Accessing Earth Science Data from the EGI Infrastructure

EGI-InSPIRE TSA3.6

EGI User Forum 2011

Horst Schwichtenberg <horst.schwichtenberg@scai.fraunhofer.de>
André Gemuend <andre.gemuend@scai.fraunhofer.de>

TSA3.6: Background

- EGI-InSPIRE WP6 TSA3.6: Services for Earth Science
- Focus on access to data infrastructures & access methods
- Attempting to bring access to data infrastructures to users of the EGI Earth Science VOs
  - find opportunities for collaboration
  - pave the way for registration / acceptance of use

TSA3.6: Background

- High number of variables typically depending on lat/lon/alt/time
  - Temperature
  - Pressure
  - Radiation
  - …
- Should be a good basis for uniform interfaces to access data, but…

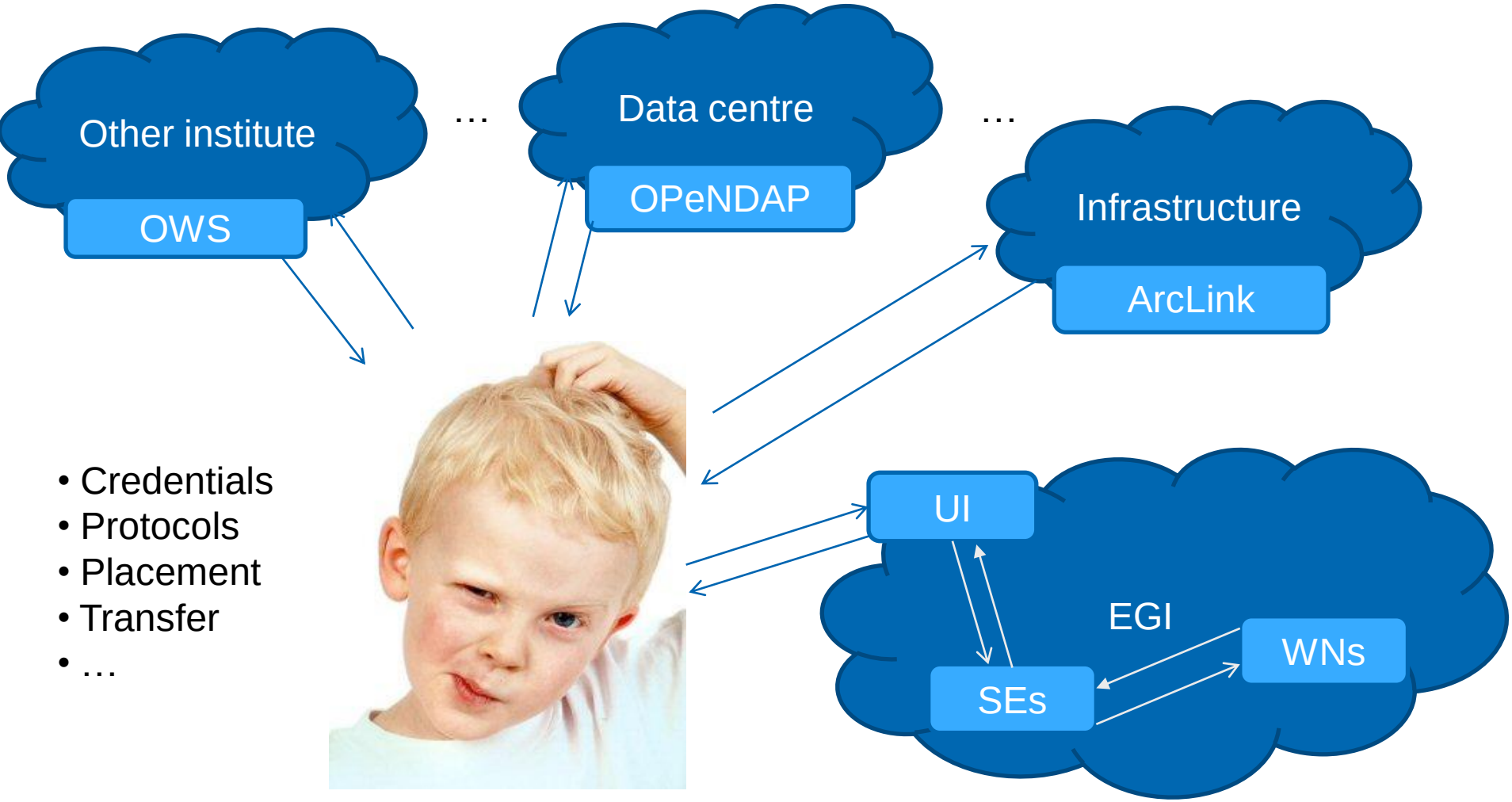
TSA3.6: Background

- High diversity in all aspects
  - Global distribution of completely separate acquisition & data centres
  - Different administrative domains, responsibilities, laws
  - Technical systems: measurement instrument, storage methods, formats, metadata, access methods…
  - Seperate data infrastructures with specific characteristics

TSA3.6: Background

- Users not satisfied with this strong separation
- Workflow for using data in EGI
  - Somehow know where to search for the data
  - Download with corresponding tool (web clients, OPeNDAP, etc.)
  - Get it to a place where you have access to Grid
  - Upload data to a storage element or put it in your input sandbox
  - For every job, sometimes from different sources

TSA3.6: Background



- Credentials
- Protocols
- Placement
- Transfer
- …

TSA3.6: Background

- Existing data infrastructures
  - usually organised around a specific goal or source, e.g.
    - satellite data (e.g. GENESI-DR initially)
    - climate data (e.g. Earth System Grid)
  - not so concerned about general interoperability
  - building (user) interfaces for this use case
- Challenges:
  - Interfaces for external access available? (how to exploit this for EGI)
  - Registration, Authentication, Authorization

TSA3.6: Facets

Discovery

- Availability
- Locality (Data center)
- Policy (rights)

Access

- Access channel / Protocols
- AuthN & AuthZ
- Performance

Processing

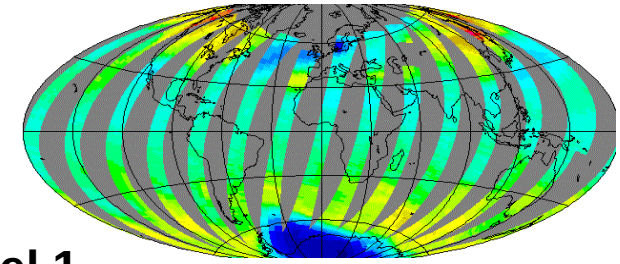
- Usage rights
- Formats
- Transformation
- Analysis

Archiving

- Traceability / Reproducibility
- Storage

GOME scenario

Raw satellite data
from the GOME instrument
(~75 GB - ~5000 orbits/y)



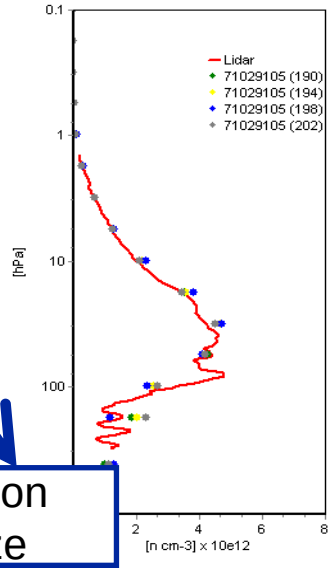
Level 1

(example of 1 day total O_3)

ESA(IT) – KNMI(NL)
Processing of raw GOME
data to ozone profiles.
2 alternative algorithms
~28000 profiles/day

Level 2

IPSL(FR)
Validate some of the
GOME ozone profiles (~10^6/y)
Coincident in space and time with
Ground-Based measurements (LIDAR)



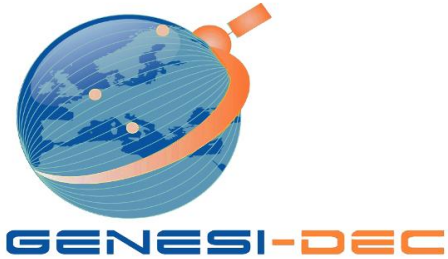
Grid environment

Visualization
& Analyze

e-collaboration in ESR VO: share data, build common metadata catalogue with geospatial extension… → 2 papers, 1PhD

TSA3.6: GENESI-DR

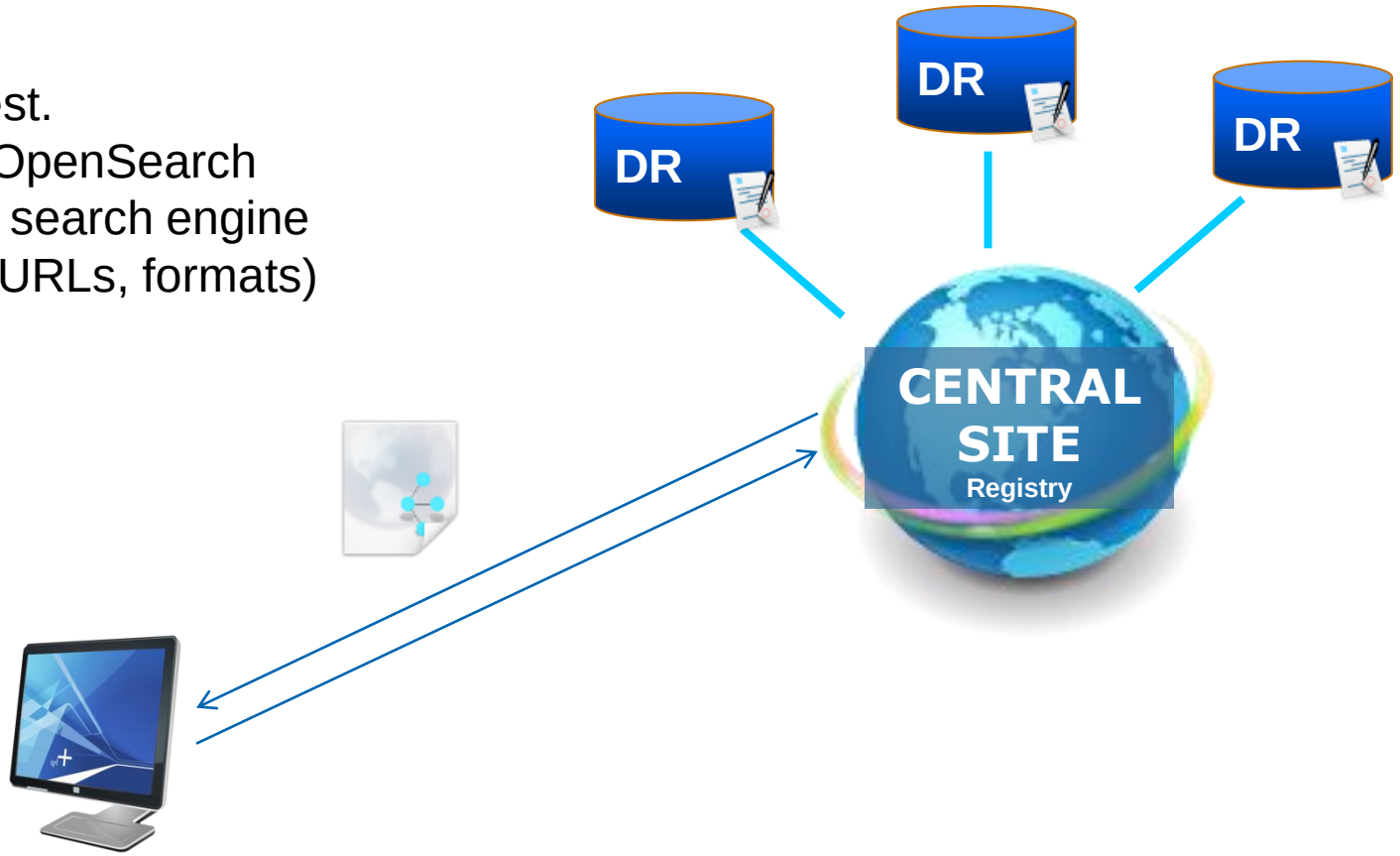
GENESI-DR → GENESI-DEC



- FP7 project led by ESA
- Find data through description and geospatial parameters
- Federated metadata search based on OpenSearch specification extended with GeoSpatial properties
- Started with earth observation satellite data
- Extending to other ES data sources and disciplines

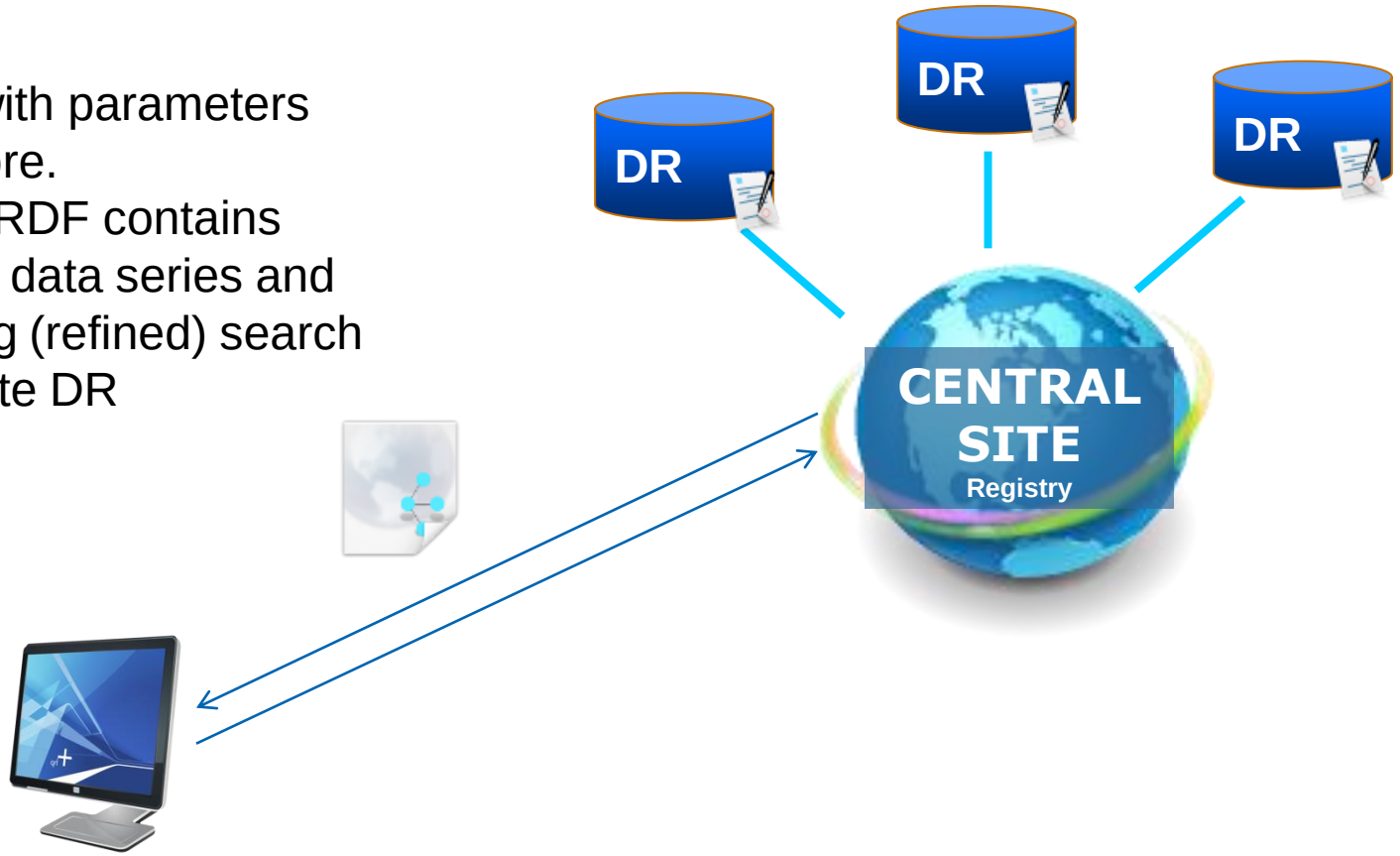
TSA3.6: GENESI-DR

Empty Request.
Response is OpenSearch description of search engine (parameters, URLs, formats)



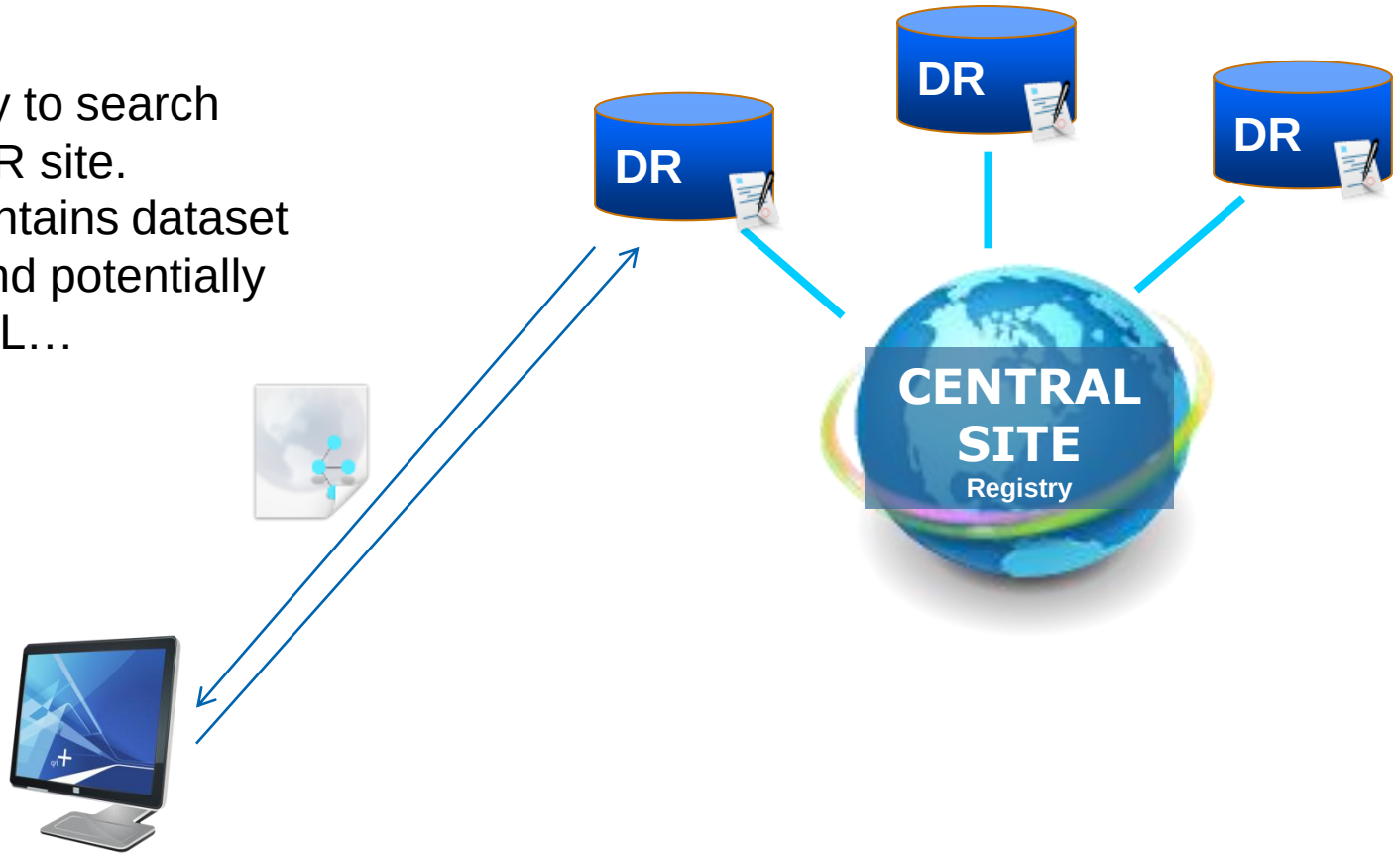
TSA3.6: GENESI-DR

Initial query with parameters provided before.
Response in RDF contains description of data series and corresponding (refined) search interface at site DR



TSA3.6: GENESI-DR

Refined query to search instance at DR site. Response contains dataset description and potentially download URL…



Example data set description:

```
<dclite4g:DataSet rdf:about="http://[…]/2011-02ALGO2.FIRE.gz.gz/xml">
    <dc:identifier>2011-02ALGO2.FIRE.gz.gz</dc:identifier>
    <dclite4g:series rdf:resource="http://dr-site.esrin.esa.int/catalogue/genesi/ATSR-WFA-2/xml"/>
    <dclite4g:onlineResource>
      <ws:HTTP rdf:about="http://due.esrin.esa.int/wfa/data/201102ALGO2.FIRE.gz.gz" />
    </dclite4g:onlineResource>
    <ical:dtstart>2011-02-01T00:00:00.000Z</ical:dtstart>
    <ical:dtend>2011-02-28T23:59:59.000Z</ical:dtend>
    <dct:spatial>POLYGON((-180 -90,-180 90,180 90,180 -90,-180 -90))</dct:spatial>
    <dclite4g:quicklook rdf:resource="http://due.esrin.esa.int/wfa/data/sg201102ALGO2.gif"/>
    <dct:created>2011-03-25T15:29:02.761Z</dct:created>
    <dct:dateSubmitted>2011-03-25T15:29:02.761Z</dct:dateSubmitted>
    <dct:modified>2011-03-25T15:29:02.761Z</dct:modified>
</dclite4g:DataSet>
```

TSA3.6: GENESI-DR

- Collaboration started in EGEE III Earth Science SDC
- Achievements
  - Scenario accessing EGI from GENESI-DR
    - Using data from EGI storage elements and data center in GENESI
    - For users of the GENESI-DR portal
  - Proof of concept to use GENESI-DR in gLite job (retrieving data sets of different series corresponding to input files)
    - Ozone profile validation with GOME / LIDAR measurements

TSA3.6: GENESI-DR

- Situation changed
  - Started with the idea of an interfacing web service
  - GENESI had own PKI-based auth with custom CA
  - changed to public plain HTTP
  - Plan on computing facilities changed, environment can‘t be accessed from EGI (only through GENESI portal)
  - GENESI-fication of own data from EGI required manual interaction, but will be automatized during GENESI-DEC

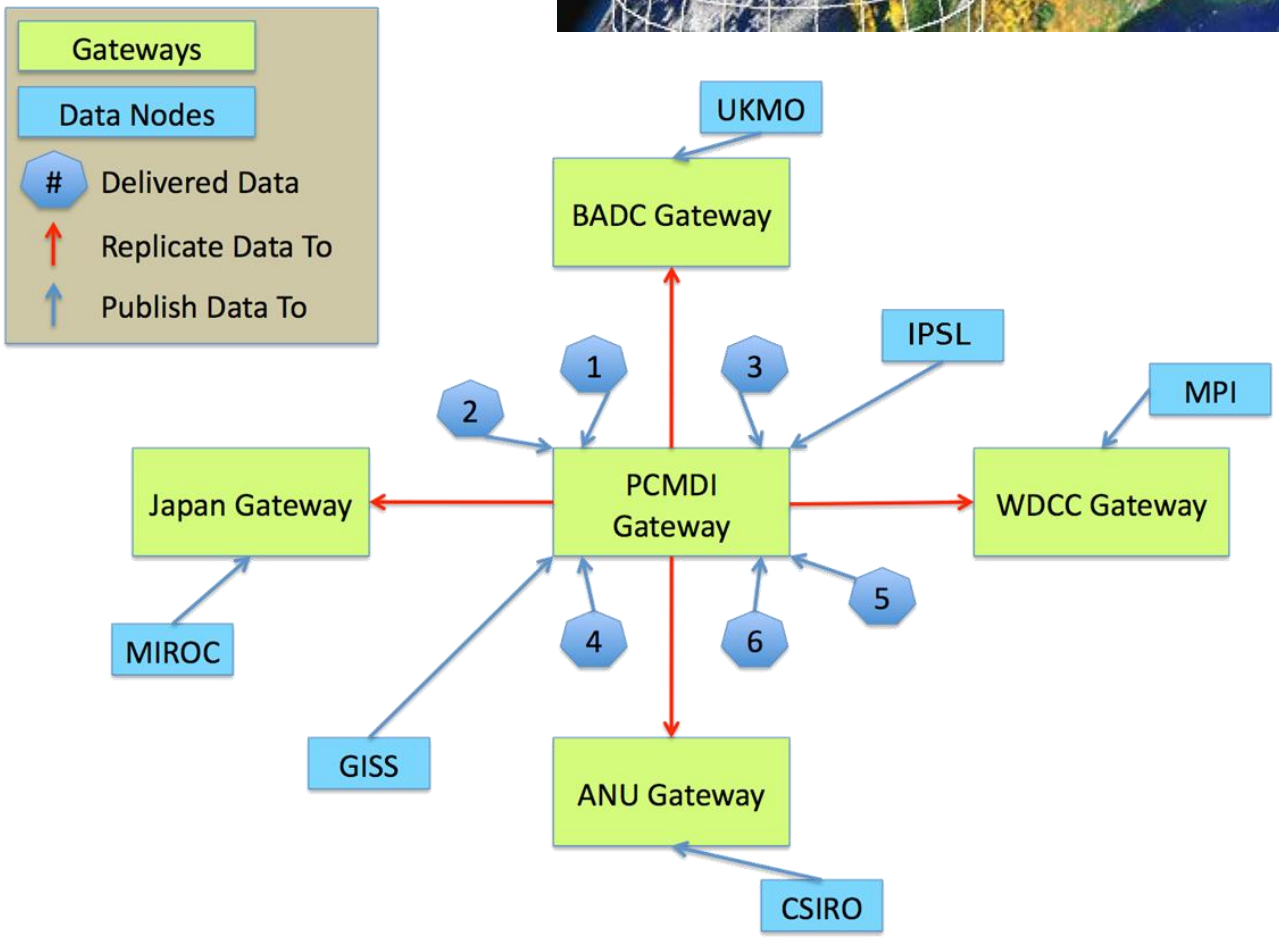
TSA3.6: GENESI-DR

- Wrote a command line client to search & download (accessible) data set files for easy integration in job scripts
- AuthN & AuthZ not dealt with in GENESI
  - different schemes at the sites, registration, access control,...
  - Can‘t solve this ultimately, but can try to facilitate
- Web user interface connected to Grid job submission
- Following GENESI-DEC developments regarding publishing to GENESI-DR from EGI results
  - Either with the same technology but EGI specific or accessible through GENESI central-site

TSA3.6: Earth System Grid

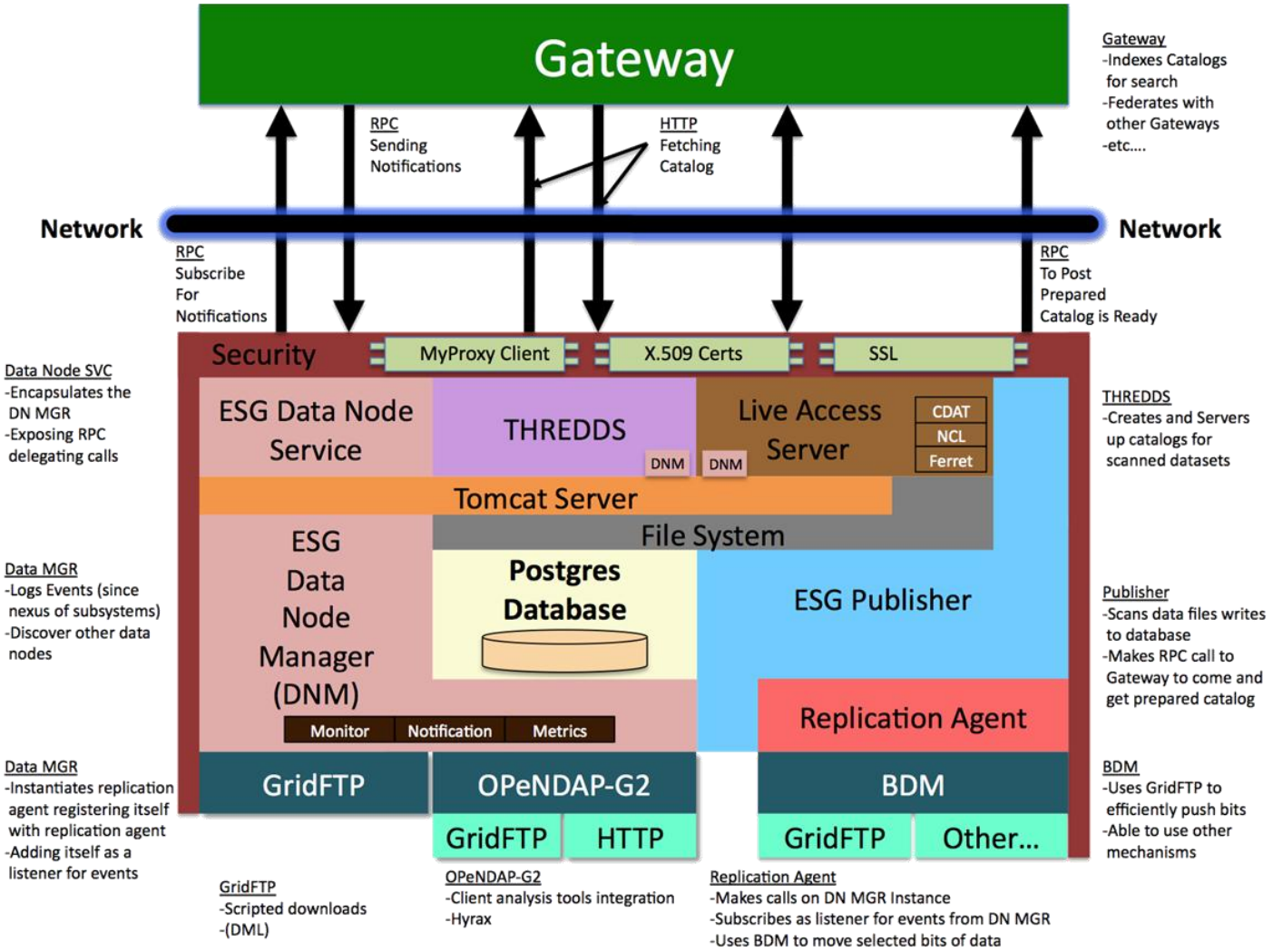
- Grid of data providers
- “Grid sites” ⇔ “data nodes”
  - Specified software components
  - Data server, catalogue, access control method, registration, etc. all uniform
- Web-based access through portals and API access for applications

TSA3.6: Earth System Grid



Source: Building on the CMIP5 effort to prepare next steps …, Sébastien Denvil (IPSL), May 2010

TSA3.6: Earth System Grid



Source: Building on the CMIP5 effort to prepare next steps …, Sébastien Denvil (IPSL), May 2010

TSA3.6: Earth System Grid

- Climate model output data
  - Model intercomparison
- Regional climate variation
- Statistical analysis
- Impact studies agriculture / insurance companies…
- Many interested users

TSA3.6: Earth System Grid

- Working group in TSA3.6
  - IPSL & SCAI
  - Bottom up approach
  - Starting with simple access to a data node
    - Scenario using a MPI application for statistical evaluation of data rows
    - Login to ESG PKI on gLite UI
    - Download data to worker nodes
    - In contact with security people from ESG Federation
    - Face to face meeting in preparation

TSA3.6: Earth System Grid

- Possible later tasks
  - Facilitate authentication
  - Cache data in Grid
  - Search / find data sets / replicas using the THREDDS catalogues or (possibly) results of the Prodiguer project
  - Managed file transfer
  - Additional functionality @ data node (e.g. subsetting)
  - Aggregate LAS functionality (or provide instance?)
- Discussing opportunities with ESG and Prodiguer participants

TSA3.6: Authentication

- Recurring problem
  - Dedicated community infrastructure (e.g. ESG)
    - Starts with user database, often evolves to OpenID authentication and SAML token authorization
    - Even if PKI authentication is respected, it will in most cases rely on custom CAs or other incompatible specifics (c.f. id string in DN)
  - No AuthN / AuthZ at all (e.g. OWS)
    - How to include it later on to access the Grid?
  - AuthN / AuthZ existing, but not unified (e.g. GENESI-DR)
    - Can we simplify this for users?

TSA3.6: Authentication

- There should really be a centralized „service“ for this
  - See for example CILogon project of Illinois
  - State of the AAI projects?
- If there is mutual trust, we can „translate“ the credentials
  - E.g. WS-Trust Security Token Service (EMI plans?)
  - Kind of „Credential store“?

TSA3.6: Authentication

- EMI is planning an STS
  - probably for year 3
  - need something in the meantime
- ESG is very open for collaboration and discussion is ongoing
  - Perhaps a MyProxy SimpleCA accepting EGI proxies to return ESG compliant short-lived end entity certificates ☺
- Open to suggestions / collaborations

Summary

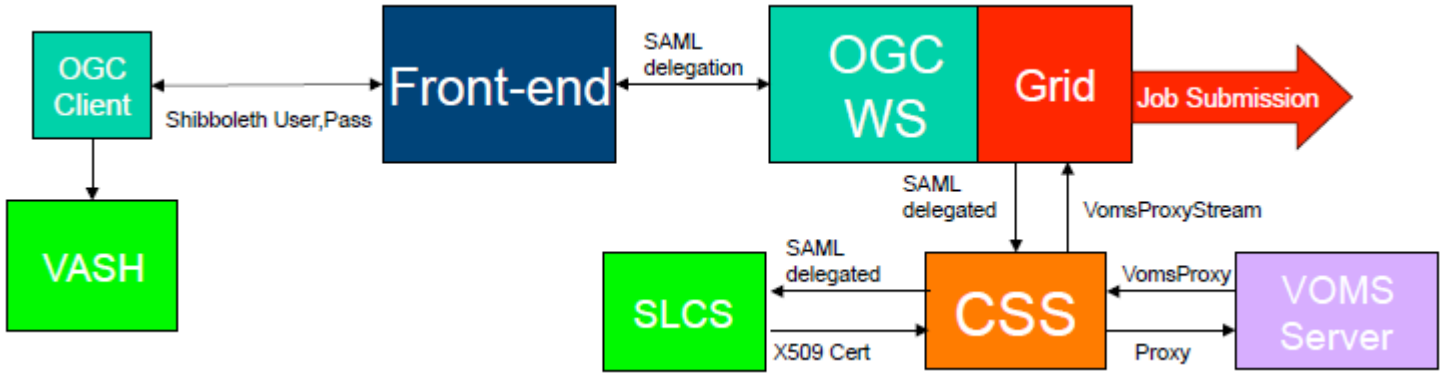
- Different aspects of problems with connecting data infrastructures to EGI
- Data providers use different technologies, own registration, authN, authZ, …
- Collaboration with ESG as an important network for earth science data can be exemplary for the future
- A lot of tasks, but good foundation and start

End

Questions?

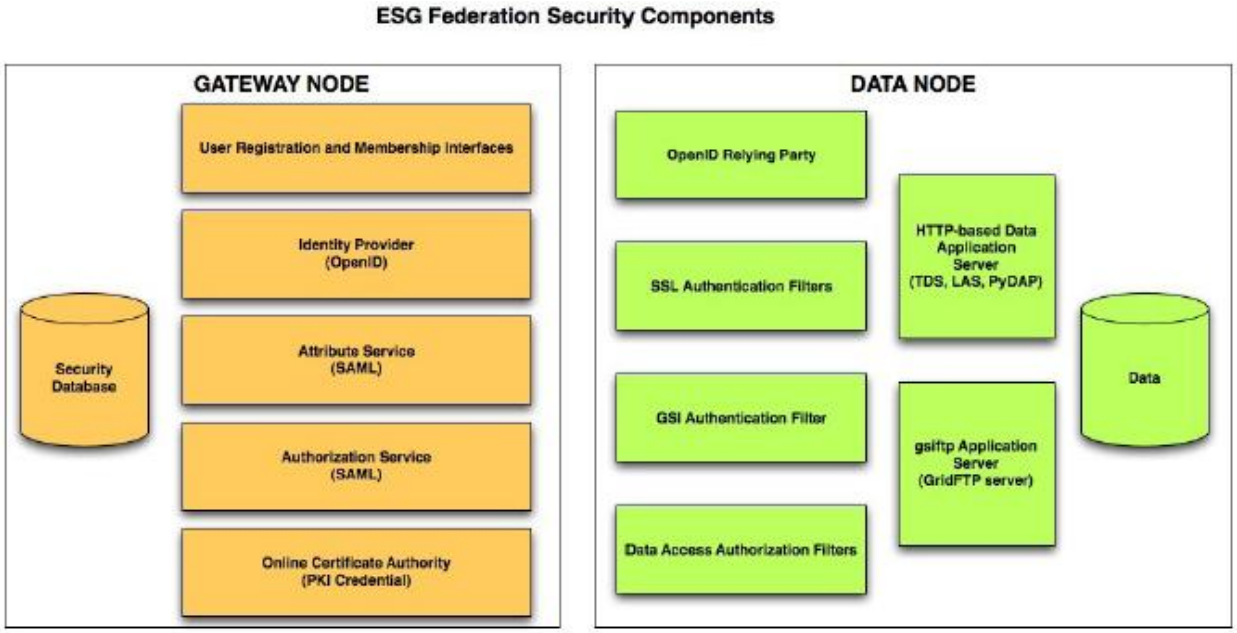
TSA3.6: Authentication

- Approach: G-OWS
  - short-lived credentials for Shibboleth identities (SLCS)
  - service to register at VOMS (VASH)
  - service to return voms proxy at runtime



TSA3.6: Authentication

- Approach: ESGF security
  - OpenID authentication for portals
  - MyProxy SimpleCA for short-lived credentials at every IdP
  - Using the same LDAP backend at the IdP
  - DN contains OpenID string, e.g. „*/O=ESG Org/OU=Climate Modeling Group/CN=http://myname.openid.esgorg.ac.uk*“



TSA3.6: Authentication

- Approach: Security Token Service
  - Part of WS-Trust standard
  - Issues WS-Security compatible tokens
  - General mechanism, kind of „negotiation“, because all involved parties can claim own requirements
  - Strong industry backing: e.g. IBM, Microsoft, RSA,...