

~~Earth~~ Science in the Data Era

How data is changing the way we
do science

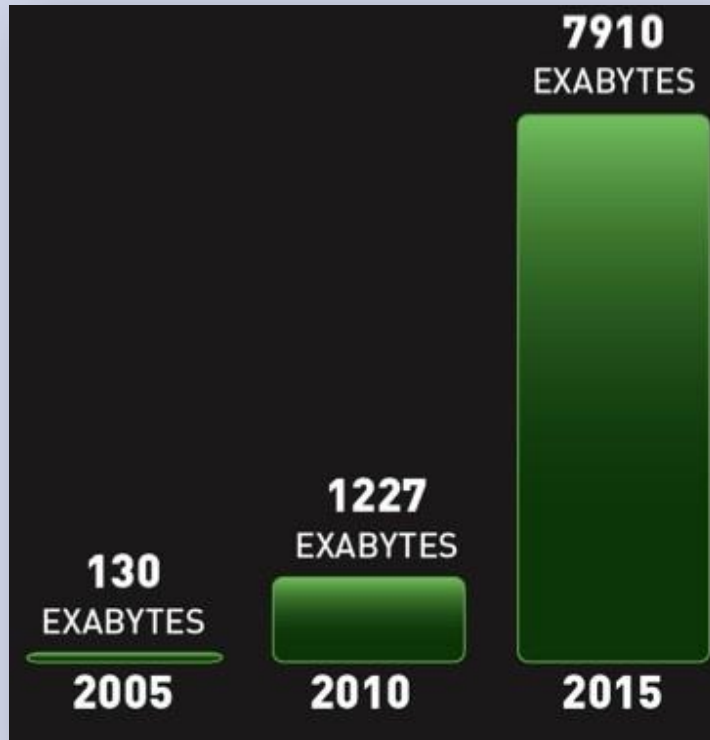
Summary

- Data & Science
- The Earth Science Case

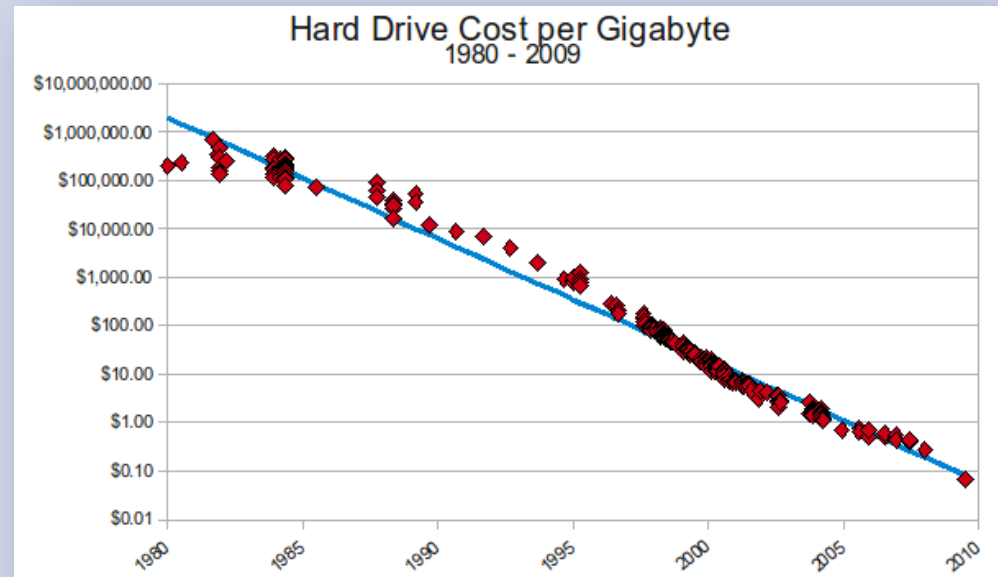
Data & Science

Where are we?
Where to go next?

Data Grow & Storage Cost

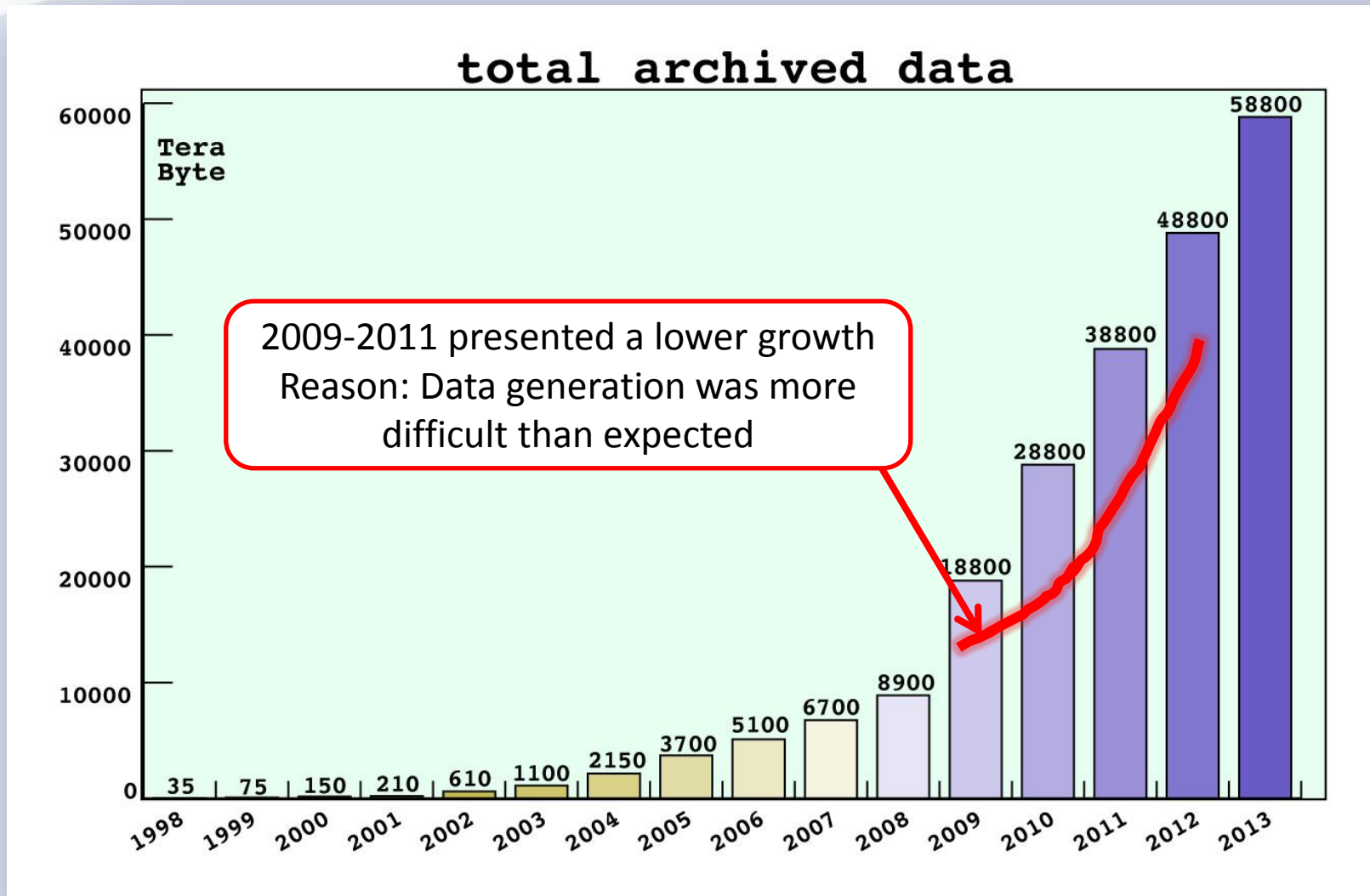


IDC's Digital Universe Study, sponsored by EMC, June 2011



<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

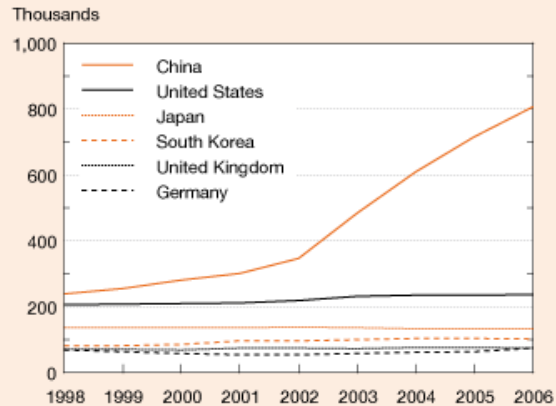
And Security, of course...

- “[...] 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

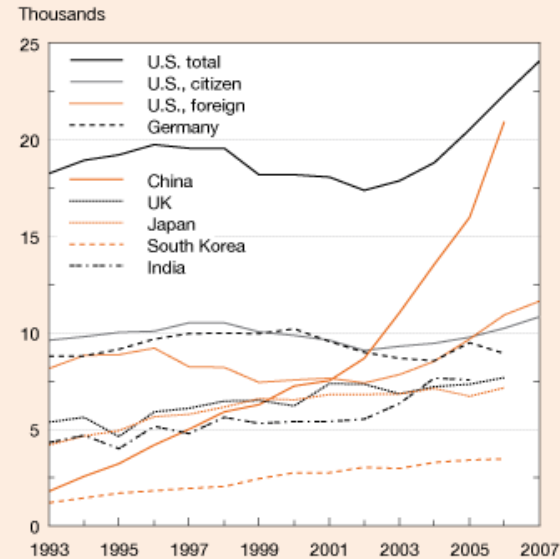


NOTE: Natural sciences include physical, biological, earth, atmospheric, ocean, agricultural, and computer sciences and mathematics.

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



UK = United Kingdom

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

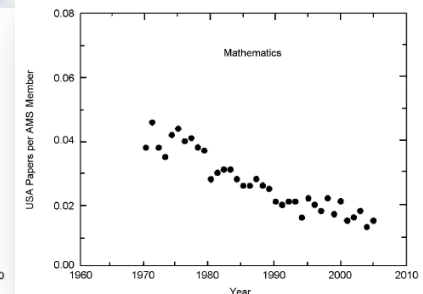
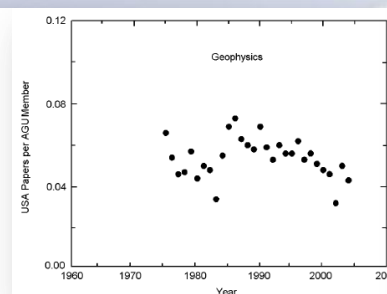
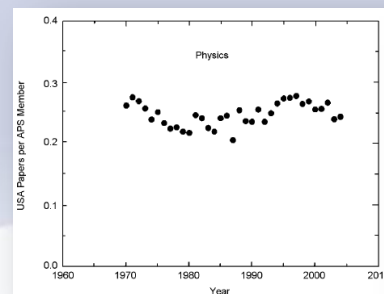
SOURCES: China—National Bureau of Statistics of China, China Statistical Yearbook, annual series (Beijing), various years; Japan—Government of Japan, Ministry of Education, Culture, Sports, Science and Technology, Higher Education Bureau, Monbusho Survey of Education; South Korea—Organisation for Economic 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

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

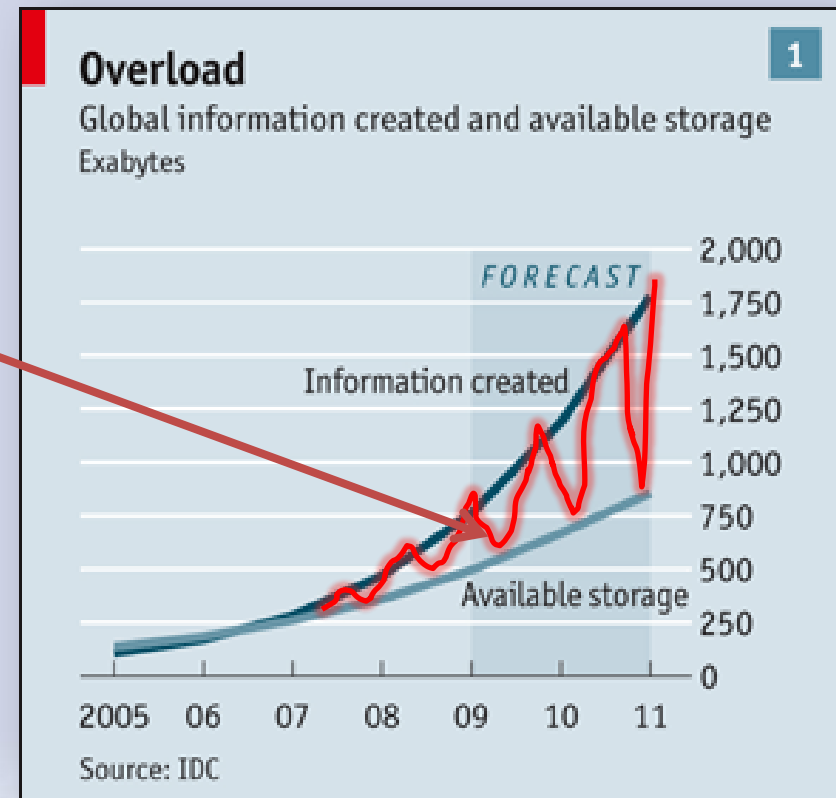
- “The generation of new information is limited by two key factors, by the incurring **economic costs** and by the capacity of the **human brain** to process and store data and information; the controlling agent needs to retain an overall understanding even when data is generated by semi-automatic processes.”

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)

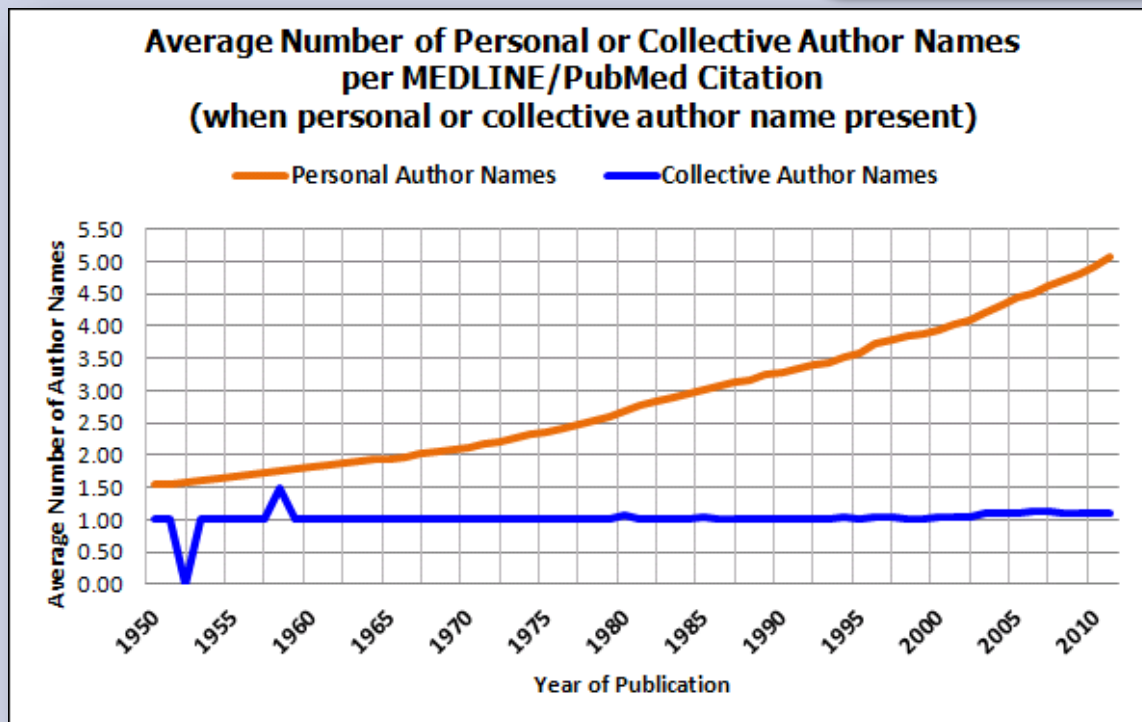
The Fourth Paradigm: Data-Intensive science

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

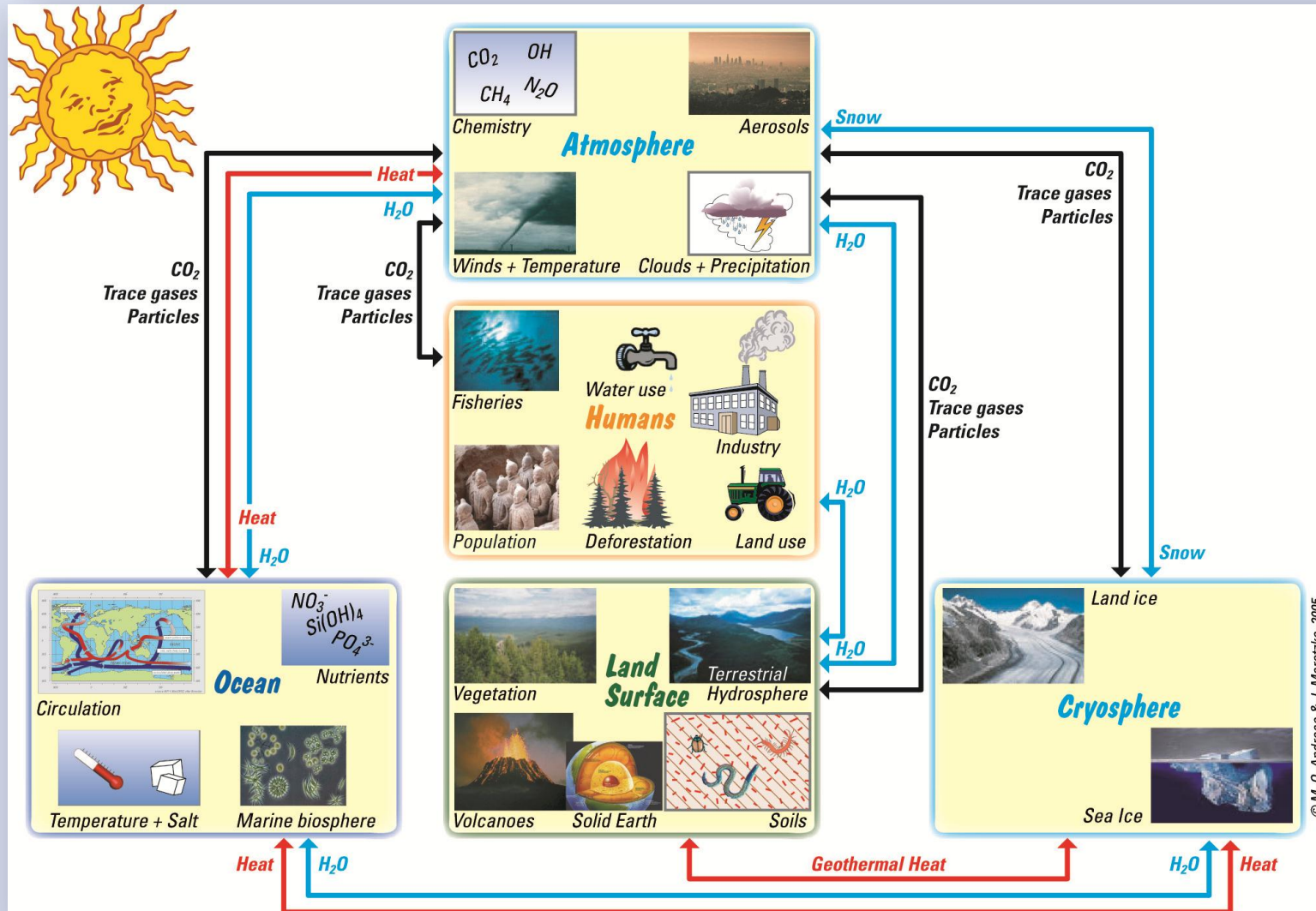
- Simulations
- Observations
 - Satellites
 - Sensor data
- ...

Coupled Model Intercomparison Project 5 (CMIP5)

- Project aimed at allowing the comparison of global coupled models

Climate Research

What are Coupled Models



© M. O. Andreae & J. Marotzke, 2005

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

- Data Volume
 - 36 TB
 - 83.000 Files
 - Downloads: ~500GB/day
- Models
 - 25 models
- Metadata
 - CF-1 + IPCC-specific
- User Community
 - Thousands of users
 - WG1, domain knowledge

28~280x

CMIP5

- Data Volume (expected)
 - 1-10 PB
 - ??? Files
 - Downloads: 10s of TB/day (+1Gbps)
- Models
 - ~35 models
 - Increased resolution
 - More experiments
 - Increased complexity (ex: biogeochemistry)
- Metadata
 - CF-1 + IPCC-specific
 - Richer set of search criteria
 - Model configuration
 - Grid specification from CF (support for native grids)
- 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

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

36x

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

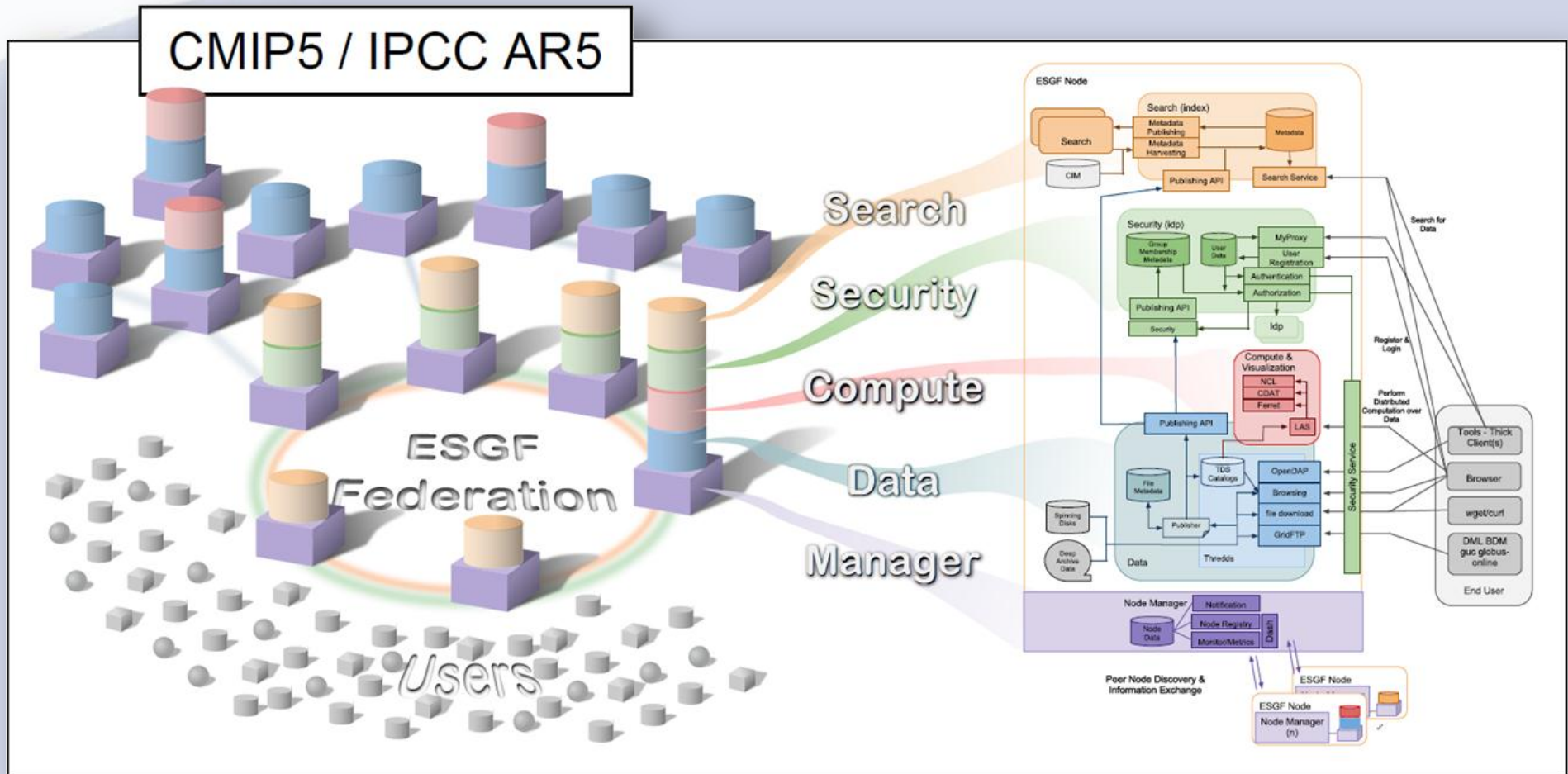
- Gateway is a monolithic (was closed-sourced) system
- Search (3Store) did not scale well
- NCAR had no resources to solve this in time, and though manifesting the intention to open the code, this was delayed for months
- Basically many communicational issues
- A group of institutions got together and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

2nd Solution

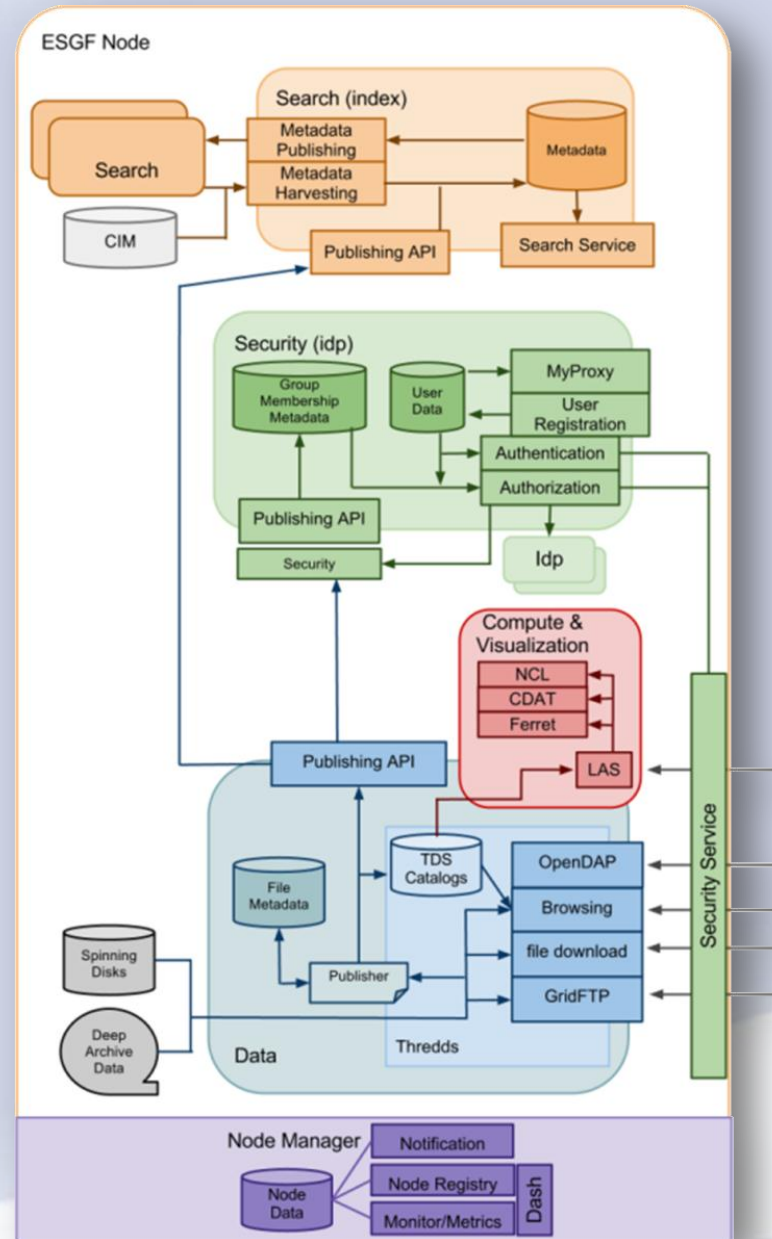
A single module-based System. It has 4 modules at this time:

- **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

Index

Solr
esgf-seach
esgf-web-fe

Idp

MyProxy
esgf-idp

Data

Thredds,
Postgres,
esgct (publisher), cdat,
esgf-orp

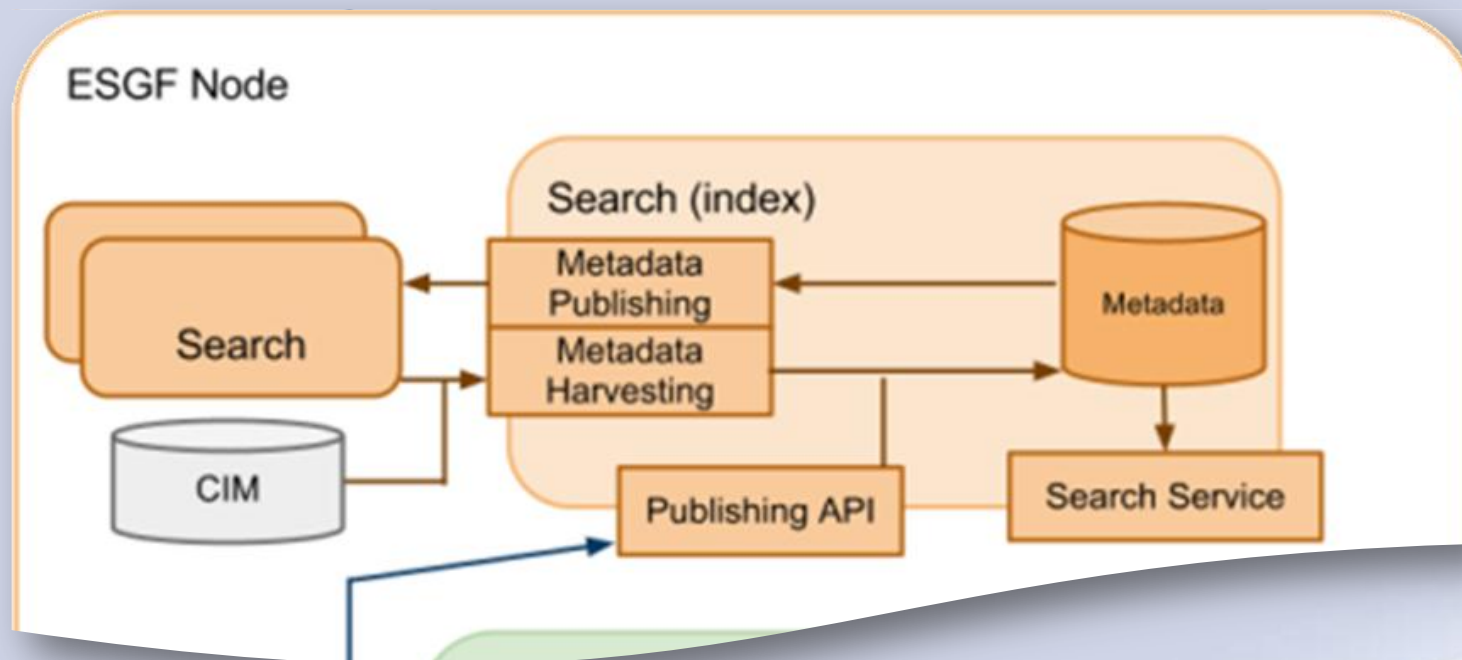
Compute

ferret
LAS

ESGF Node Manager

Java, Git, Ant, Tomcat, curl, openssl
esgf-node-manager, esgf-dashboard

ESGF P2P Search



ESGF P2P Search

ESGF
Earth System Grid Federation

Home Search Tools Login

Welcome to the ESGF P2P Node @ DKRZ

Quick Search

Keyword:

Advanced Search (Category, Geospatial, Temporal, and more)...

Peer Nodes

- ANL Node
- CMCC Node
- DKRZ Node
- IPSL Node
- NASA-JPL Node
- ORNL Node
- PCMDI Node
- PNNL Node

About DKRZ TEST P2P node

DKRZ, the German Climate Computing Centre, provides tools and the associated services which are needed to investigate the processes in the climate system: Computer power, data management and guidance to use these tools efficiently. As a national service provider, DKRZ operates a supercomputer center to enable climate simulation and provides the scientific users with the technical infrastructure needed for processing and analysis of climate data. This also includes support for related application software, advice and support in data processing issues. Finally, DKRZ also participates in national and international joint projects with the aim of improving the infrastructure for climate modeling.

Learn More

Resources

Quick Links

- Create Account
- MyProxyLogin
- Expert Search (XML)
- Wget Script Generator
- ESGF aggregated RSS feed
- Contact ESGF

Instructions

- ESGF Full User Guide
- Search Help
- Wget Scripts FAQ
- Wget Scripting
- Tutorial: Download Strategies
- Using Globus Online
- Subscribing to RSS Notification

Guest User | ESGF P2P Version 1.3.2-26-g00d5632-devel [7e-2.0.4] [Privacy Policy](#) | [Terms of Use](#) | [Contact ESGF](#)

ESGF
Earth System Grid Federation

Home Search Tools Login

Current Selections

(x) project:CMIP5

Search

Temporal Search
Geospatial Search
Browse Categories
Search Help

Search All Sites Show All Replicas Show All Versions

< 1 2 3 ... 4366 4367 > displaying 1 to 10 of 4366 search results

Search Categories

Project

Institute

Model

Instrument

Experiment Family

Experiment

Time Frequency

3hr (1520)

6hr (1635)

day (3685)

fx (1641)

mon (27751)

monClim (80)

subhr (25)

yr (658)

Product

Realm

Variable

Variable Long Name

MIP Table

CF Standard Name

Ensemble

Results Data Cart

project=CMIP5 / IPCC Fifth Assessment Report, model=Institute for Numerical Mathematics, experiment=ESM historical, time_frequency=fx, modeling_realm=atmos, ensemble=r0i0p0, version=20110927
Data Node: pomdi9.llnl.gov
Version: 20110927
Description: Inmcm4 model output prepared for CMIP5 ESM historical
Further options: [Add To Cart](#) [Visualize and Analyze](#) [CIM Metadata](#)

project=CMIP5 / IPCC Fifth Assessment Report, model=Institute for Numerical Mathematics, experiment=1 percent per year CO2, time_frequency=day, modeling_realm=atmos, ensemble=r1i1p1, version=20110323
Data Node: pomdi9.llnl.gov
Version: 20110323
Description: Inmcm4 model output prepared for CMIP5 1 percent per year CO2
Further options: [Add To Cart](#) [Visualize and Analyze](#) [CIM Metadata](#)

project=CMIP5 / IPCC Fifth Assessment Report, model=Institute for Numerical Mathematics, experiment=1 percent per year CO2, time_frequency=day, modeling_realm=ocean, ensemble=r1i1p1, version=20110323
Data Node: pomdi9.llnl.gov
Version: 20110323
Description: Inmcm4 model output prepared for CMIP5 1 percent per year CO2
Further options: [Add To Cart](#) [Visualize and Analyze](#) [CIM Metadata](#)

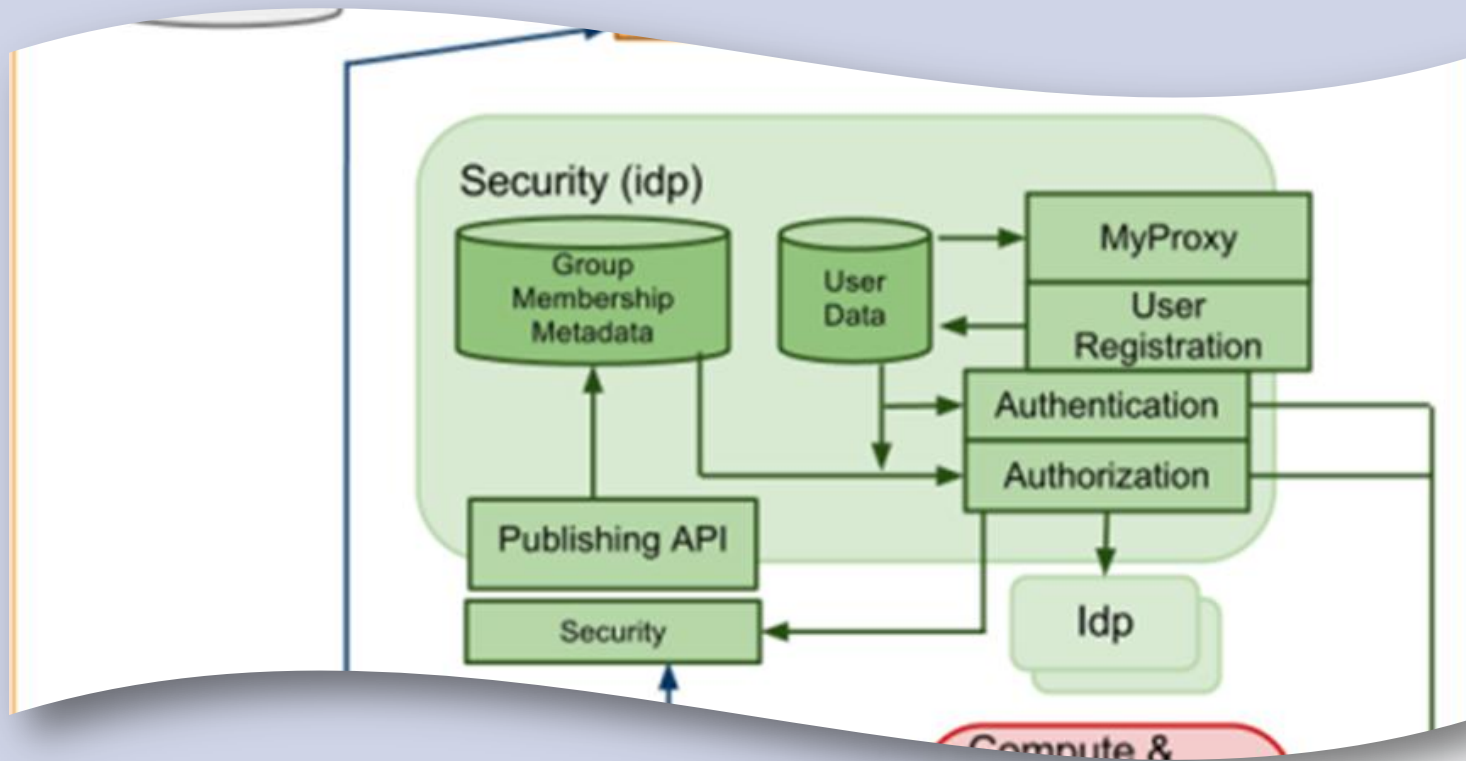
project=CMIP5 / IPCC Fifth Assessment Report, model=Institute for Numerical Mathematics, experiment=1 percent per year CO2, time_frequency=mon, modeling_realm=atmos, ensemble=r1i1p1, version=20110323
Data Node: pomdi9.llnl.gov
Version: 20110323
Description: Inmcm4 model output prepared for CMIP5 1 percent per year CO2
Further options: [Add To Cart](#) [Visualize and Analyze](#) [CIM Metadata](#)

project=CMIP5 / IPCC Fifth Assessment Report, model=Institute for Numerical Mathematics,

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

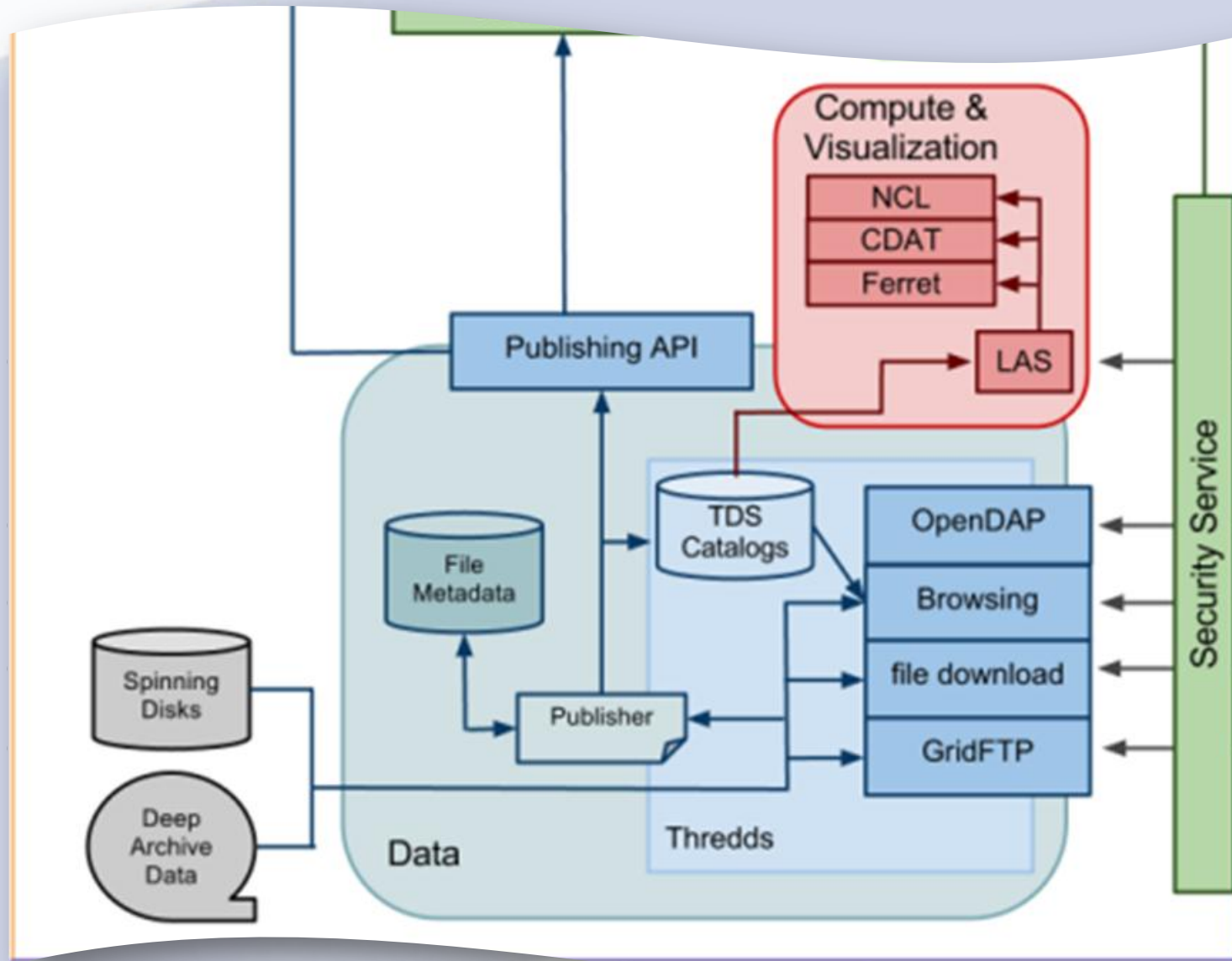
ESGF P2P Security



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/esgcat/catalog.html>

WDCC Test TDS

THREDDS Data Server

Dataset **Size** **Last Modified**

Catalog <http://cmip3.dkrz.de/thredds/esgcat/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: [gsftftp://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](ftp://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)

Catalog <http://cmip3.dkrz.de/thredds/esgcat/2/cmip5.output1.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

cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.huss.1.aggregation

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

cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.psl.1.aggregation

tas_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc

cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tas.1.aggregation

tasmax_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc

cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tasmax.1.aggregation

tasmin_day_bcc-csm1-1_1pctCO2_r1i1p1_01600101-02991231.nc

cmip5.output1.BCC.bcc-csm1-1.1pctCO2.day.atmos.day.r1i1p1.tasmin.1.aggregation

1.675
Gbytes

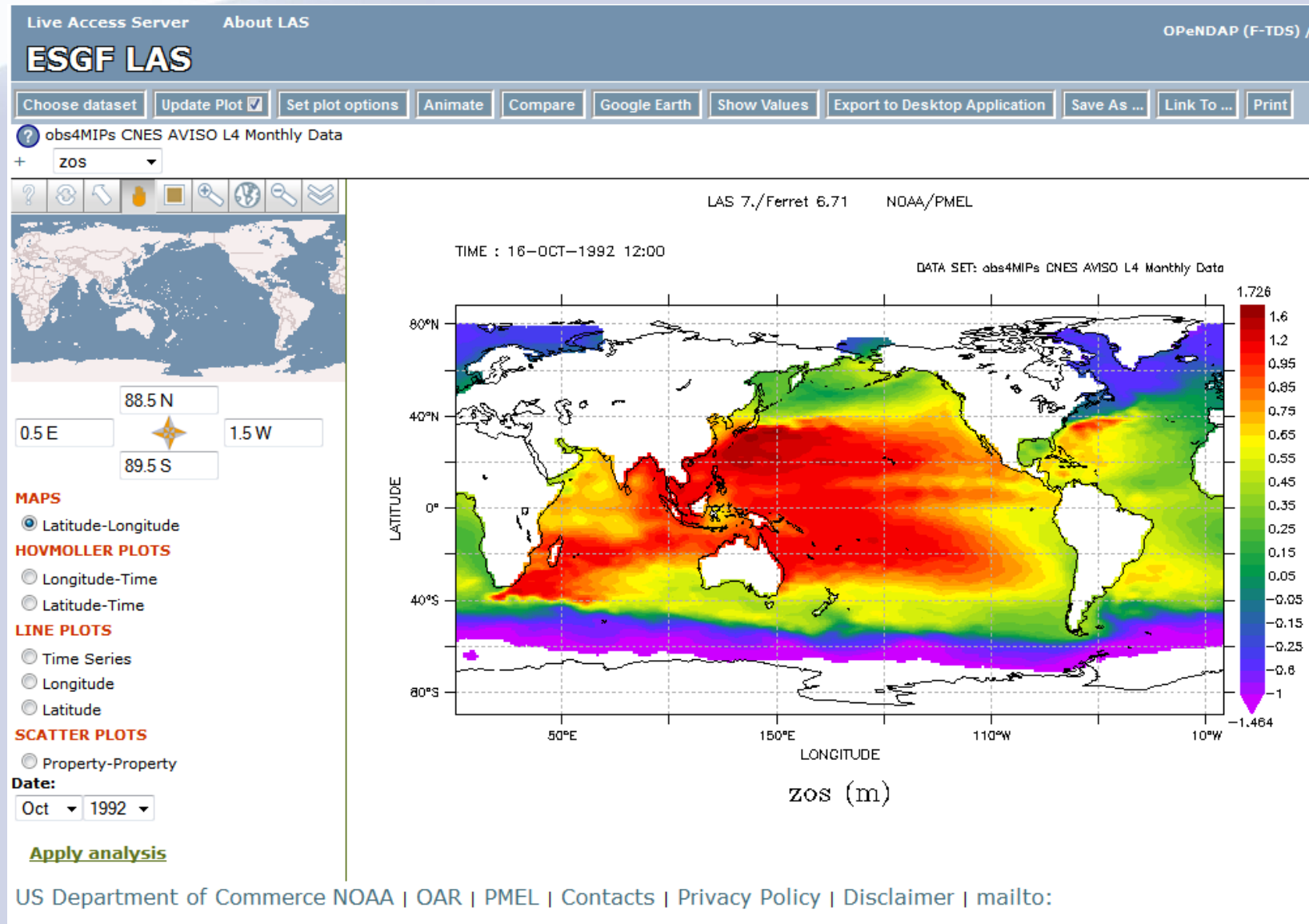
1.675
Gbytes

1.675
Gbytes

1.675
Gbytes



ESGF P2P Data + Compute - LAS



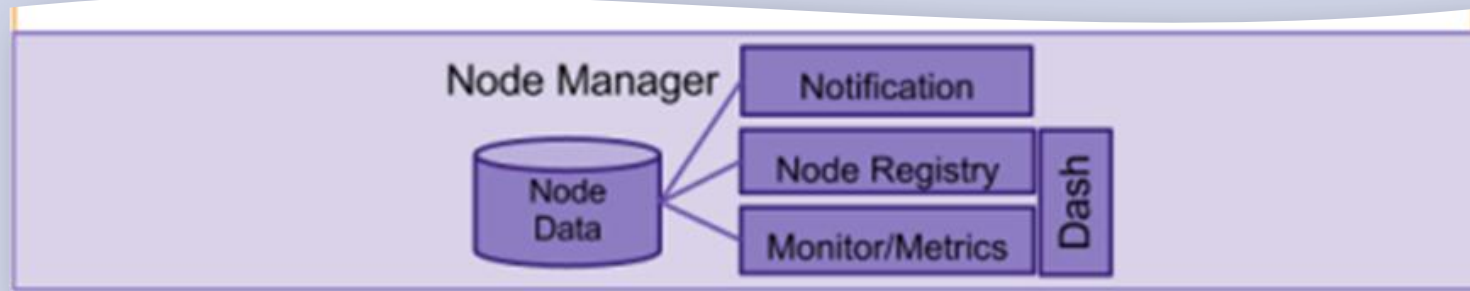
ESGF P2P Data + Compute

- Thredds Data Server (TDS)
 - Hosts metadata via xml catalogs
 - Allows HTML browsing
 - Manages file access
- GridFTP
 - Preferred 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



Peer Group: esgf-prod Host: esgf-dev.dkrz.de Application Server

Home Search Tools Account Dashboard Admin Logout

Peer Group: esgf-prod

Peer Group information & Statistics

Peer Group Maps

Availability Registered Users Node Type



Hosts List [Reference date 06/18/2012 10:54]

Visible	Alias	Host Name	Last 5 minutes a...	Last hour availab...	Last day availabil...	Last week availa...	Last month av
<input checked="" type="checkbox"/>	esg01.nersc.gov	128.55.80.79	100%	100%	99.67%	99.33%	99.33%
<input checked="" type="checkbox"/>	pcmdi9.llnl.gov	198.128.245.159	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	vesg.ipsl.polytech...	129.104.53.4	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	esgf-node.ipsl.fr	134.157.176.115	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	pcmdi11.llnl.gov	198.128.245.161	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	ssa.lasg.ac.cn	210.75.240.163	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	esgdata.ofcl.noa...	140.208.31.117	100%	100%	100%	99%	99%
<input checked="" type="checkbox"/>	suclipse1.dkrz.de	136.172.30.13	100%	100%	100%	98.67%	98.33%

Map



Host information

Host: esgf-dev.dkrz.de
 IP: 136.172.30.117
 Peer Group: esgf-dev.dkrz.de
 City: Hamburg
 Total services: 1 [Show Details]
 Permalink: code
 Time Interval: 5 minutes

Host Availability
 [06/18/2012 10:51 - 06/18/2012 10:56]

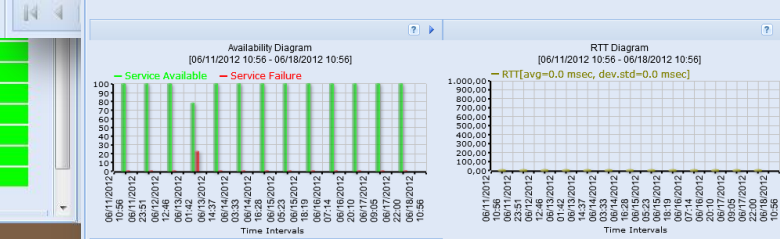
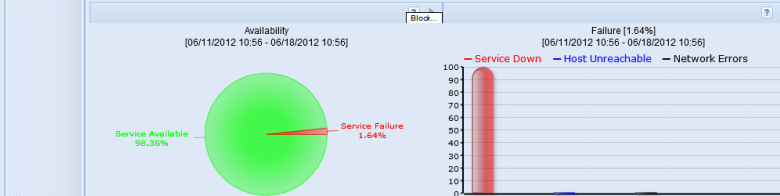


Services List [Reference date 06/18/2012 10:56]

Peer Group: esgf-prod Host: esgf-dev.dkrz.de Application Server

Service

Settings Service Dashboard Availability Failure Availability Diagram RTT Diagram Summary



Summary

Service Name	Application Server	Host	esgf-dev.dkrz.de
Port	80	Peer Group	esgf-dev.dkrz.de
E-Mail	admin@esgf-dev.dkrz.de	Institution	dkrz
Start Date	05/03/2012 10:12	End Date	
Availability (period)	98.36%	Availability (total)	98.53%
RTT (period)	0 msec	RTT (total)	0 msec
Dev.Std. RTT (period)	0 msec	Dev.Std. RTT (total)	0 msec

...ity Last week availabi... Last month availa...
 98% 98%

Displaying topics 1 - 1 of 1

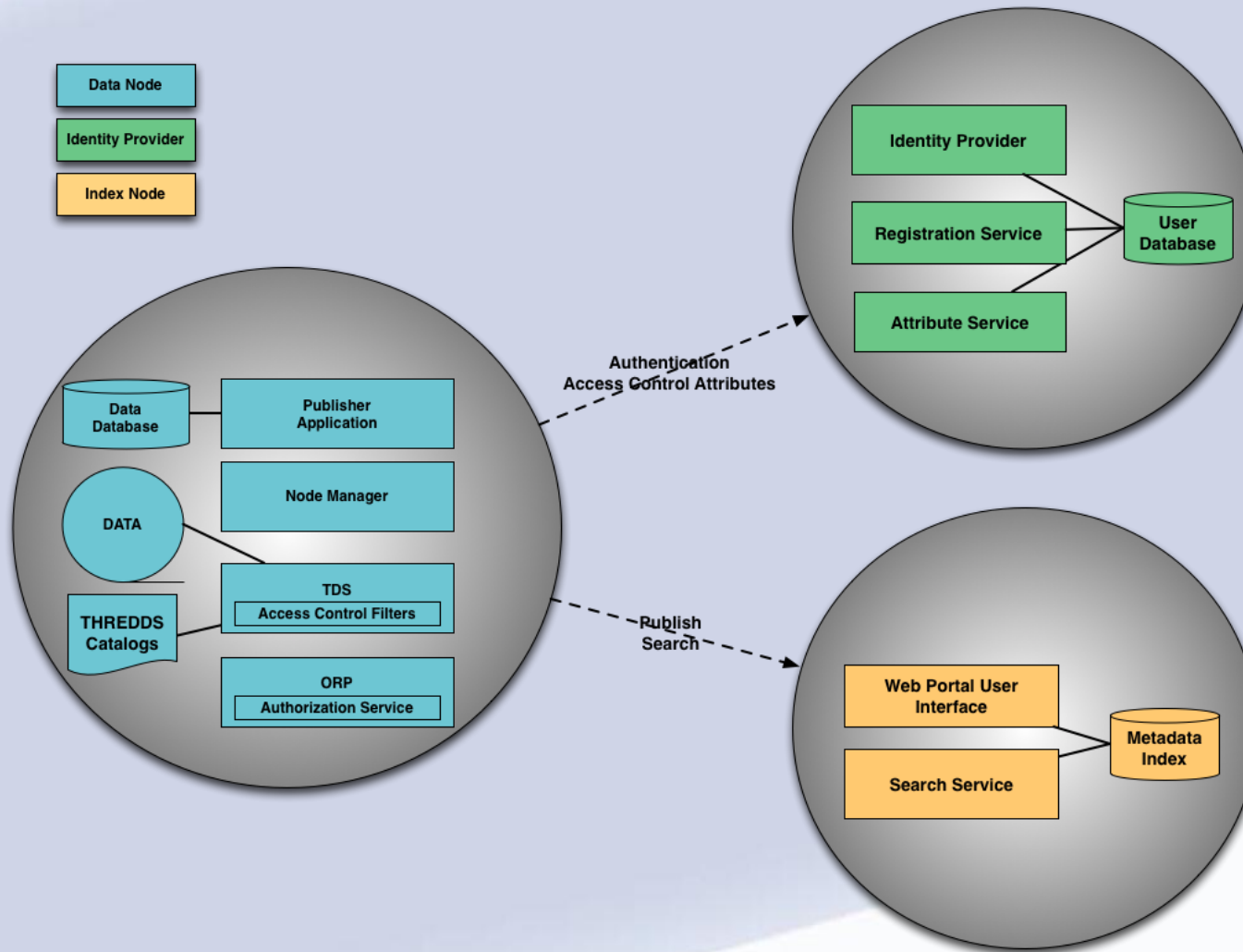
User: https://esgf-dev.dkrz.de/esgf-ldp/openid/rootAdmin
 ESGF P2P Version 1.3.2-66-g2abf5a3-devel [fe-2.1.8] Privacy Policy & Legal Notice | Contact ESGF



ESGF P2P Manager

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

ESGF P2P Interactions



Thank you.