Data Centre Inter-Operability
– DCIO –
a practical exchange approach

Yasjka Meijer *et al.*
European Space Agency
Frascati, Italy

An initiative to stimulate Data Centre Inter-Operability

- Developed by data centres in close contact with data providers; community approach
- ESA has interest through GECA project:
  1. to harmonise cal/val data exchange,
  2. to benefit from data available from different sources
- GECA requires access to correlative datasets from multiple EO domains

DCIO objectives

Objectives:

- Expose data in your DC to more users
- Get access to a wide range of datasets:
  - Exchange catalogue information
  - Exchange data files

Explore to

- Harmonise data exchange agreements
- Harmonise metadata standards

Motivation & requirements

- Respect DCs' integrity; data protocol, etc.
  - no data copying or duplication across DCs
- Allow expandability of services
- Automated metadata exchange
- Automated data file exchange,
  i.e. exchange data location → URL
- Single-sign on to facilitate data access
- Feedback mechanism on data usage

DCIO working set-up 1/2

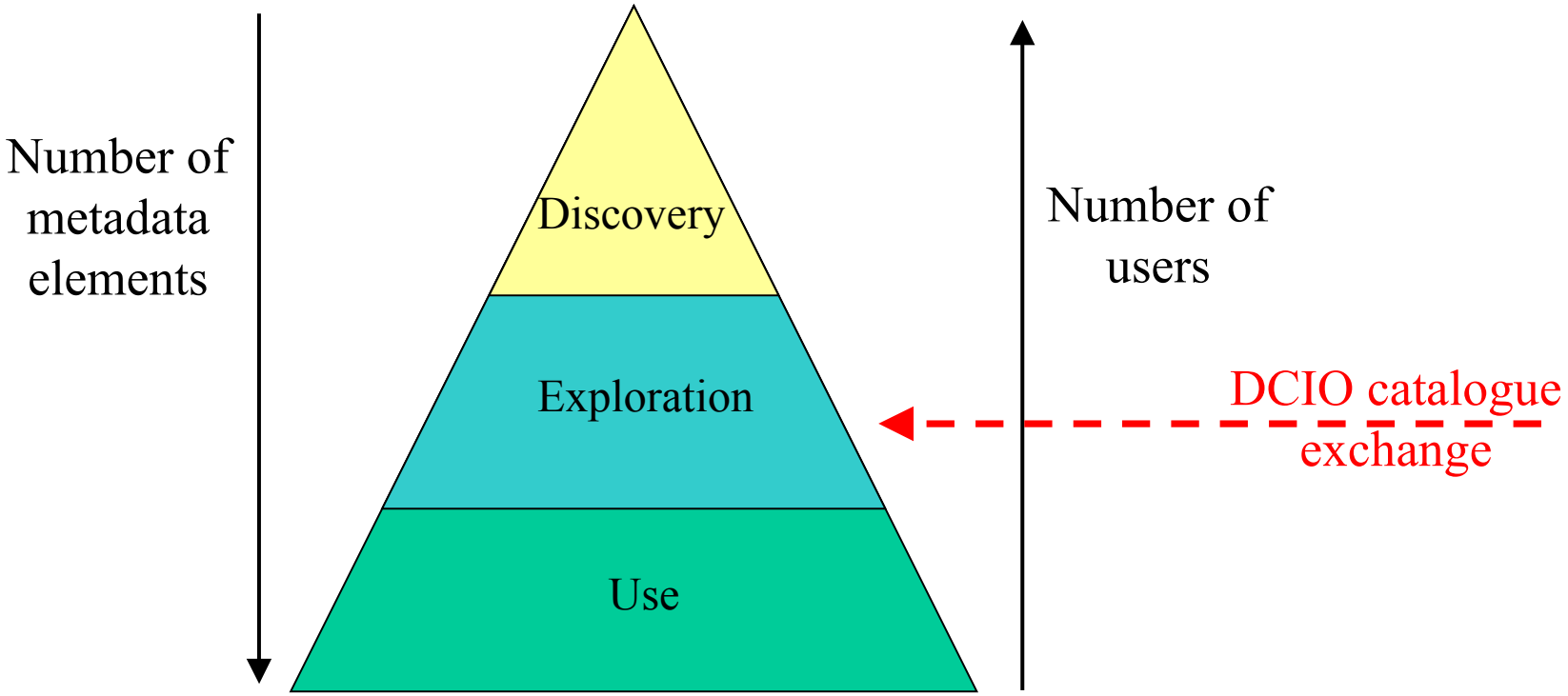
Initiative is led by ESA, started in 12-2008

- 26 participants and growing
- 13 data centres & exchange initiatives
- Now had 14 telecons and 1 meeting
- Every 1–2 months a telecon
  (using toll-free numbers)
- Every 1–2 years a meeting, preferably coinciding with another event
- Email exchange on specific topics

DCIO working set-up 2/2

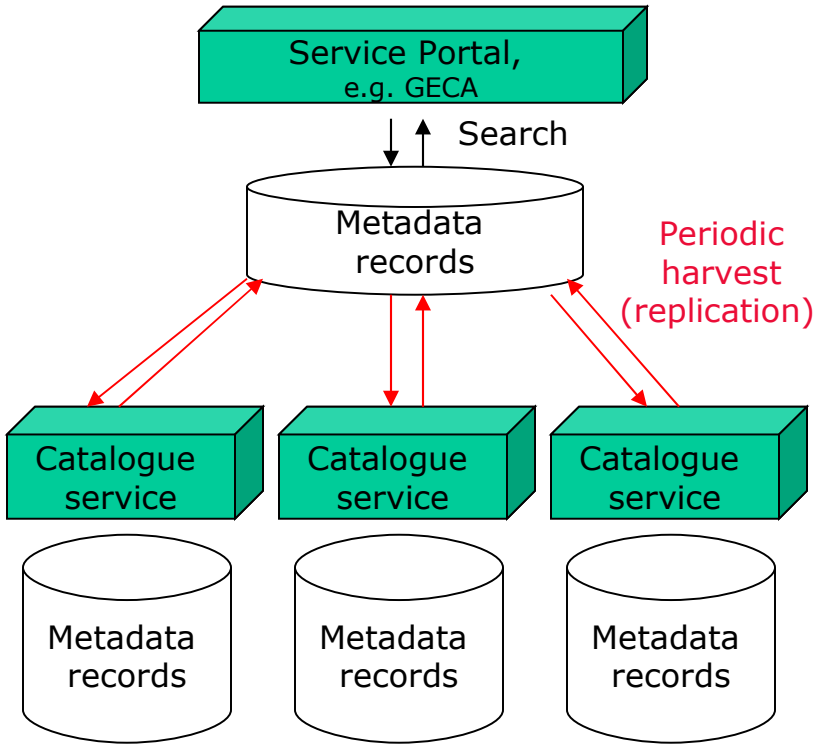
Data Centre	Main focus
AVDC (NASA)	Satellite validation
AERONET (NASA) *	Research and monitoring
Ceilometer Network (German)	Research and monitoring
Earlinet (European)	Research and monitoring
EVDC (NILU/ESA)	Satellite validation
GeoMON (European)	Monitoring; data exchange/exploitation
GEOSS (Internat.)	Data exchange/exploitation
GlobWAVE (ESA) *	Data exchange/exploitation
MyOcean (EU) *	Long-term monitoring/ Support to validation
NDACC (Internat.)	Long-term monitoring/ Support to validation
Wegener Center, RO sat.	Satellite validation
WIS (WMO & Internat.)	Data exchange/exploitation
WOUDC (Internat.)	Research and monitoring

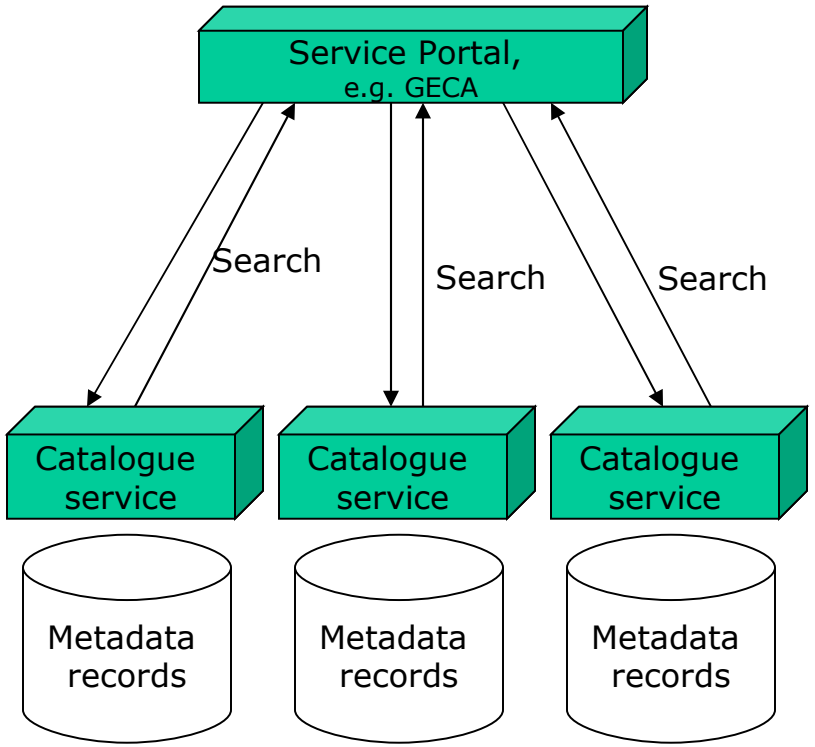
Metadata levels



Context: 1) catalogue metadata,
2) metadata standard,
3) data file format

Metadata Harvesting vs Distributed Search





- **Harvest**
  - Advantages: quick searches and no need for peer to support querying of all metadata fields
  - Disadvantage: metadata duplication
- **Distributed Search**
  - Advantage: metadata maintained closer to source and no duplication
  - Disadvantage: searches takes longer to complete, are more frequent requests and have more chances to be incomplete

Open Archives Initiative – Protocol for Metadata Harvesting

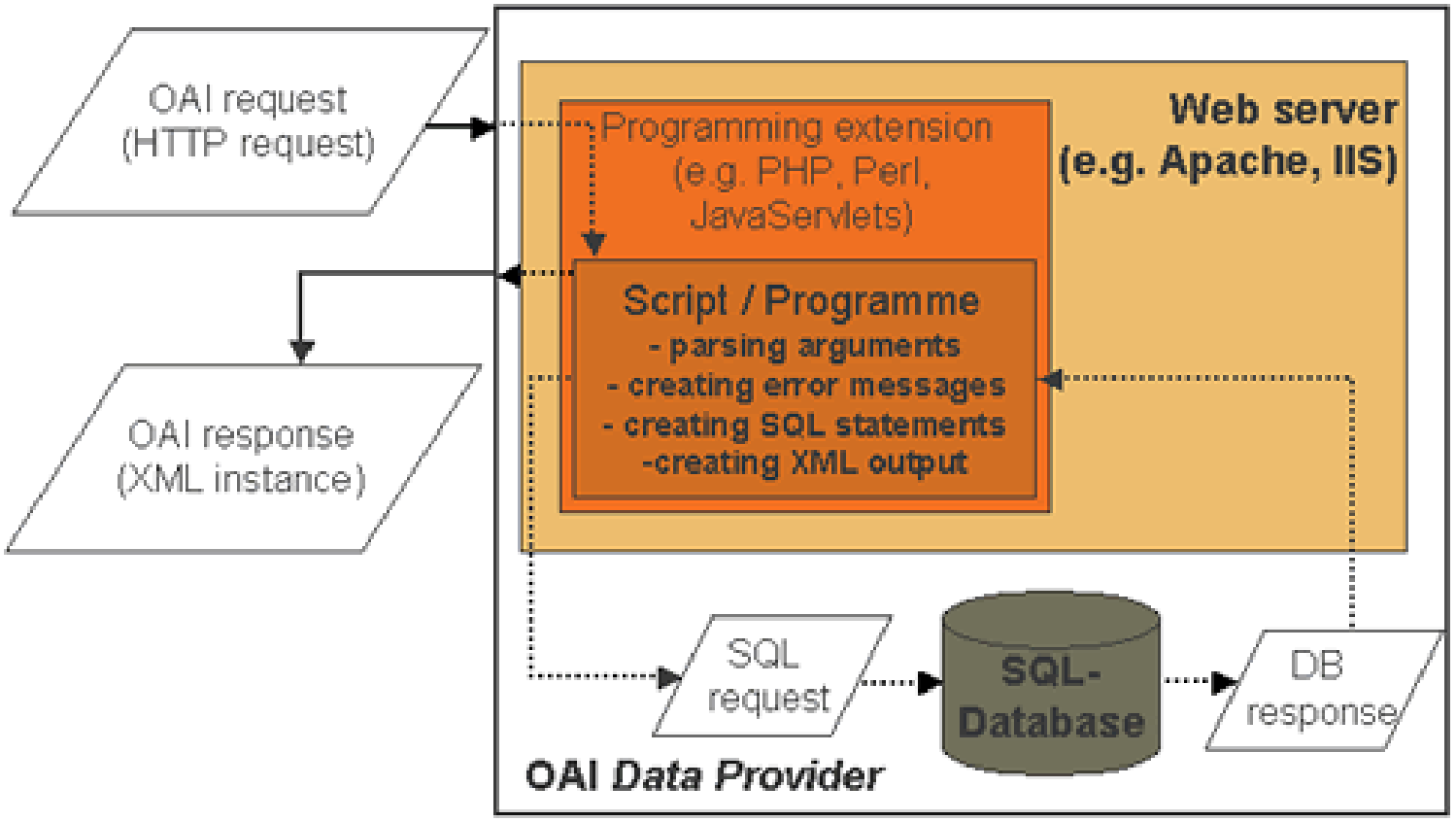
- Simple web service protocol for replication of catalogue content
- Employs XML formatted metadata over HTTP → firewall-friendly
- Metadata format:
  - Mandatory = return of Dublin Core metadata
  - Specific communities to develop specific metadata models & formats
- Version 1 in 2001; version 2 in 2002; no changes since → mature

- Originated in world of scientific "e-prints"
  but widely applicable by using different metadata models

OAI – PMH overview 2/2

- 2 “participants”
  - Data provider: exposes metadata
  - Service provider: uses harvested metadata
- 2 Software Components:
  - A repository is the server application that can process OAI-PMH requests
  - A harvester is the client application that issues OAI-PMH requests
- Open source tools exist but mapping from specific databases to specific community XML format to be programmed

Data provider “architecture”



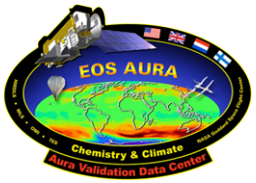
Data access & OpenID

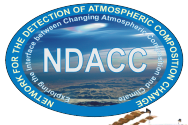
- Data files remain at source until needed
- URL in catalogue metadata
- Direct access with OpenID; so-called single-sign-on user authentication
- Users require access credentials with both the data centres and OpenID
  - No passwords are exchanged
- OpenID uses a central identity provider

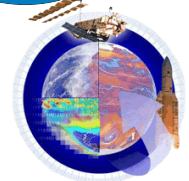
Data exchange agreement

- Initial steps were made toward a joint Data Exchange Agreement
  - Many overlapping usage rules
  - Data usage not charged
  - Data ownership remains to data originator
  - Notification about intended publications
  - Registration of usage
  - Acknowledgement of data owner
- DCIO participants allow direct access as long as usage statistics are provided

GEOMS:
Generic EO Metadata Standard



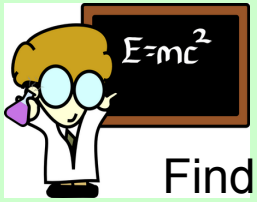




- GEOMS is a dedicated metadata standard for EO Cal/Val activities
- GEOMS has been established in collaboration with AVDC (NASA), EVDC (NILU/ESA), ESA, BIRA and NDACC
  - Initial focus on atmosphere,
  - BUT now also broadened to other domains
- GECA will adopt GEOMS format

Example: Catalogue access

Use case:



Find data

GECA server

Collocation Catalogue

Collocation tool

Correlative Data Catalogue

Satellite Data Catalogue

or anbody else, e.g., Earlinet or GALION

OAI-PMH

Periodic harvest of catalogue metadata

GENESI-DR

Periodic harvest of catalogue metadata

EOS AURA

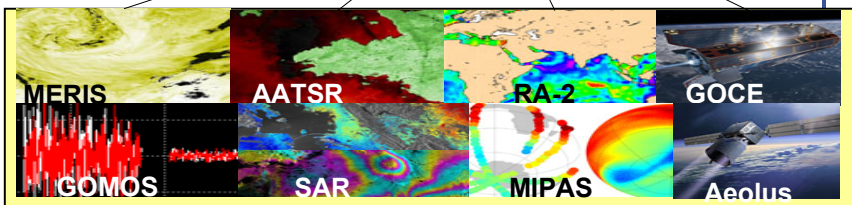
ENVISAT-1

NDACC

EARLINET

+?

Correlative data - DCIO



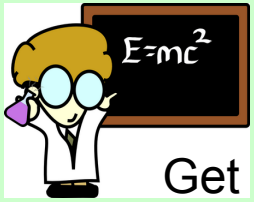
ESA satellite data

European Space Agency
Agence spatiale européenne

the Living Planet

Example: Data access

Use case:



Get data

Data can be used on server or downloaded to the user

GECA server

Collocation Catalogue

Satellite Overpass Catalogue

Correlative Data Catalogue

Satellite Data Catalogue

Data download access

Child products database

or anbody else, e.g., Earlinet or GALION

Agreements & OpenID will allow downloading from DCIO databases

GENESI-DR provide access details to ESA data repositories

Overpass IPF

EOS AURA Chemistry & Climate Aura Validation Data Center

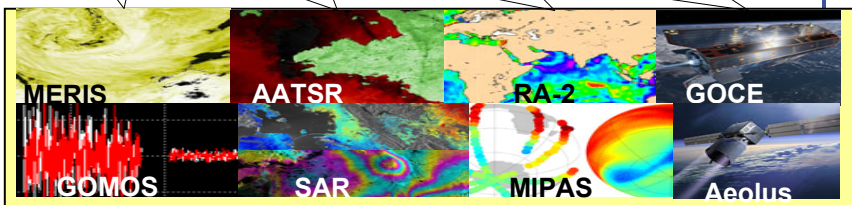
ENVISAT-1

NDACC

EARLINET

+?

Correlative data - DCIO



ESA satellite data

DCIO status of implementation

OAI – PMH catalogue exchange:

- AVDC: OAI-cat operational since 05-2011
- EVDC: OAI-cat operational since 06-2011
- Earlinet: discussions well-advance, possibly OAI-cat by end of 2011
- NDACC: implementation will start in 10-2011
- GECA: harvester tested

OpenID:

- GECA will host identity provider
- AVDC will adopt it; GECA to exploit it for all users
- Other DBs to be confirmed but technically easy
- Expected operational early 2011

Conclusion & invitation

- DCIO offers:
  - access to more data
  - visibility of your data
  - full control of your data files
  - different levels of exposure
    1. Just join discussions
    2. Allow catalogue exchange via OAI-PMH
    3. Allow public data access or via OpenID

- Participate to DCIO meetings by emailing Yasjka.Meijer@esa.int

Thank you!

DCIO: Data Centre Interoperability

- GECA will host correlative datasets of multiple EO domains
  - Requirement for interoperability between data centres
- Initiation of DCIO activity
  - Access to wider range of correlative datasets
- Current DCIO partners: AVDC, EVDC, NDACC and Earlinet
- Prototypes are working for exchange of catalogue meta-data
- Metadata catalogue in GECA will allow data in peer data centres to be visible
- Opportunity to join DCIO !