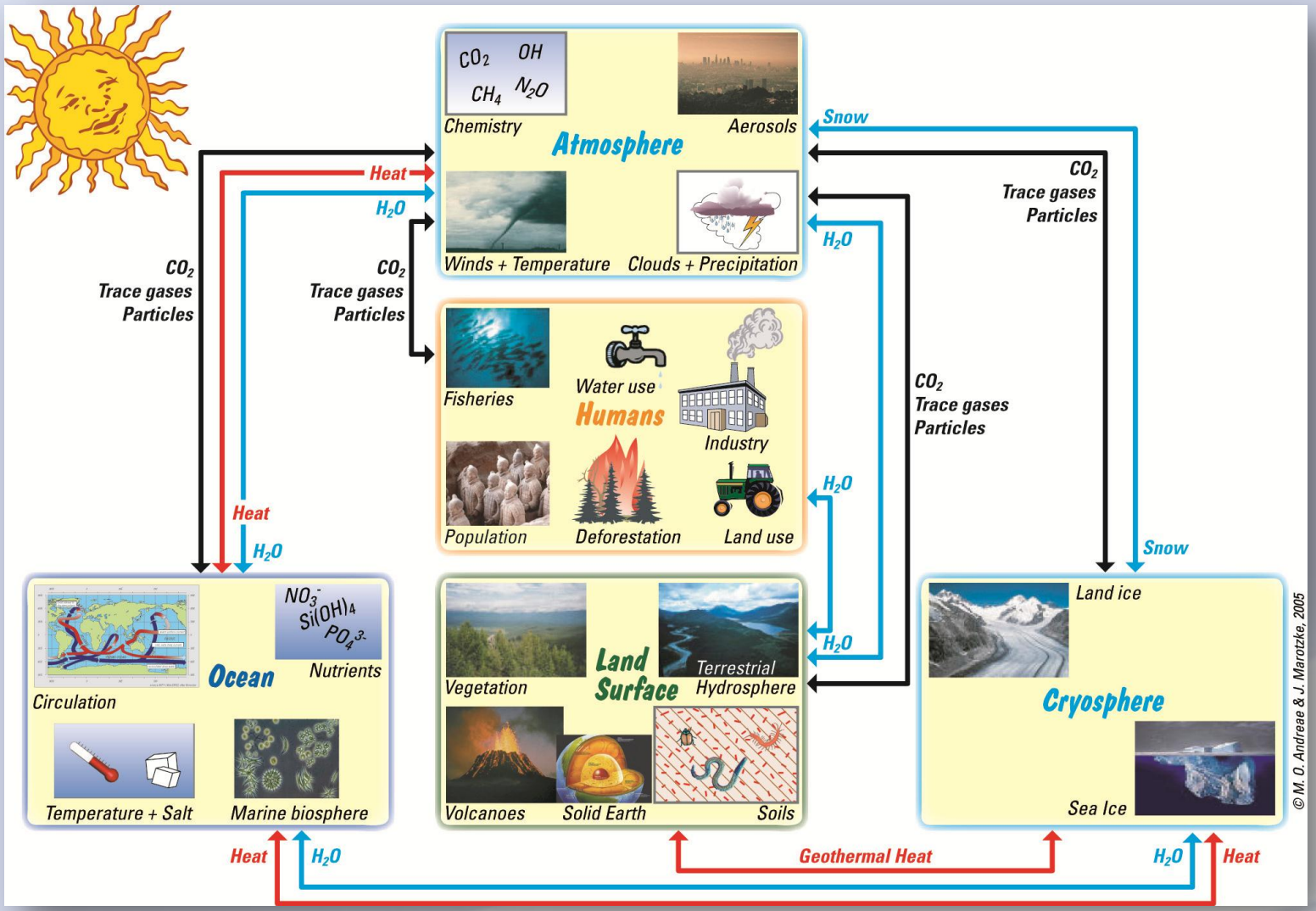
Data in Earth Science

- Simulations
- Observations
  - Satellites
  - Sensor data
- …

Coupled Model Intercomparison Project 5 (CMIP5)

- Project aimed at allowing the comparison of global coupled models

Climate Research
What are Coupled Models



Coupled Model Intercomparison Project 5 (CMIP5)

- Project aimed at allowing the comparison of global coupled models
- The main source of data for the Intergovernmental Panel on Climate Change (IPCC) Assessment Report
- Defines and standardizes vocabulary, experiments and data output required for data comparison
- In this iteration the expected audience is broader than the climate community (this implies more metadata!)
- This iteration also adds quality control
- … and DOIs for data citation

Comparison to last iteration

CMIP3	CMIP5
• Data Volume – 36 TB – 83.000 Files – Downloads: ~500GB/day	• Data Volume (expected) – 1-10 PB – ??? Files – Downloads: 10s of TB/day (+1Gbps)
• Models – 25 models	• Models – ~35 models – Increased resolution – More experiments – Increased complexity (ex: biogeochemistry)
• Metadata – CF-1 + IPCC-specific	• Metadata – CF-1 + IPCC-specific – Richer set of search criteria – Model configuration – Grid specification from CF (support for native grids)
• User Community – Thousands of users – WG1, domain knowledge	• User Community – 10s of thousands of users – Wider range of user groups will require better descriptions of data, attention to ease-of-use

CMIP5 current status

Summary	
Modeling centers	32
Models	59
Experiments	109
Data nodes	22
Gateways	5
Datasets	51907
Size	1,304.04 TB
Files	3,054,622

Data Accessible from Federated System – Not all data has been published yet - 18.6.2012

CMIP5 Data Requirements

- Integrated model metadata (data generation)
- Fast data search
- Distributed user management
- Security & Licensing
- Versioning – QC/DOI
- Notification System
- Long Term Archival
- Replication

1st Solution

A two-systems solution:

- **Data node**: data access management (PCMDI)
- **Gateway**: user management, search, metadata (NCAR)

1st Solution - Problem

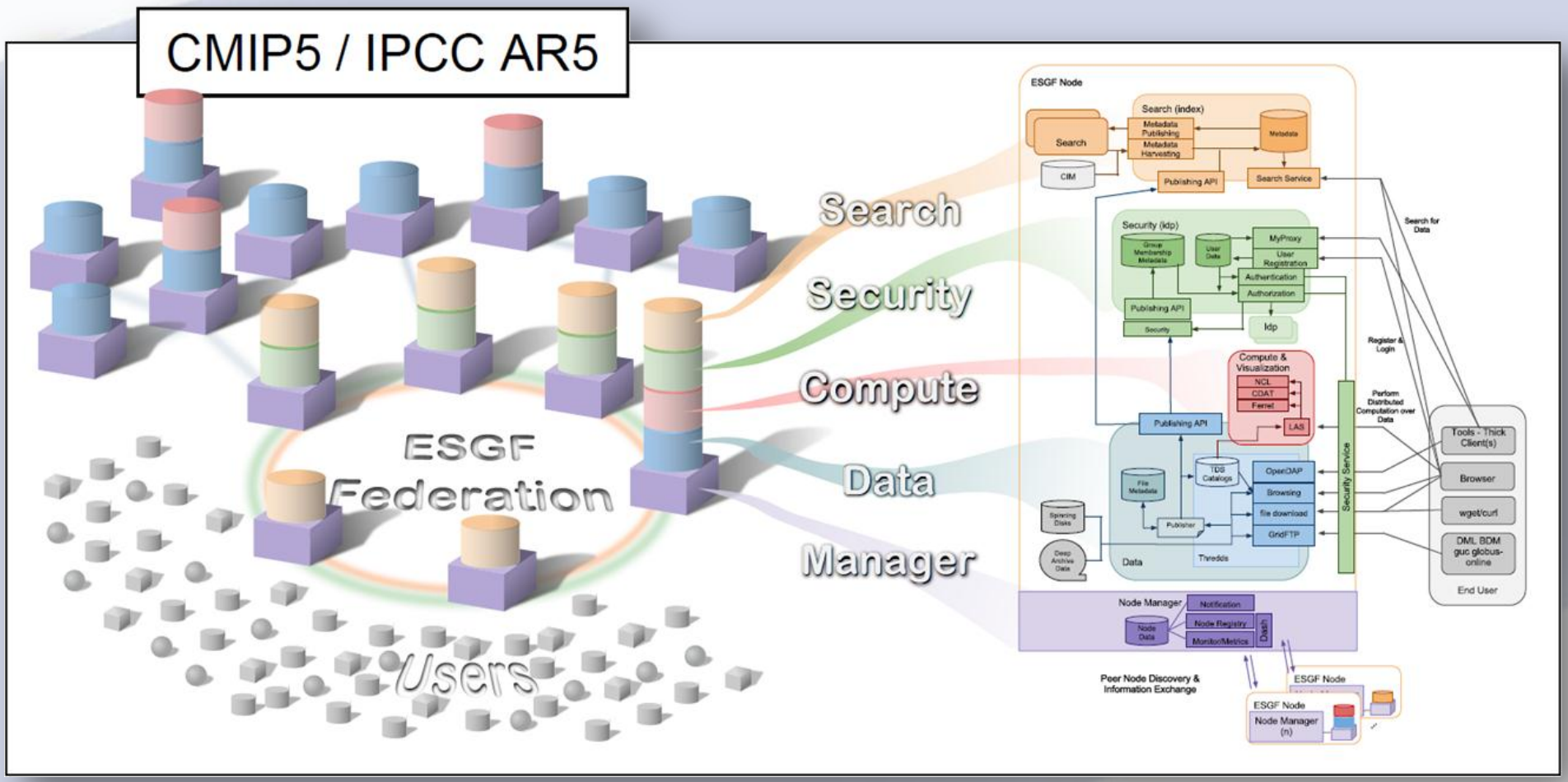
- Gateway is a monolithic (was closed-sourced) system
- Search (3Store) did not scale well
- NCAR had no resources to solve this in time, and though manifesting the intention to open the code, this was delayed for months
- Basically many communicational issues
- A group of institutions got together and decided to join efforts in solving these shortcomings
- The ESGF group was born and a parallel development started

2nd Solution

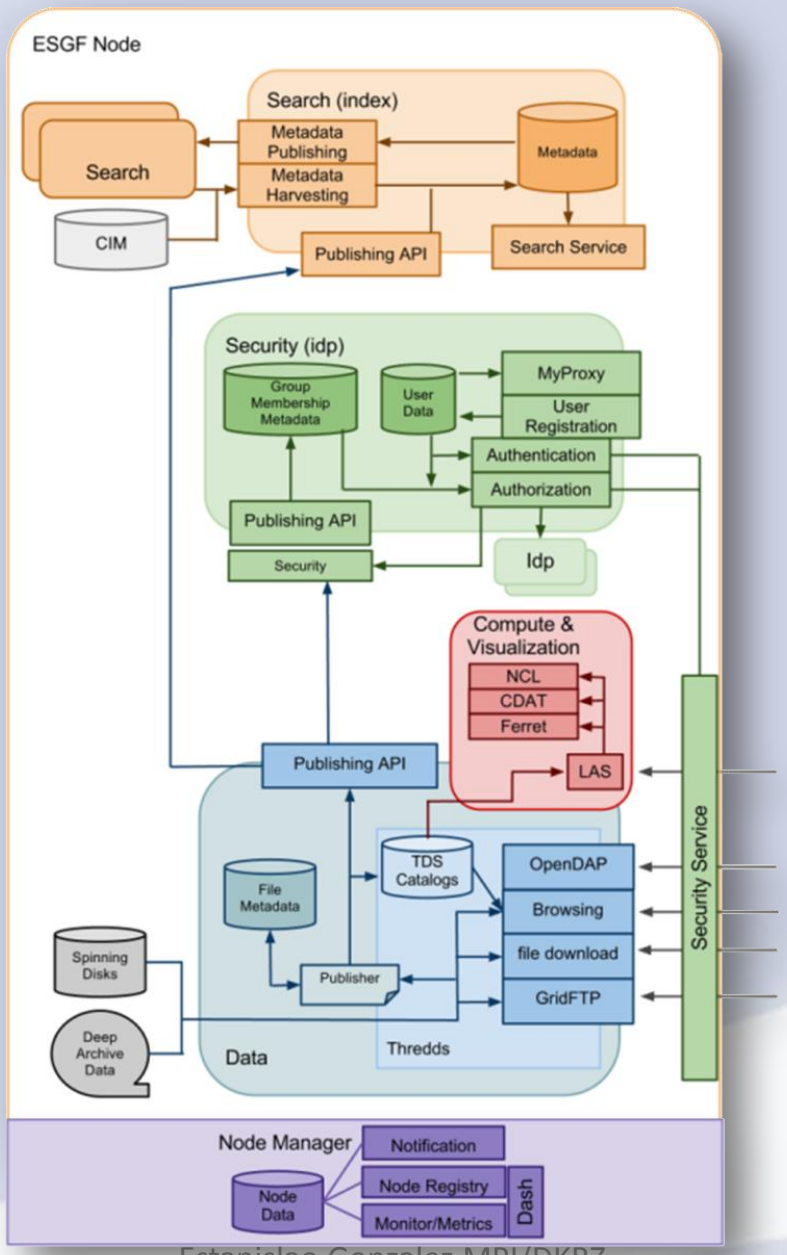
A single module-based System. It has 4 modules at this time:

- **Data**: same as the older data node
- **Computation**: allows visualization and basic computation operators at the data node
- **Index**: user interface and distributed search
- **Idp**: Identity provider for managing authentication/authorization at a federation level

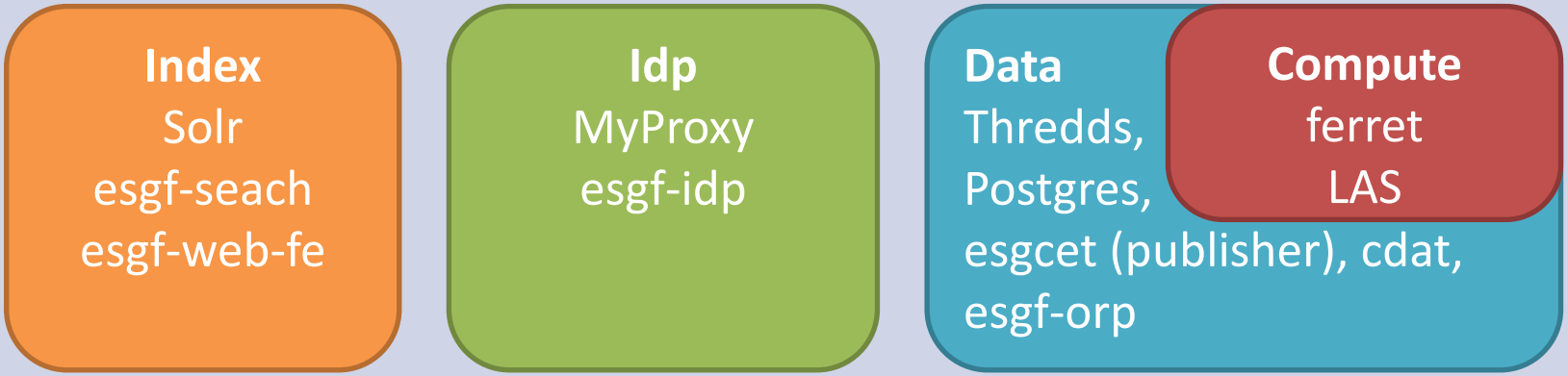
ESGF P2P Federation



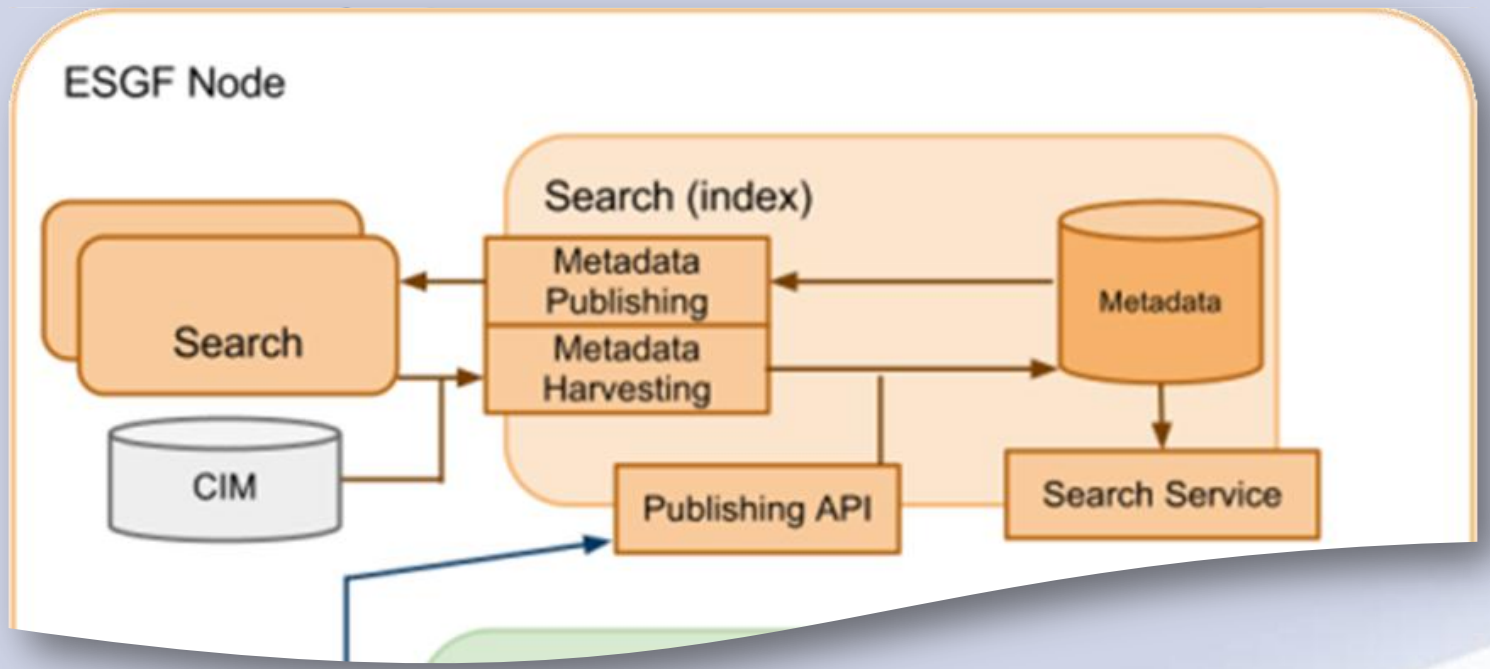
ESGF P2P Architecture



ESGF P2P Components



ESGF P2P Search

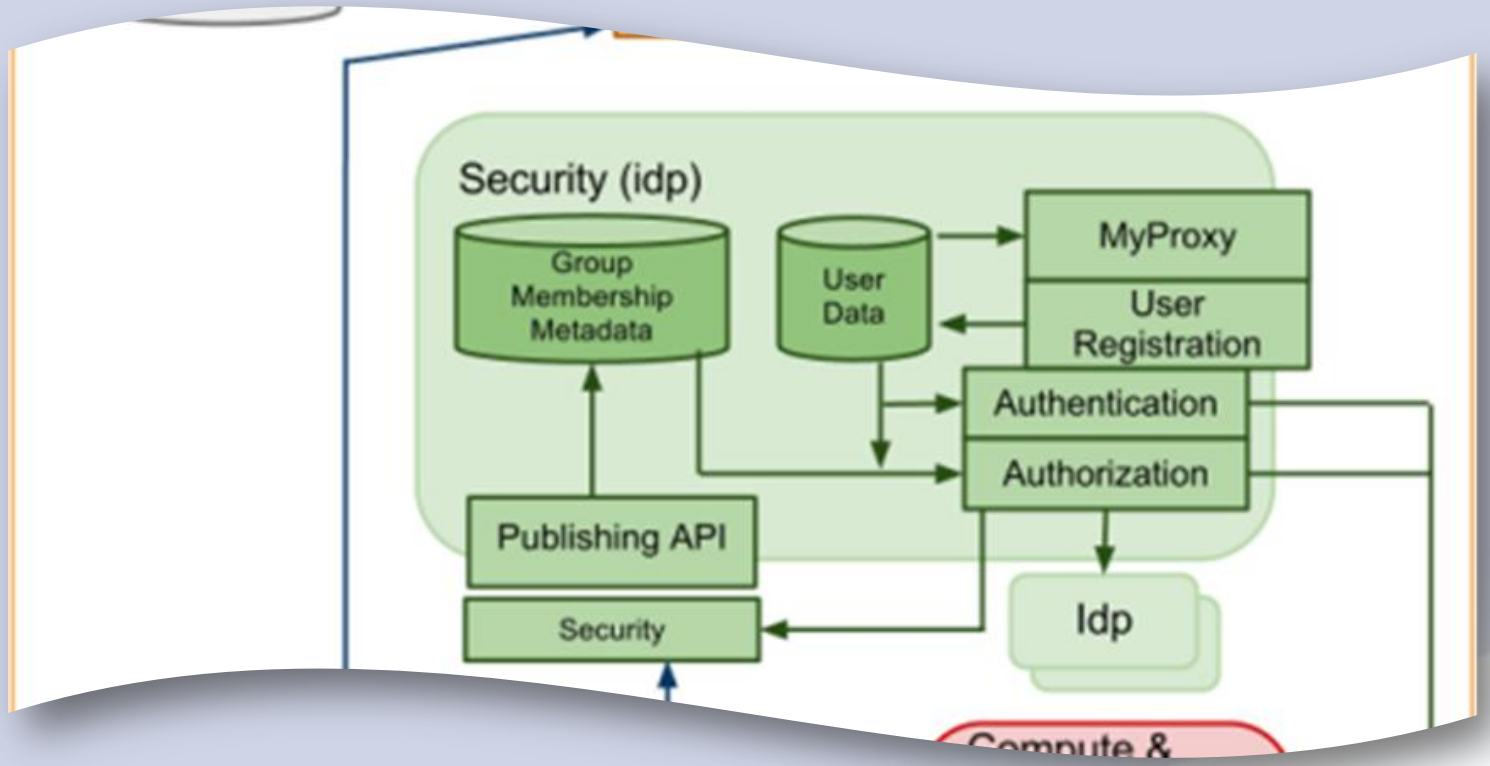


ESGF P2P Search

ESGF P2P Search

- Solr Backend
- esgf-search
  - provides standardized API
    - query=temperature&model!=MPI-ESM-LR&…
  - returns Solr results (XML/JSON) or a bash script for simple downloading
- esgf-web-fe (front end)
  - Implements GUI for faceted search
  - Displays CIM metadata (link to external source)
- No security is required for searching, basic concept is:
  - Metadata is free and unsecured, data isn't

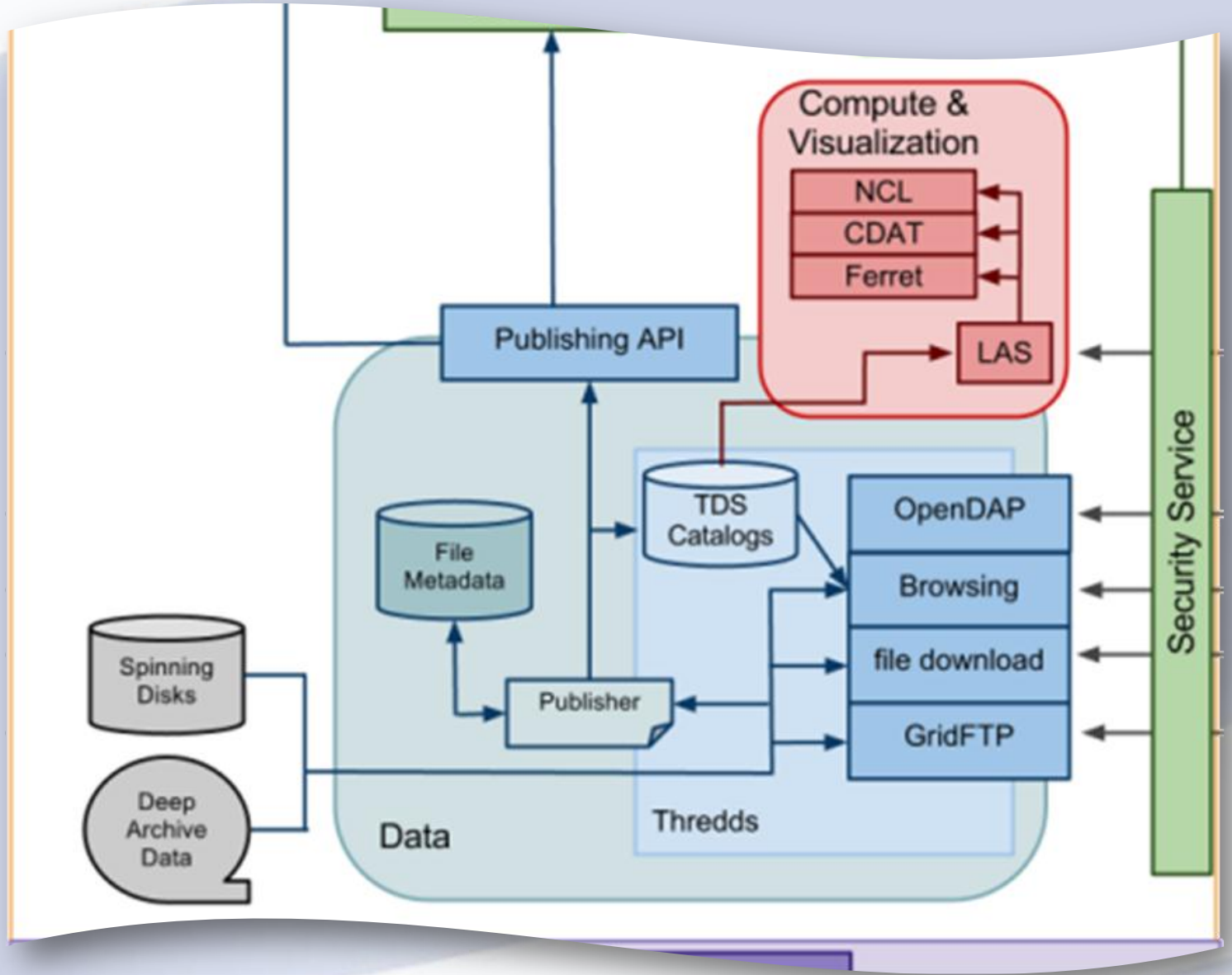
ESGF P2P Security



ESGF P2P Security

- esgf-idp
  - OpenId provider (authentication)
  - Attribute service (authorization)
  - Registration service (Joining a group, license, etc)
- MyProxy
  - automates generation of short lived X509 Certificates (72Hs) for non interactive system access
- esgf-orp (on the data side)
  - Set of Tomcat filters used for securing resources and handling authorization & authentication (X509, OpenID)

ESGF P2P Data + Compute



ESGF P2P Data + Compute - TDS

Catalog http://cmip3.dkrz.de/thredds/esgcet/catalog.html

Dataset	Size	Last Modif
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.ocean.day.r1i1p1.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.seaIce.day.r1i1p1.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.fx.atmos.fx.r0i0p0.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.fx.land.fx.r0i0p0.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.fx.ocean.fx.r0i0p0.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.mon.atmos.Amon.r1i1p1.v1/		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.monClim.atmos.Amon.r1i1p1.v1/		

Catalog http://cmip3.dkrz.de/thredds/esgcet/2/cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.v1.html

Dataset

project=CMIP5, model=BCC-CSM1.1, Beijing Climate Center, China Meteorolog
experiment=1 percent per year CO2, time_frequency=day, modeling realm=atmos, en

huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.huss.1.aggregati		
pr_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc		
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.pr.1.aggregation		--
psl_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc	1.675 Gbytes	--
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.psl.1.aggregation		--
tas_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc	1.675 Gbytes	--
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tas.1.aggregation		--
tasmax_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc	1.675 Gbytes	--
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tasmax.1.aggregation		--
tasmin_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc	1.675 Gbytes	--
cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tasmin.1.aggregation		--

WDCC Test TDS

THREDDS Data Server

Catalog http://cmip3.dkrz.de/thredds/esgcet/2/cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.v1.html

Dataset: project=CMIP5, model=BCC-CSM1.1, Beijing Climate Center, China Meteorological Administration, experiment=1 percent per year CO2, time_frequency=day, modeling realm=atmos, ensemble=r1i1p1, version=1/huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc

- *Data format:* NetCDF
- *Data size:* 1.675 Gbytes
- *Data type:* GRID
- *ID:* cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.v1.huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc
- *RestrictAccess:* esg-user

Access:

1. **HTTPServer:** /thredds/fileServer/cmip5/output1/BCC/bcc-csm1-1/1pctCO2/day/atmos/day/r1i1p1/v1/huss/huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc
2. **GridFTP:** gsiftp://cmip3.dkrz.de:2811//cmip5/output1/BCC/bcc-csm1-1/1pctCO2/day/atmos/day/r1i1p1/v1/huss/huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc
3. **OPENDAP:** /thredds/dodsC/cmip5/output1/BCC/bcc-csm1-1/1pctCO2/day/atmos/day/r1i1p1/v1/huss/huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc

Variables:

- *Vocabulary* [CF-1.0]:
  - **huss** = *Near-Surface Specific Humidity* = specific_humidity (1)

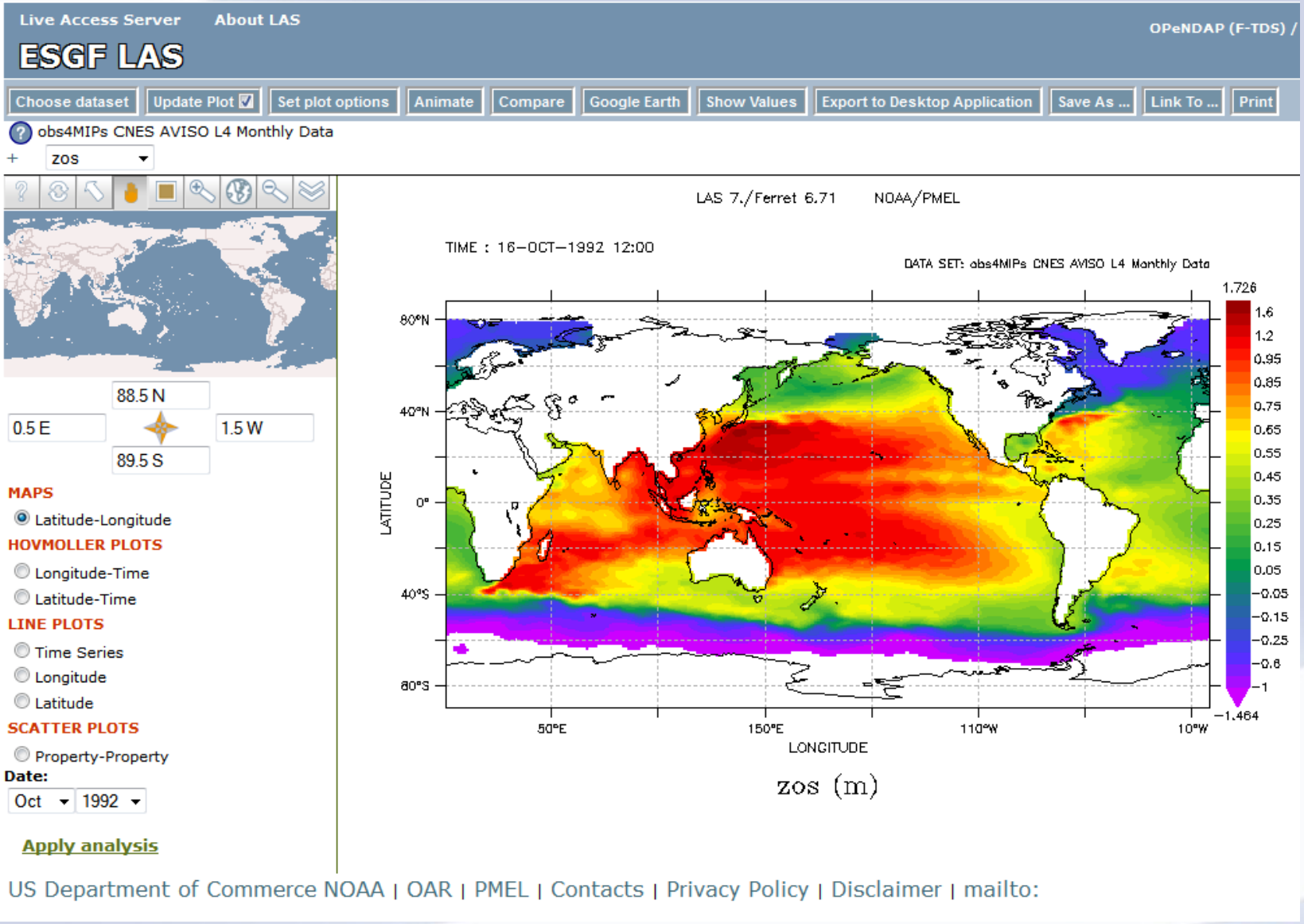
Properties:

- file_id = "cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.huss_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc"
- file_version = "1"
- size = "1675683884"
- tracking_id = "d76ed0b5-3cf2-4e7d-8e38-ecfad0b9dc31"
- mod_time = "2011-04-19 09:39:25"
- checksum = "17eca0a1707d63cbacb12ec11818adb3"
- checksum_type = "md5"

Viewers:

- Integrated Data Viewer (IDV) (webstart)
- NetCDF-Java ToolsUI (webstart)

ESGF P2P Data + Compute - LAS



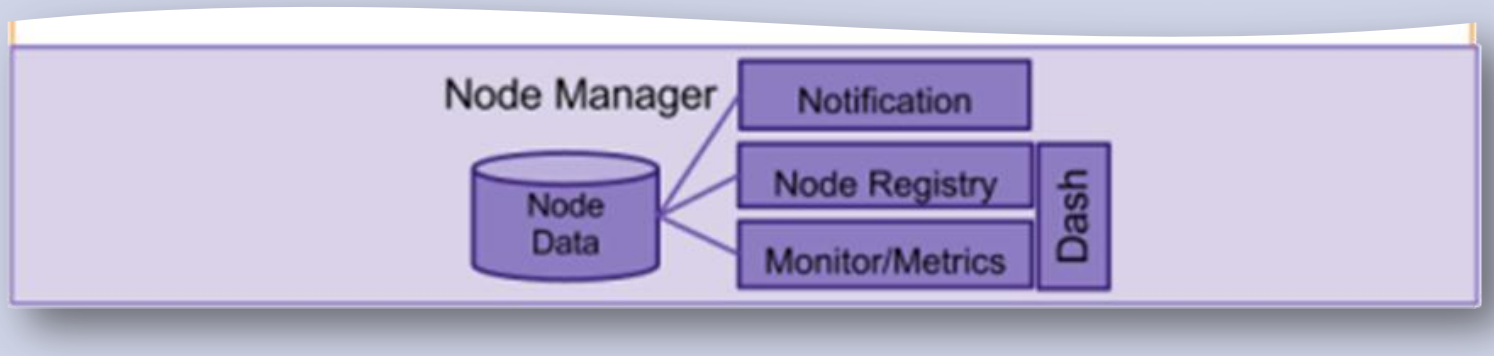
ESGF P2P Data + Compute

- Thredds Data Server (TDS)
  - Hosts metadata via xml catalogs
  - Allows HTML browsing
  - Manages file access
- GridFTP
  - Prefered method for downloading
- esgcet
  - Set of tools to handle publication (versioning, metadata extraction, catalog generation, etc)

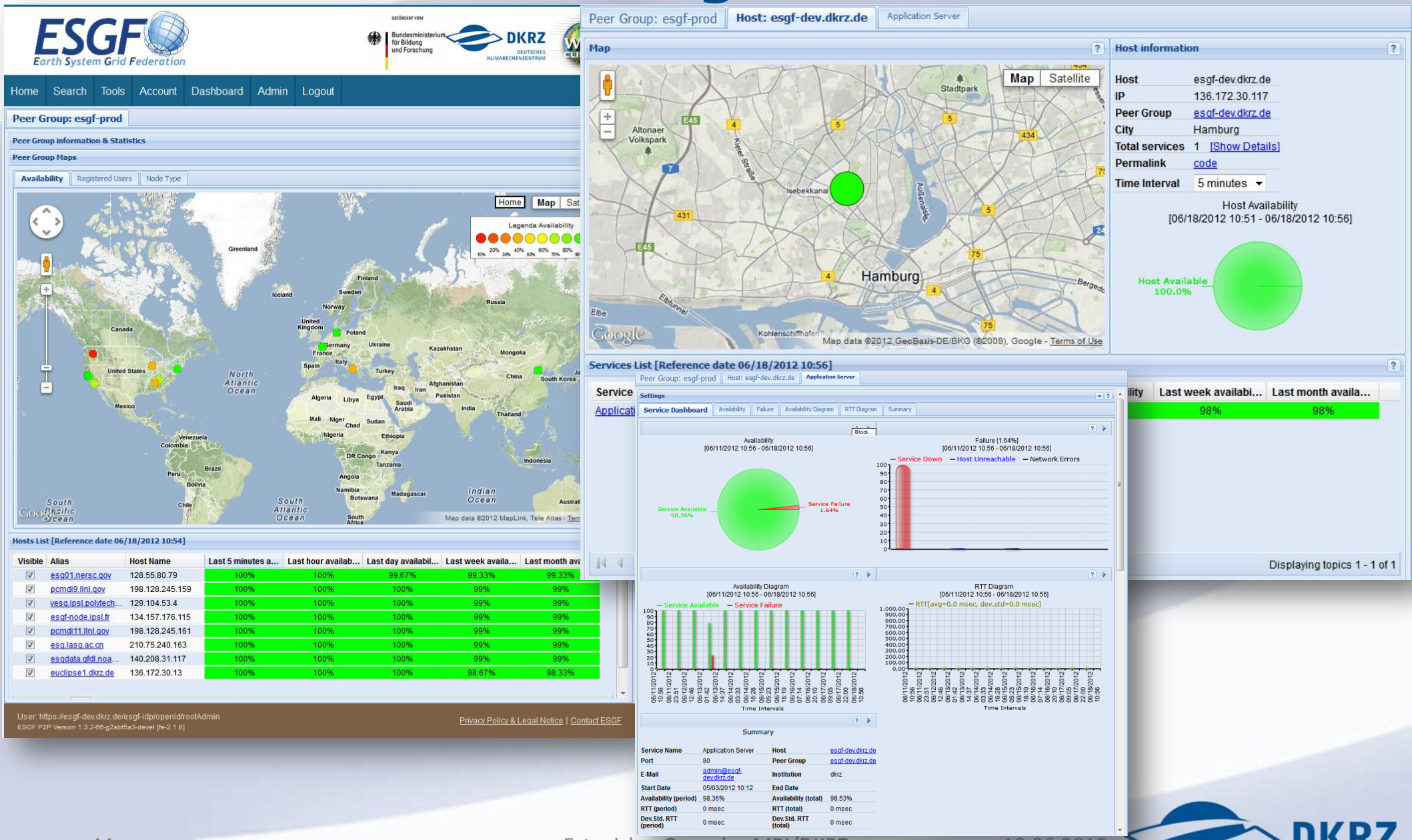
ESGF P2P Data + Compute

- LAS
  - Data visualization
  - Simple analysis tools (min, mean, max)
  - Data comparison

ESGF P2P Manager



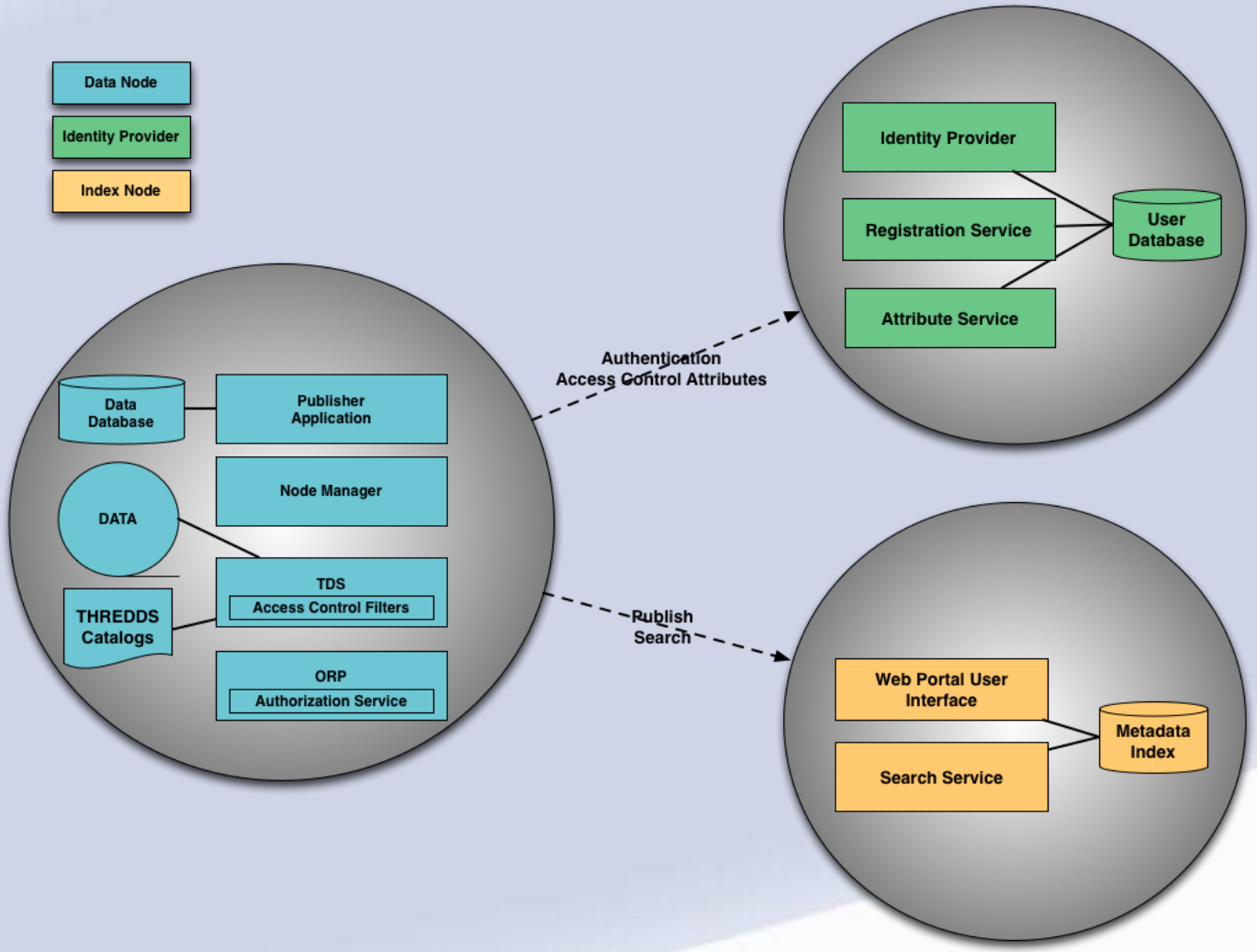
ESGF P2P Manager - Dashboard



ESGF P2P Manager

- esgf-node-manager
  - Registry (of all available services)
  - P2P protocoll (peer discovery, etc)
- esgf-dashboard
  - Visualization of node status and metrics

ESGF P2P Interactions



Thank you.