



EcoGrid Design & Architecture

the resource-access fabric for the

Science Environment for Ecological Knowledge

Arcot Rajasekar
San Diego SupercomputerCenter





What are Grids and Data Grids?

- ***Grids** are distributed (computational) systems that enable*
 - *sharing and aggregation of **resources***
 - *distributed across "multiple" administrative domains*
 - *based on availability and capability, and*
 - *meeting cost and quality-of-service requirements.*
-
- ***Data Grids** enable sharing of data and information resources*

■ Massive data scales

- Petabytes of data
- Billions of data sets
- Millions of data collections
- Hundreds of sites/resources,
- Thousands of users



■ Challenges

- Foster Collaboration
- Support Data-intensive Computing
- Ease Discovery and Usage



Data Grids Enable Collaboration

- Foster collaboration
 - Provide controlled data sharing
- Enable data-intensive computing
 - Ease data movement and management – multiple caches
 - Support super-computing and visualization
- Provide data persistence – library functionality
 - Long term storage and availability
- Support massive data scales
 - Secure access to Peta bytes of data
 - Serve several 100s of millions of data objects
- Facilitate discovery and usage
 - Metadata searching and attribute-based access
 - Unified data access system

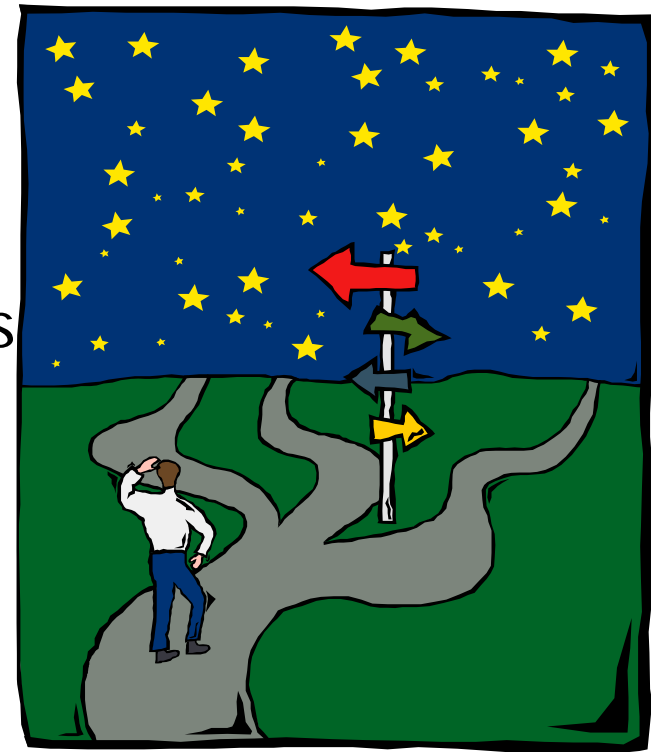
Data Handling Problems

- Large Datasets; Large Number of Datasets; Scaling
- Distributed, Heterogeneous Storage
- Collaboration, Access Control, Authentication, Security
- Replication, Coherency
- Fault Tolerance and Load Balancing
- Scheduling, Caching & Data Access
- Data Migration over Time & Space
- Data/Collection Curation
- Uniform Name Space
- Handling Legacy Data and Data/Resource Evolution



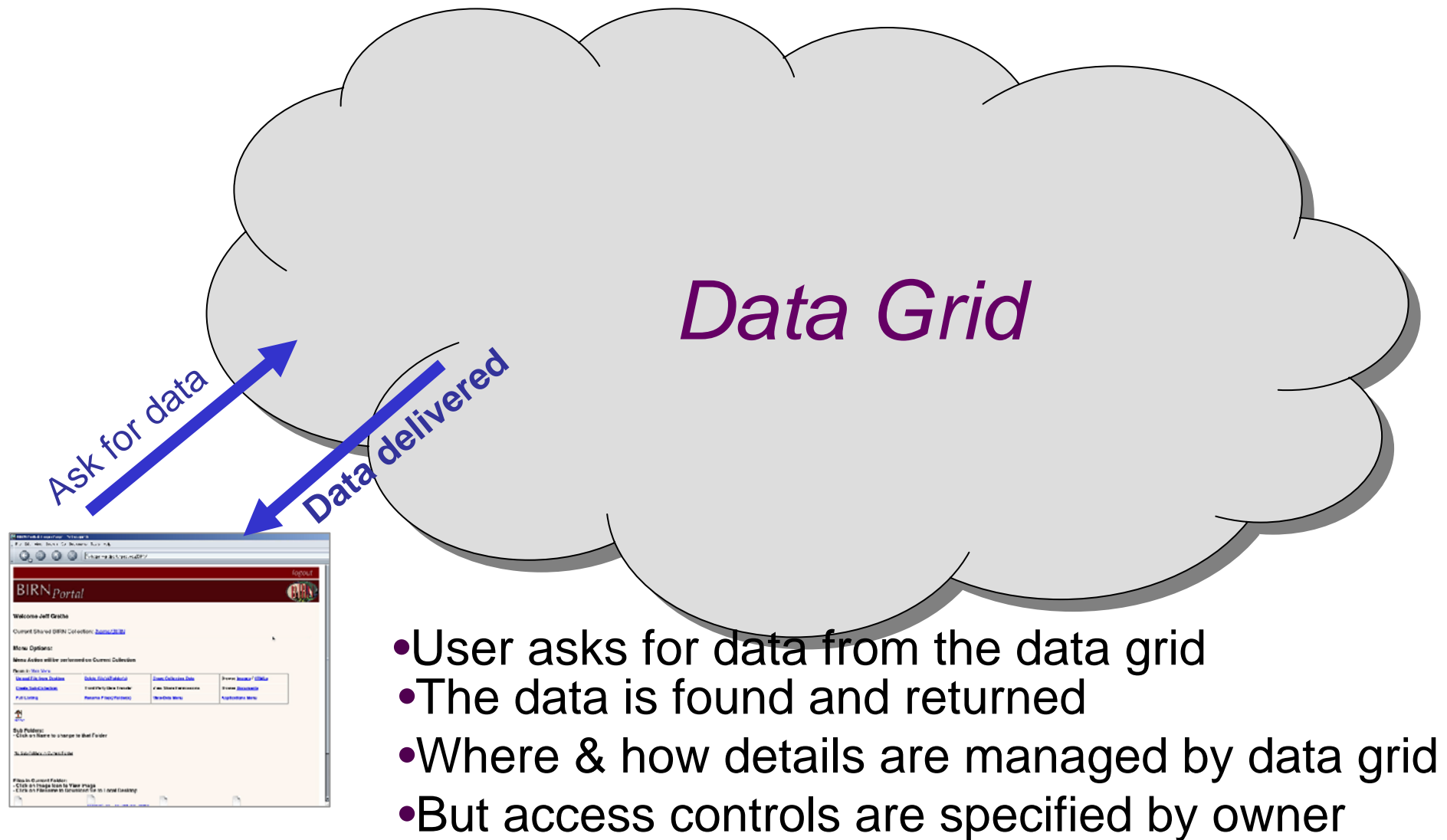
Metadata Problems

- Types of Metadata – Relational to XML to unstructured
- Standardized to User-defined Metadata
- Large Number of Attributes;
- Large Size; Scaling
- Federation - integration over space
- Evolution - integration over time
- Evolution - integration over contexts
- Discovery and Search
- Presentation – user friendly
- Extraction and Maintenance



- Low-level Semantic Integration
 - Schema Integration and Crosswalks
 - Ontological Differences & Translations
 - Context Dependency
 - Attribute Mappings
 - Value, Unit, Type/Semantic Conversions
 - Inter-domain & Intra-domain Integration

Using a Data Grid – *in Abstract*



- User asks for data from the data grid
- The data is found and returned
- Where & how details are managed by data grid
- But access controls are specified by owner

- Science Environment for Ecological Knowledge (SEEK)
 - Multidisciplinary research project to create:
 - Distributed data network (EcoGrid)
 - Environmental, ecological, and systematics data
 - Scalable systems for scientific analysis (workflow systems)
 - Systems for semi-automated data and model integration
 - Collaborators
 - NCEAS, UNM, SDSC, U Kansas
 - Vermont, Napier, ASU, UNC

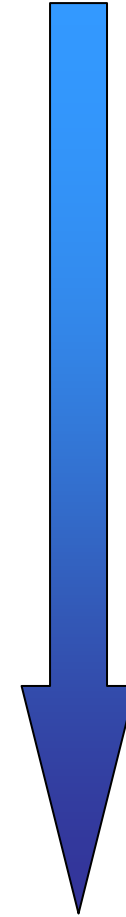


What is EcoGrid?

- Seamless access service to distributed data and metadata
 - scalability, multiplicity of platforms and storage devices
 - authentication through single sign-on authentication
 - multi-level access control,
- Maintain a registry for data, metadata, ontologies, services
- Allow rapid incorporation of new data sources as well as decades of legacy ecological data,
- Provide extensible, ecologically-relevant metadata based on the Ecological Metadata Language,
- Replicate and version of data to provide fault tolerance, disaster recovery and load balancing.
- Enable execution of applications as workflows
- Help semantic integration of data resources

Principal foci for SEEK

Planning
Metadata Entry
Data Acquisition
Quality Assurance
Storage and Access
Data Integration
Analysis and Modeling
Synthesis





```

graph LR
    Input[Raw Data  
1, 2, 3, 4, 5  
6, 7, 8, 9, 10  
11, 12, 13, 14, 15  
16, 17, 18, 19, 20  
21, 22, 23, 24, 25  
Metadata  
id, name, description  
1, John, Data Scientist  
2, Jane, Data Analyst  
3, Bob, Data Engineer  
4, Alice, Data Manager  
5, Charlie, Data Scientist] --> Step1[Metadata Parsing and Data Ingestion]
    Step1 --> Step2[Data Cleaning and Transformation]
    Step2 --> Step3[Analysis Algorithm 1]
    Step3 --> Step4[Analysis Algorithm 2]
    Step4 --> Step5[Output Generation and Visualization]
    Step5 --> Output1[ ]
    Step5 --> Output2[ ]
  
```

[illegible]

Raw Data x, y, z 1, 3, 6 3, 8, 5 1, 9, 0	Raw Data a, b, c, d 1, 3, 6, 4 3, 8, 5, 1 1, 9, 0, 1	Raw Data m, n, o 1, 3, 6 3, 8, 5 1, 9, 0	Raw Data p, q 1, 3 3, 8 1, 9
Metadata <pre><eml> <dataset> <title>Dataset 1</title> ...</pre>	Metadata <pre><eml> <dataset> <title>Dataset 2</title> ...</pre>	Metadata <pre><eml> <dataset> <title>Dataset 3</title> ...</pre>	Metadata <pre><eml> <dataset> <title>Dataset 4</title> ...</pre>

The diagram illustrates four different types of sequence data sources, each represented by a green rounded rectangle containing a white box with a blue triangle pointing right. The sources are labeled as follows:

- Genbank**: A white box with a blue triangle pointing right, labeled "Genbank".
- Eml200DataSource**: A white box with a blue triangle pointing right, labeled "Eml200DataSource".
- GISEquence**: A white box with a blue triangle pointing right, labeled "GISEquence".
- Diagrams**: A yellow box with a blue triangle pointing right, labeled "Diagrams".

Science Environment for Ecological Knowledge

- EcoGrid
 - Uniform interfaces to manage environmental data
- Kepler
 - Modeling scientific workflows
- Sparrow
 - “Smart” data discovery and integration

- Knowledge Representation
- Classification and Nomenclature
- Biodiversity and Ecological Analysis and Modeling

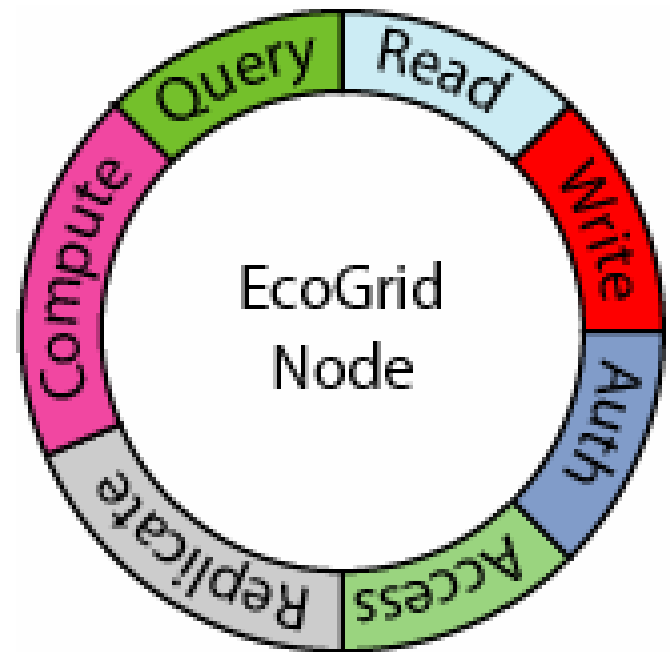


- Goal: standardize interfaces (using web and grid services)
 - We have standardized data via EML
 - Integrate diverse data networks from ecology, biodiversity, and environmental sciences

- Grid-standardized interfaces
 - Uniform interface to:
 - Metacat, SRB, DiGIR, Xanthoria, etc.
 - Anyone can implement these interfaces
 - Hides complexity of underlying systems

- Metadata-mediated data access
 - Supports multiple metadata standards
 - EML, Darwin Core as foci

- Computational services
 - Pre-defined analytical services
 - On-the-fly analytical services



EcoGrid client interactions

■ Modes of interaction

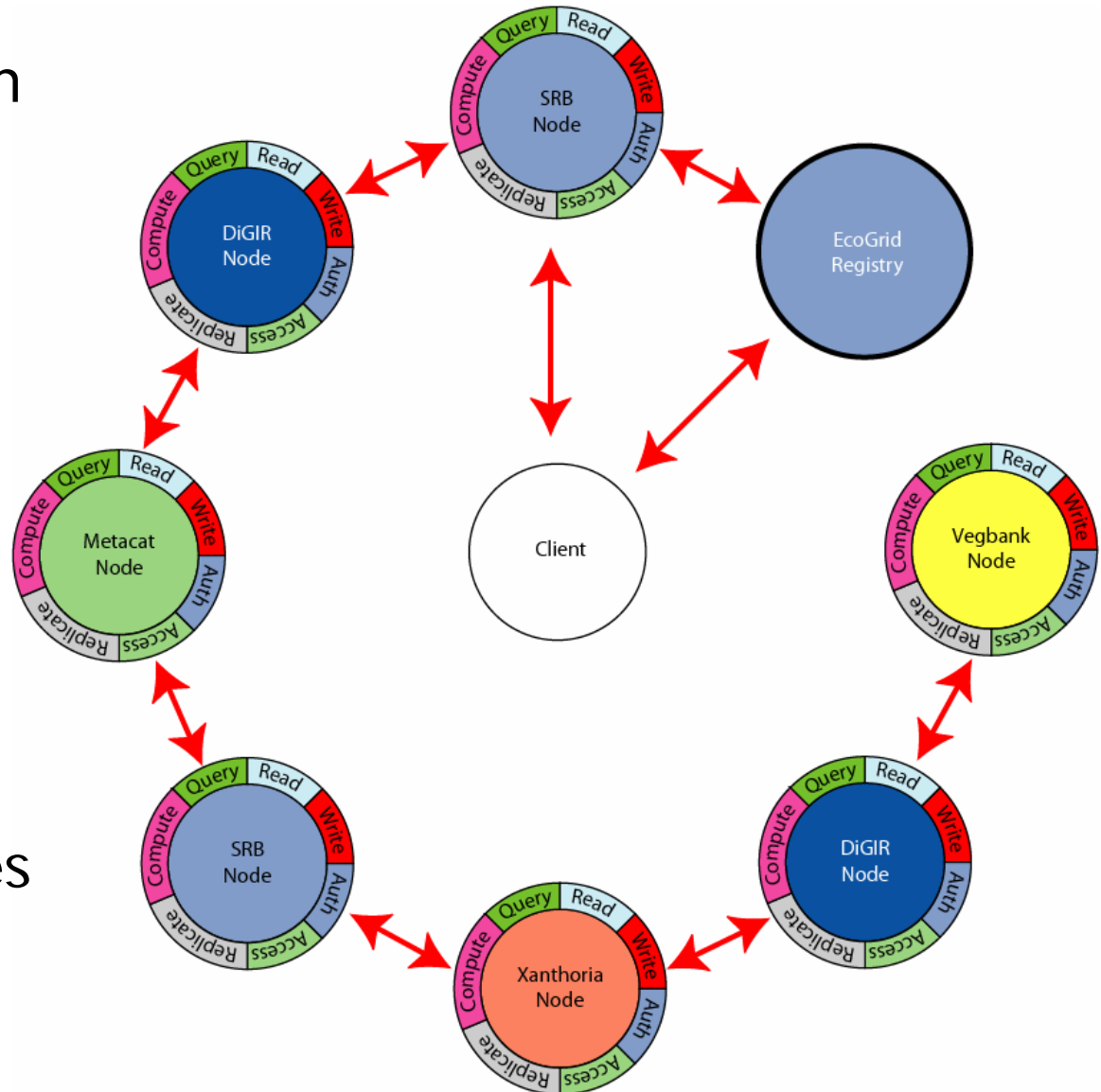
- Client-server
- Fully distributed
- Peer-to-peer

■ EcoGrid Registry

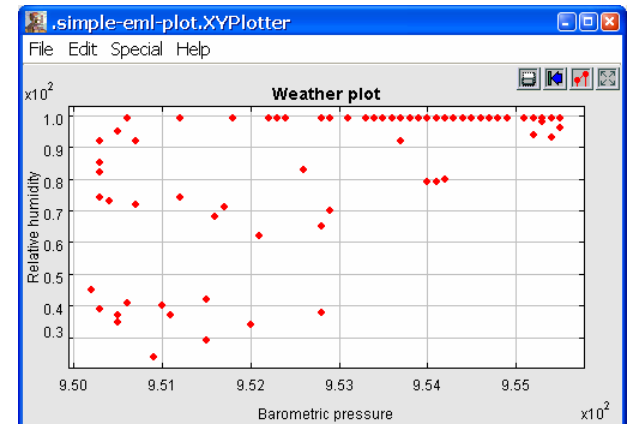
- Node discovery
- Service discovery

■ Aggregation services

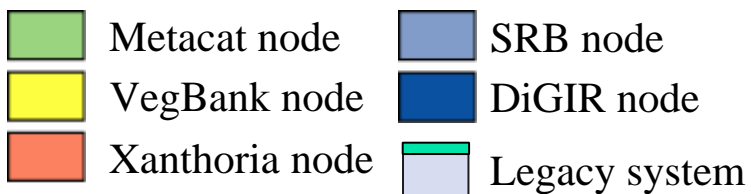
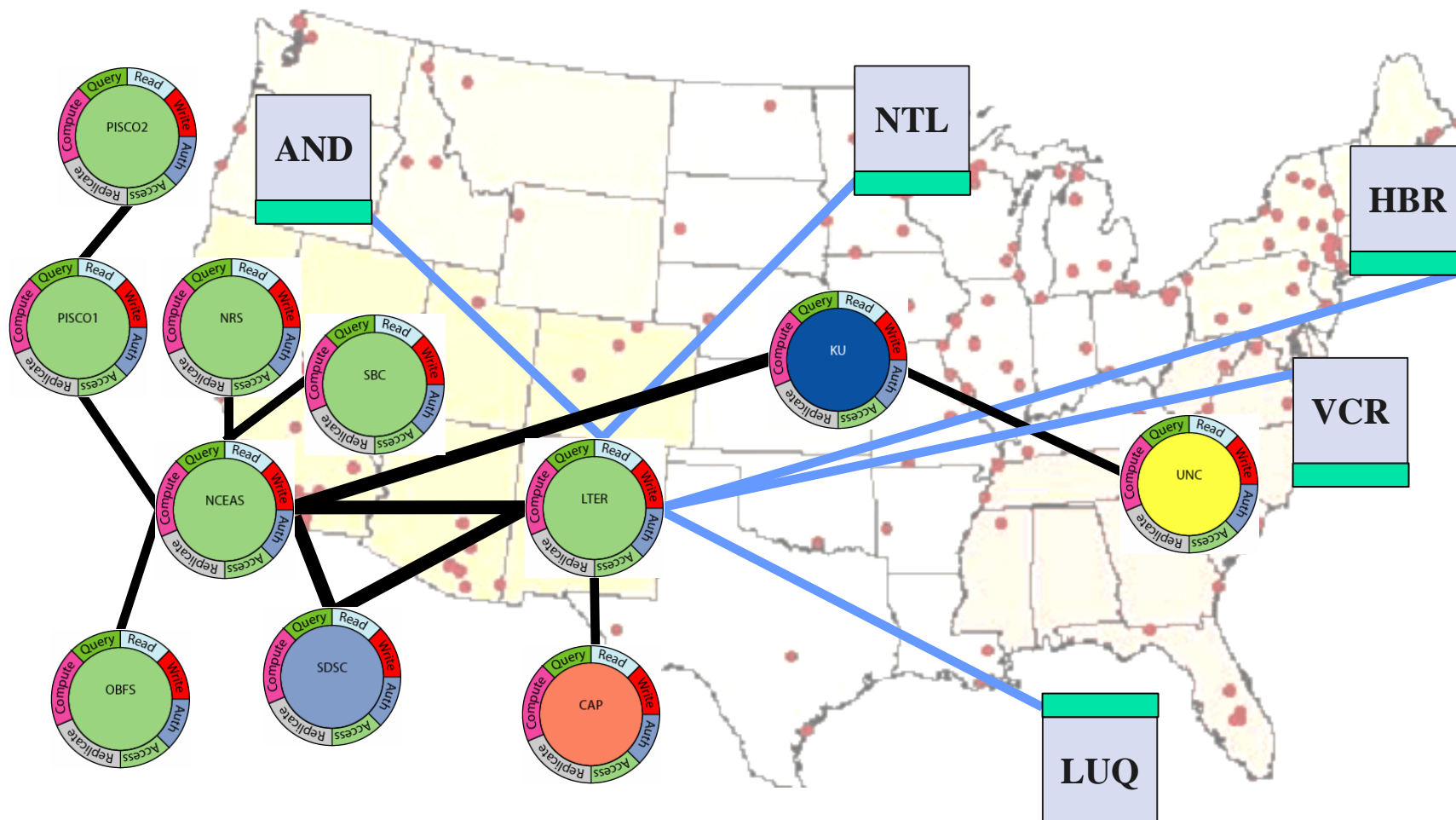
- Centralized access
- Reliability
- Data preservation



- Data and Metadata
 - Distributed Data
 - XML-based Metadata
- Service to Semantic Mediation Layer
 - Access to Ontologies and Taxon Services
 - Helping with Semantic Data Integration
- Service to Analysis and Modelling Layer
 - Interaction with Kepler - Workflows
 - Interaction with Grid Computing Facilities
- Access to Legacy Apps
 - LifeMapper
 - Spatial Data Workbench



EcoGrid Resources



LTER Network (24) Natural History Collections (>> 100)
 Organization of Biological Field Stations (180)
 UC Natural Reserve System (36)
 Partnership for Interdisciplinary Studies of Coastal Oceans (4)
 Multi-agency Rocky Intertidal Network (60)

- Metadata: a means to manage ecological data
 - There is no universal data model for ecology
 - Accommodate heterogeneity and dispersion

- EML
 - Common language for archiving and transporting data
 - Discovery information
 - Creator, Title, Abstract, Keyword, etc.
 - Content
 - Context
 - Physical, logical structure

 - SEEK will add semantic structure



```
<?xml version="1.0"?>
<eml:eml packageId="piscoUCSB.5.20" system="knb"
xmlns:eml="eml://ecoinformatics.org/eml-2.0.0">
<dataset>
  <shortName>Alegria Temperatures</shortName>
  <title>PISCO: Intertidal Temperature Data:
    Alegria, California: 1996-1997</title>
  <creator id="C.Blanchette">
    <individualName>
      <givenName>Carol</givenName>
      <surName>Blanchette</surName>
    </individualName>
    <organizationName>PISCO</organizationName>
    <address>
      <deliveryPoint>UCSB Marine Science
        Institute</deliveryPoint>
      <city>Santa Barbara</city>
      <administrativeArea>CA</administrativeArea>
      <postalCode>93106</postalCode>
    </address>
  </creator>
  <abstract>
    <para>These temperature data were collected
      at Alegria Beach, California, and were ...
    </para>
  </abstract>
  <keywordSet>
    <keyword>OceanographicSensorData</keyword>
    <keyword>Thermistor</keyword>
  </keywordSet>
  <keywordThesaurus>
    PISCOCategories
  </keywordThesaurus>
  </keywordSet>
  <intellectualRights><para>Please contact the
    authors for permission to use these data.
    Please also acknowledge the authors in any
    publications.</para>
  </intellectualRights>
  <contact>
    <references>C.Blanchette</references>
  </contact>
</dataset>
</eml:eml>
```

Transform



DATA CATALOG
Search

data catalog
login
search
insert

Metadata Identifier: piscoUCSB.5.20

Short Name: Alegria Temperatures

Title: PISCO: Intertidal Temperature Data: Alegria, California: 1996-1997

Individual: Dr. Carol Blanchette

Organization: PISCO

Address: UCSB Marine Science Institute,
Santa Barbara,
CA 93106

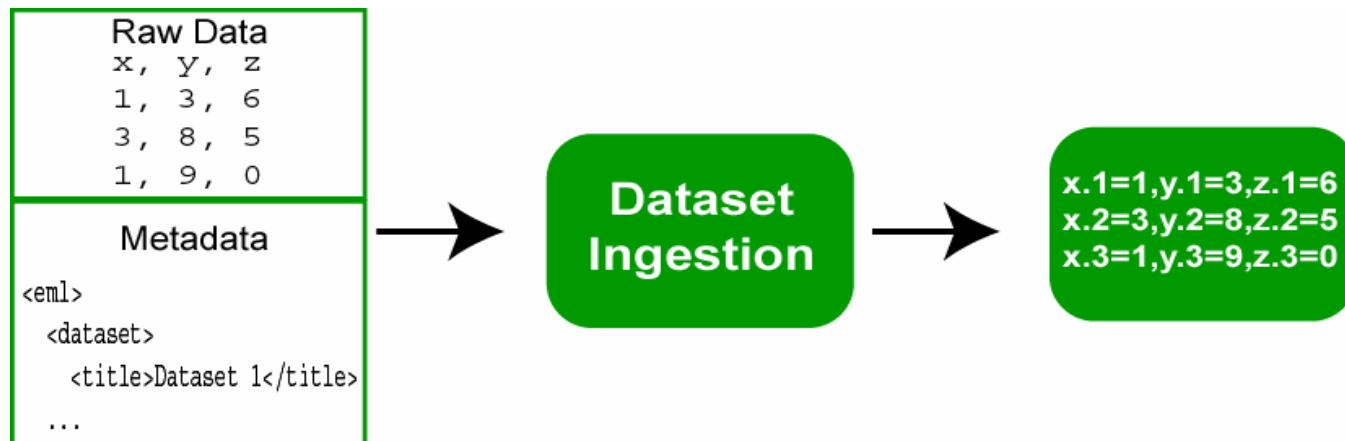
Abstract:

These temperature data were collected at Alegria Beach, California, and were part of an array of intertidal temperature sensors that extend from Piedras Blancas, on the California Central Coast, to Port Hueneme, on the California South Coast. Sensors are set to a frequency of 20 minutes per measurement, and the units are collected approximately once every couple of months (up to ~6 months). The data are downloaded and saved as BoxCar(tm). DTF files, and are then converted to ASCII Comma Separated Values files for archive purposes. Please see the methods section for detailed processing descriptions.

Keywords:

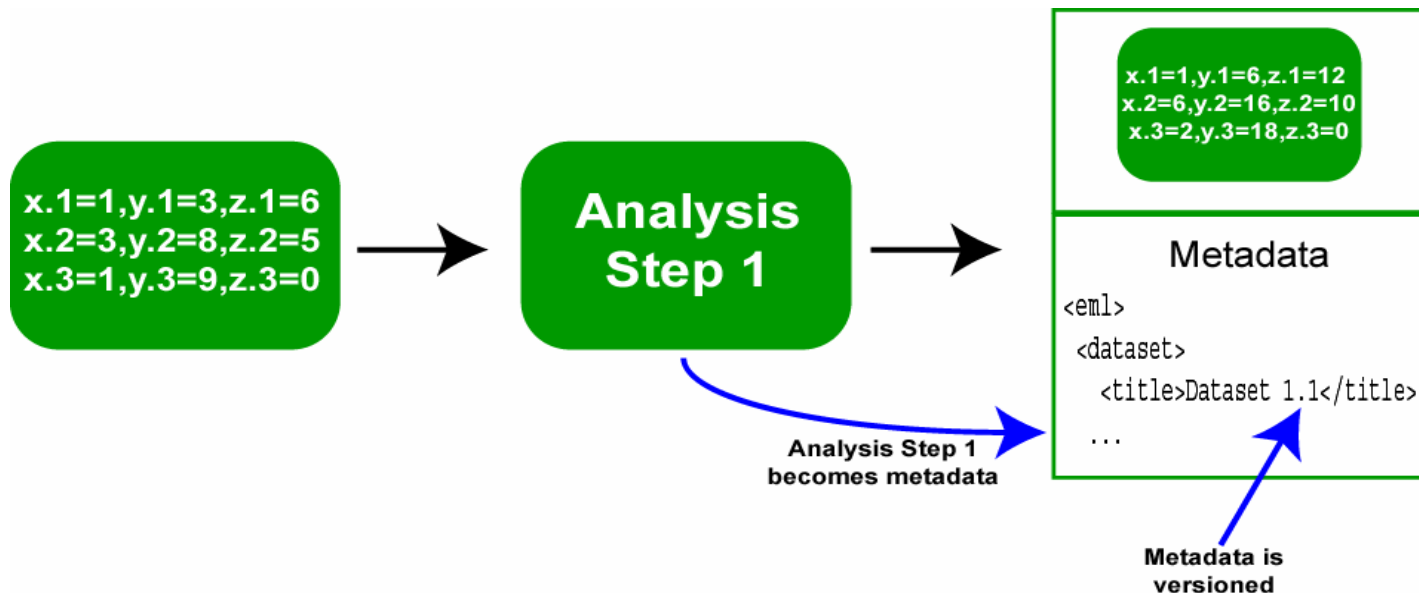
- intertidal
- temperature

- Key information needed to read and machine process a data file is in the metadata
 - File descriptors (CSV, Excel, RDBMS, etc.)
 - Entity (table) and Attribute (column) descriptions
 - Name
 - Type (integer, float, string, etc.)
 - Codes (missing values, nulls, etc.)
 - Integrity constraints
 - In the future, this will include semantic typing

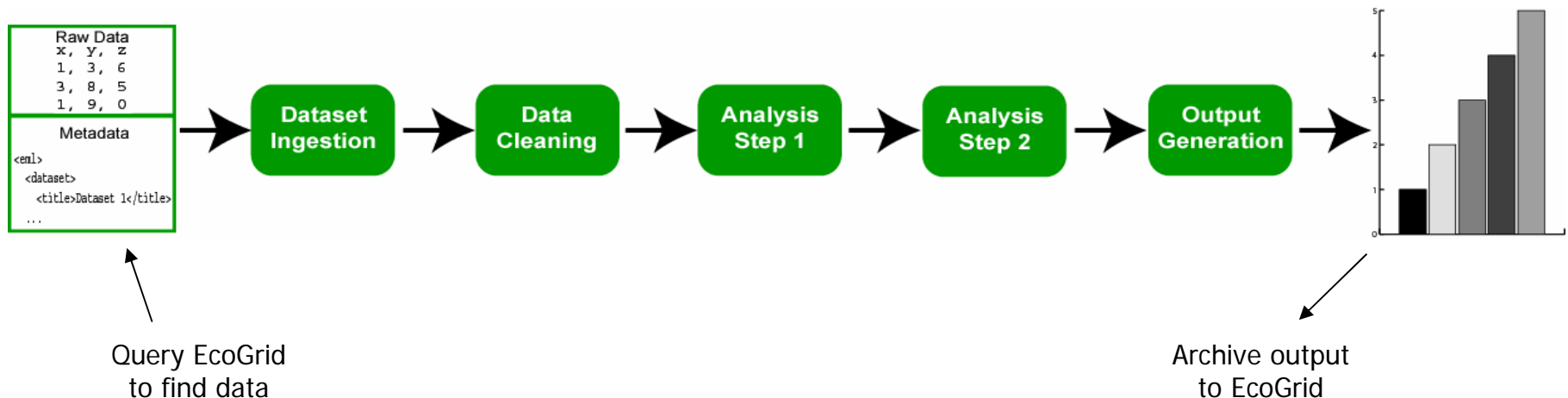


Metadata revision

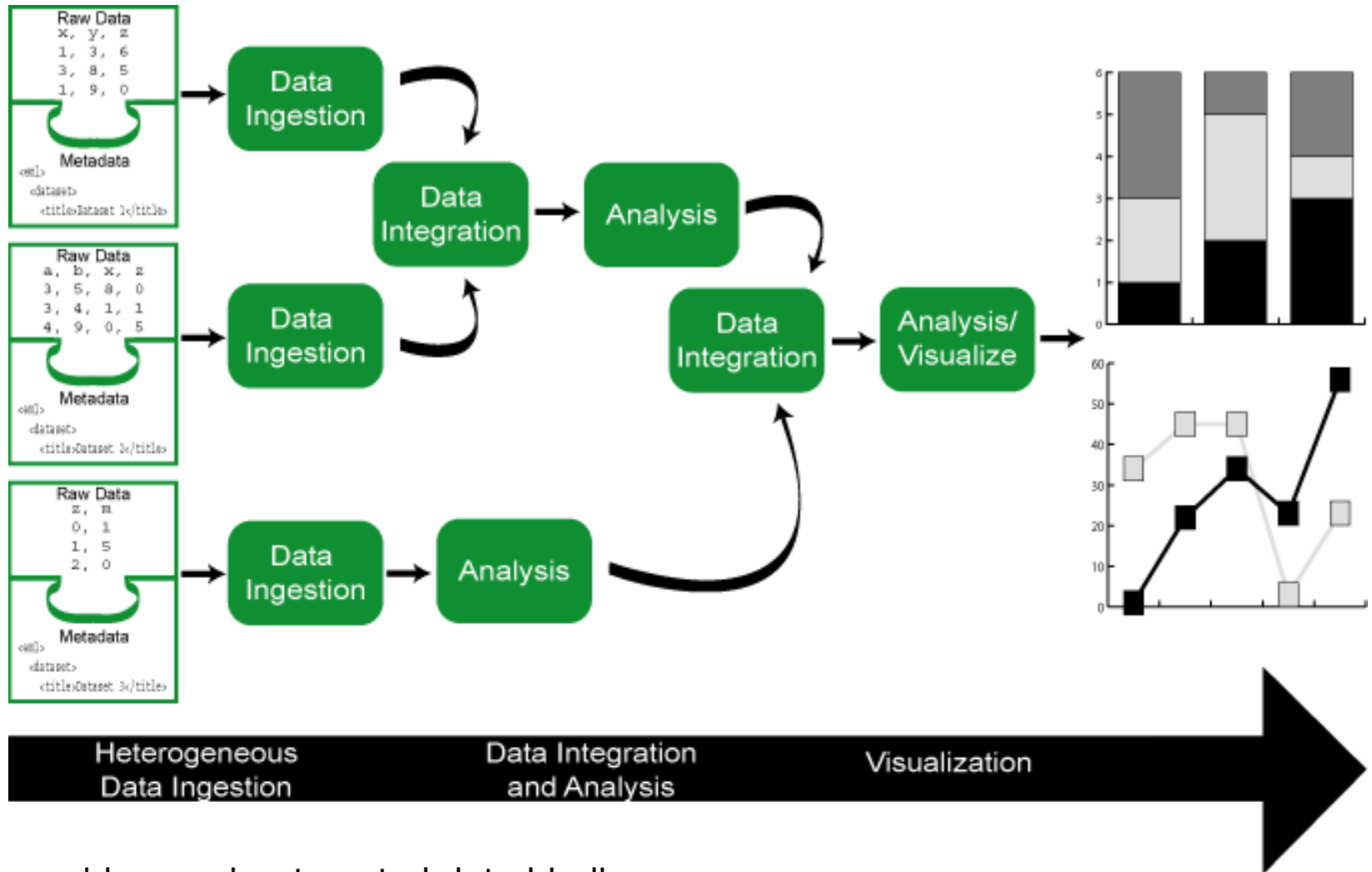
- Metadata needs to be revised following any transformation
- Versioning of metadata and data is important to reuse/repeatability
- The process describes the data lineage as it has been transformed
- Derived data sets can be stored in EcoGrid with provenance



- In the SEEK model, data ingestion/cleaning can be metadata driven (specifically with EML)
- Output generation includes creating appropriate metadata
- The analysis pipeline itself becomes metadata



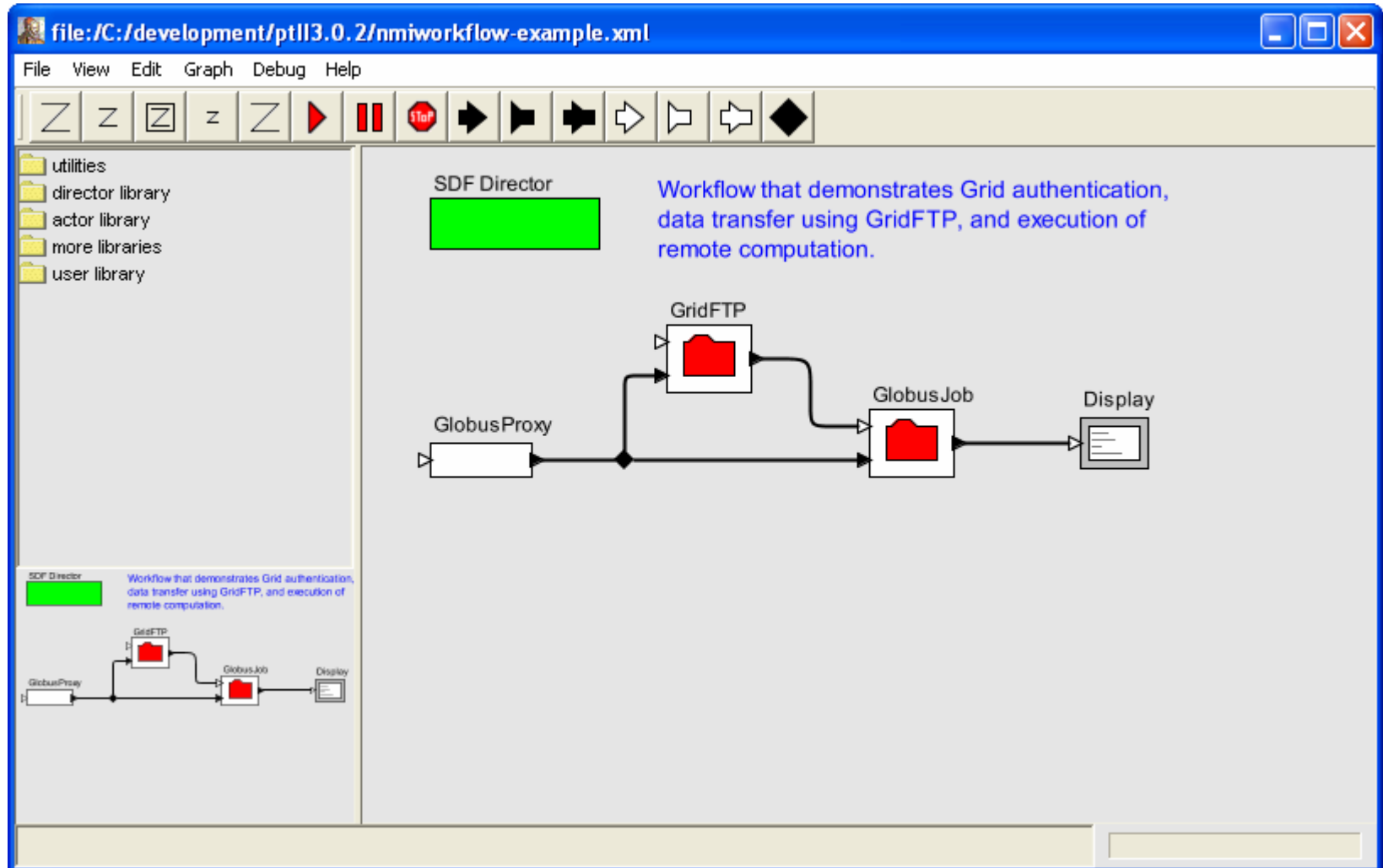
Kepler: scientific workflows

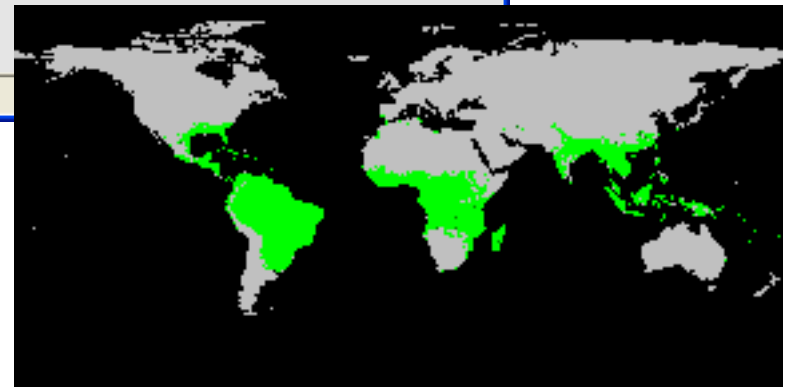
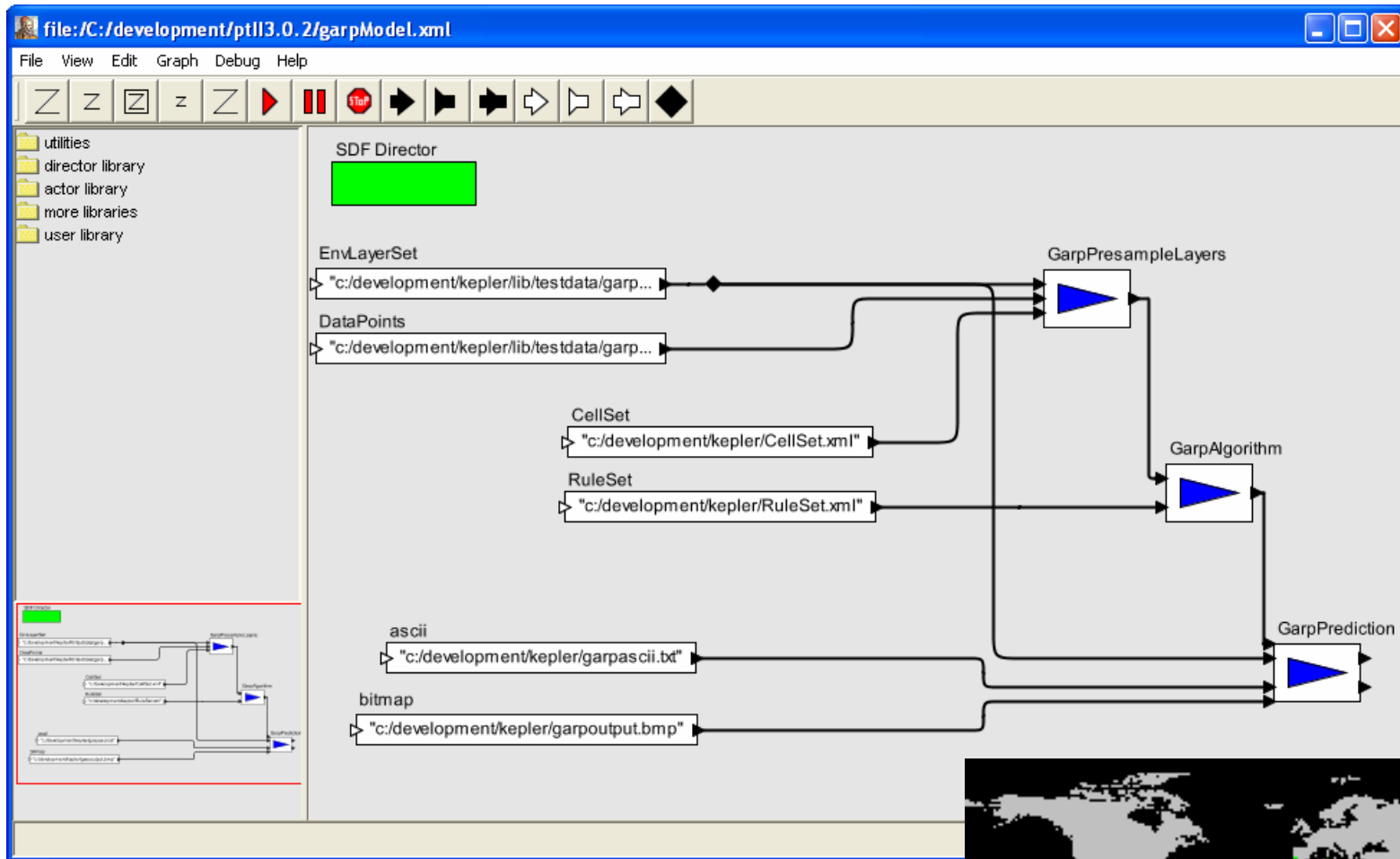


EML provides semi-automated data binding

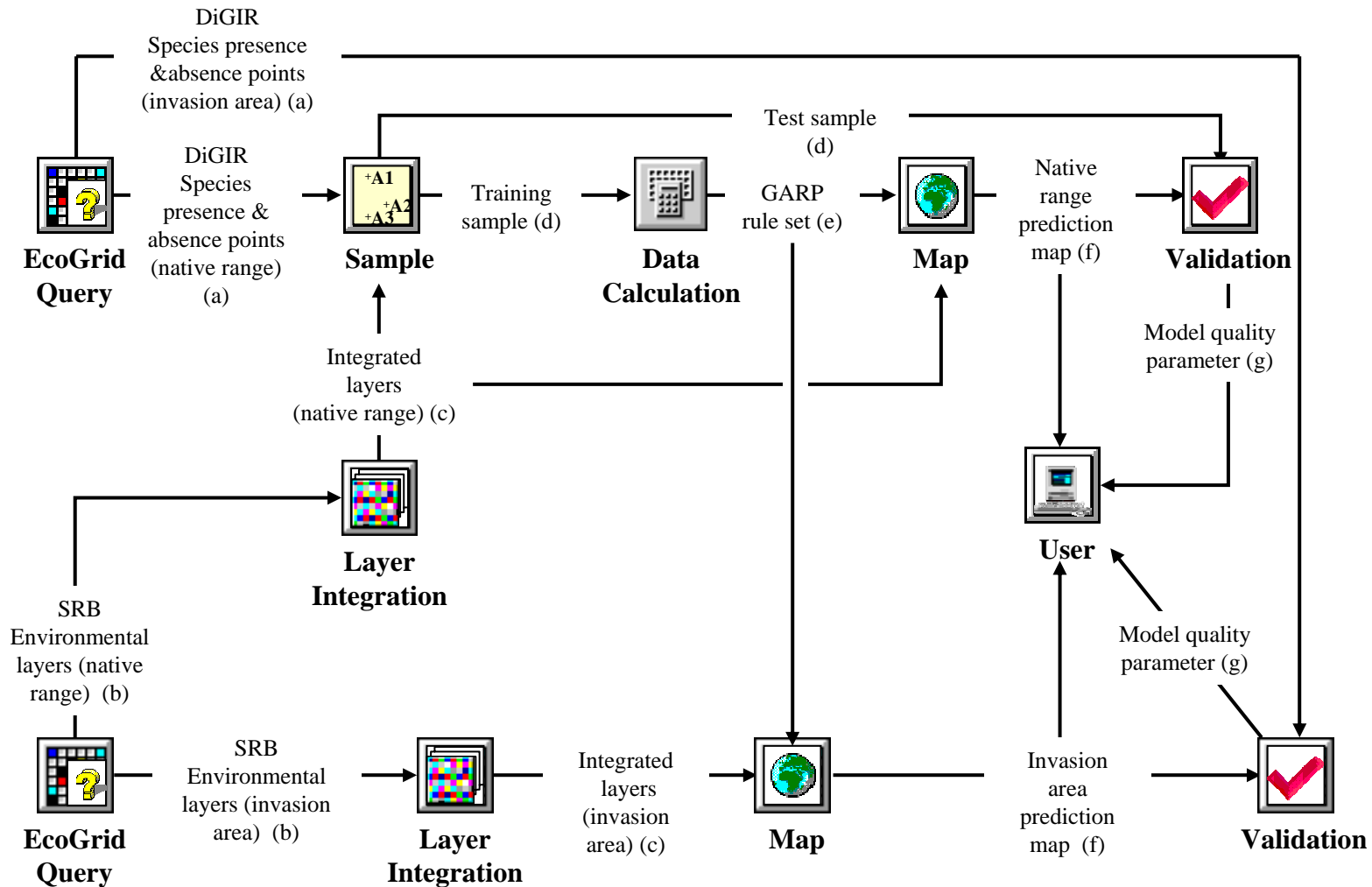
Scientific workflows represent knowledge about the process; Kepler captures this knowledge

Kepler: grid services access



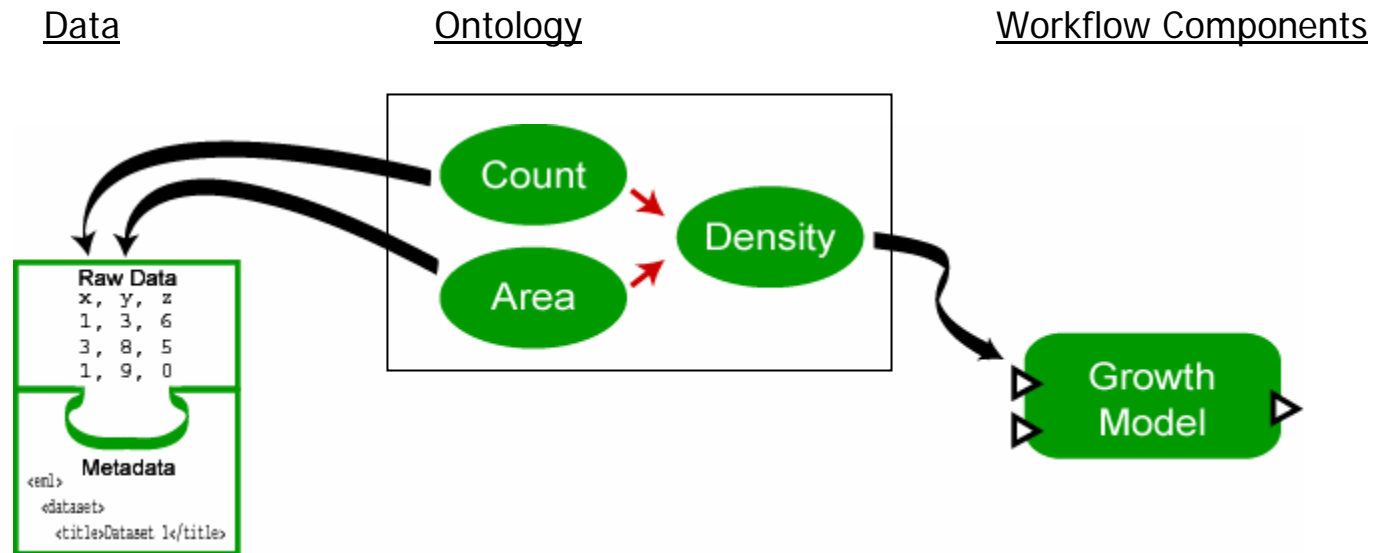


GARP Invasive Species Model



Scientific workflows represent knowledge about the process; AMS captures this knowledge

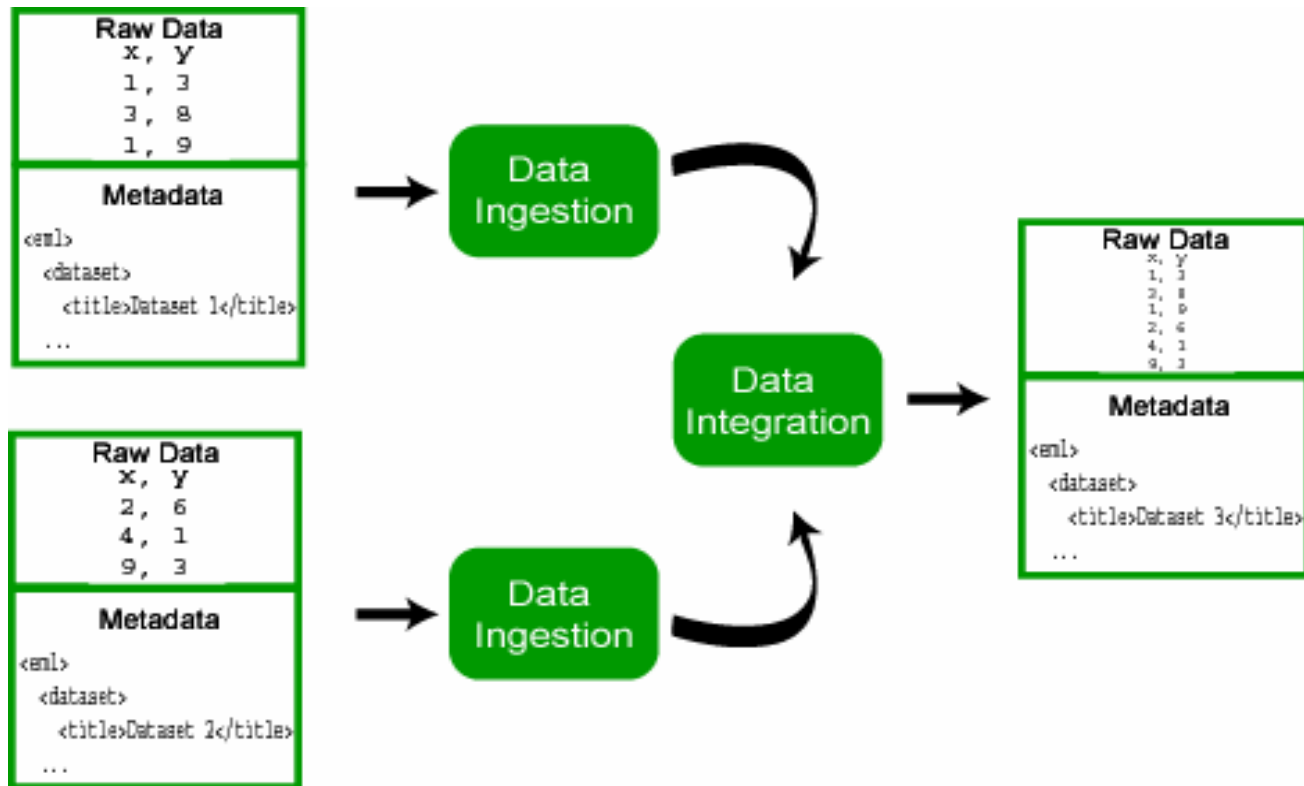
- Label data with semantic types
- Label inputs and outputs of analytical components with semantic types



- Use reasoning engines to generate transformation steps
 - Beware analytical constraints
- Use reasoning engine to discover relevant components

Homogeneous data integration

- Integration of homogeneous or mostly homogeneous data via EML metadata is relatively straightforward





- The diagram illustrates the process of integrating two datasets into a single product table. It shows two input tables, 'Study A' and 'Study B', which are combined into an 'INTEGRATED DATA PRODUCT' table. Green arrows indicate the flow of data from the input tables to the integrated product table.

METADATA (from EML)

Study A = White Mountains
 PIRU=Picea rubens
 BEPA=Betula papyifera
 Area column units = square meter

Study B = Green Mountains
 picrub=Picea rubens
 betpap=Betula papyifera
 Area sampled = 1 square meter

DATA

Date	Site	Species	Area	Count
10/1/1993	N654	PIRU	2	26
10/3/1994	N654	PIRU	2	29
10/1/1993	N654	BEPA	1	3

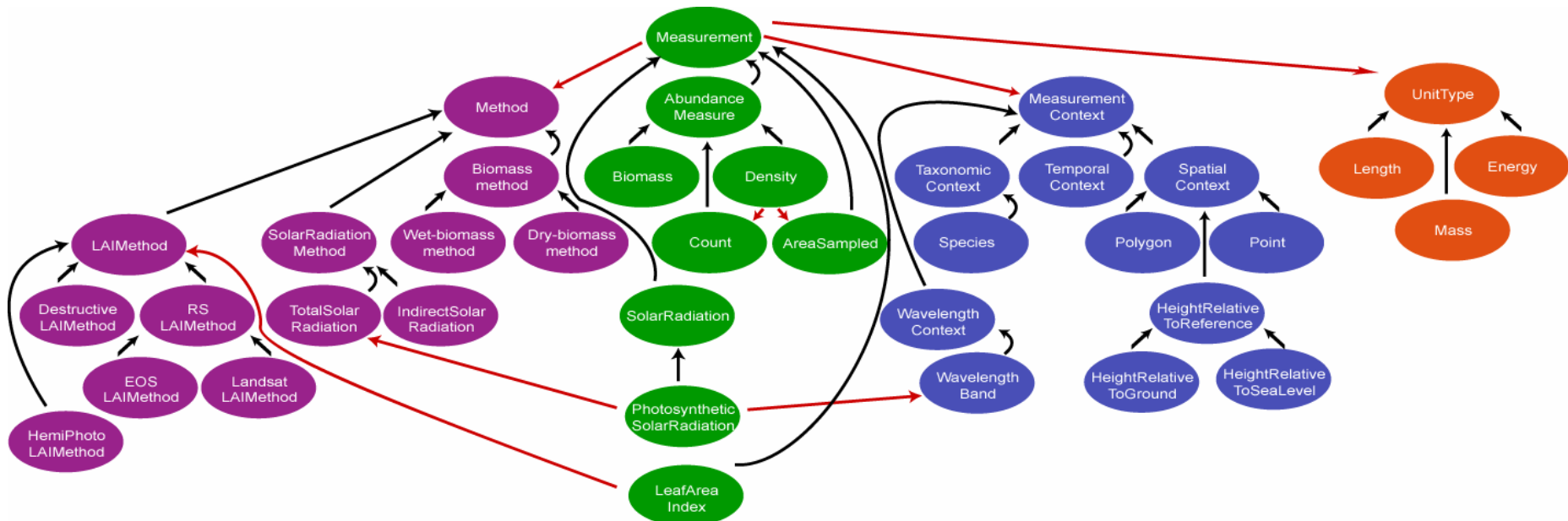
Date	Site	picrub	betpap
31Oct1993	1	13.5	1.6
14Nov1994	1	8.4	1.8

INTEGRATED DATA PRODUCT

Study	Date	Site	Species	Density
A	10/1/1993	N654	Picea rubens	13
A	10/3/1994	N654	Picea rubens	14.5
A	10/1/1993	N654	Betula papyifera	3
B	10/31/1993	1	Picea rubens	13.5
B	10/31/1993	1	Betula papyifera	1.6
B	11/14/1994	1	Picea rubens	8.4
B	11/14/1994	1	Betula papyifera	1.8

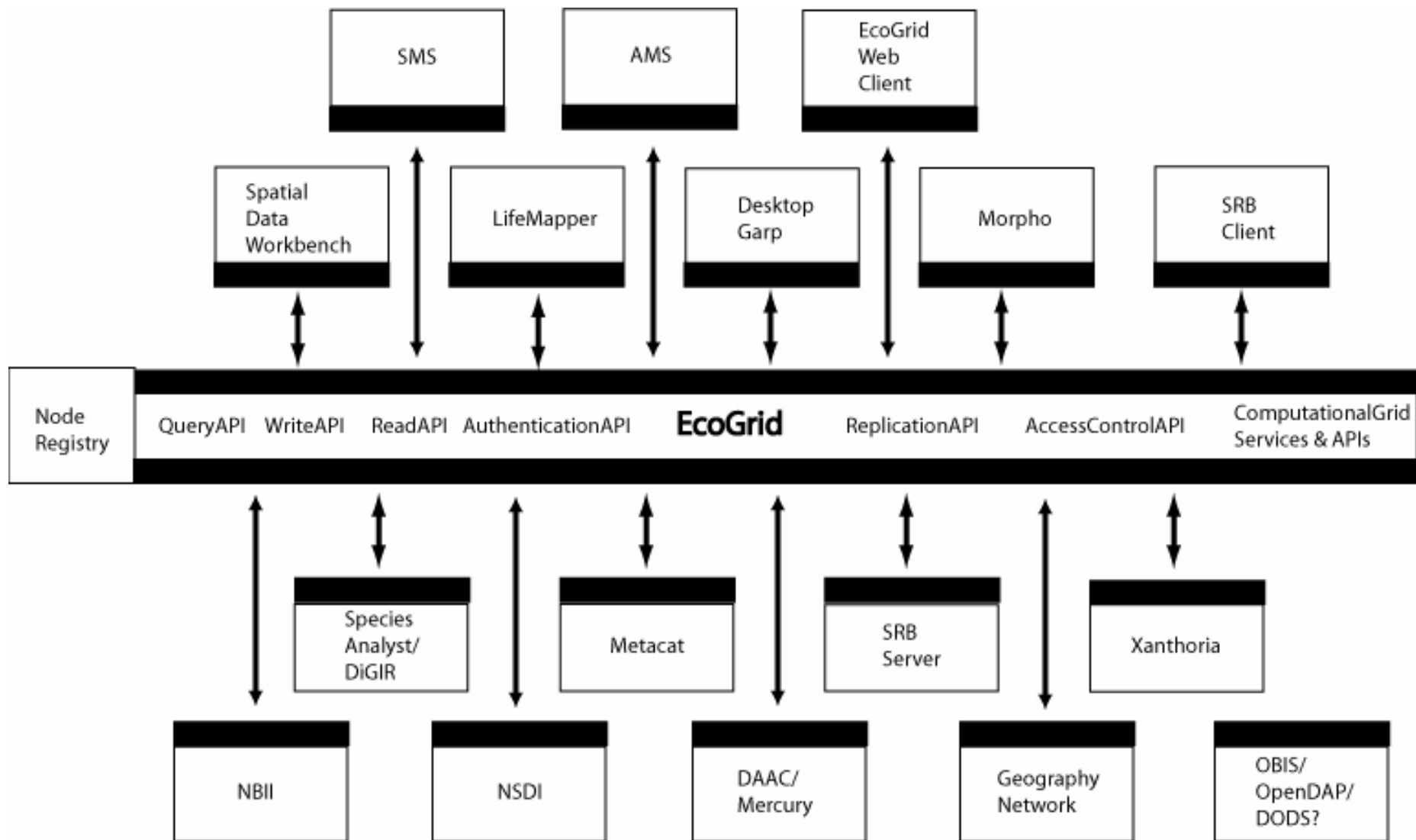
Ecological ontologies

- What was measured (e.g., biomass)
 - Type of measurement (e.g., Energy)
 - Context of measurement (e.g., *Psychotria limonensis*)
 - How it was measured (e.g., dry weight)
- SEEK intends to enable community-created ecological ontologies using OWL
 - Represents a controlled vocabulary for ecological metadata

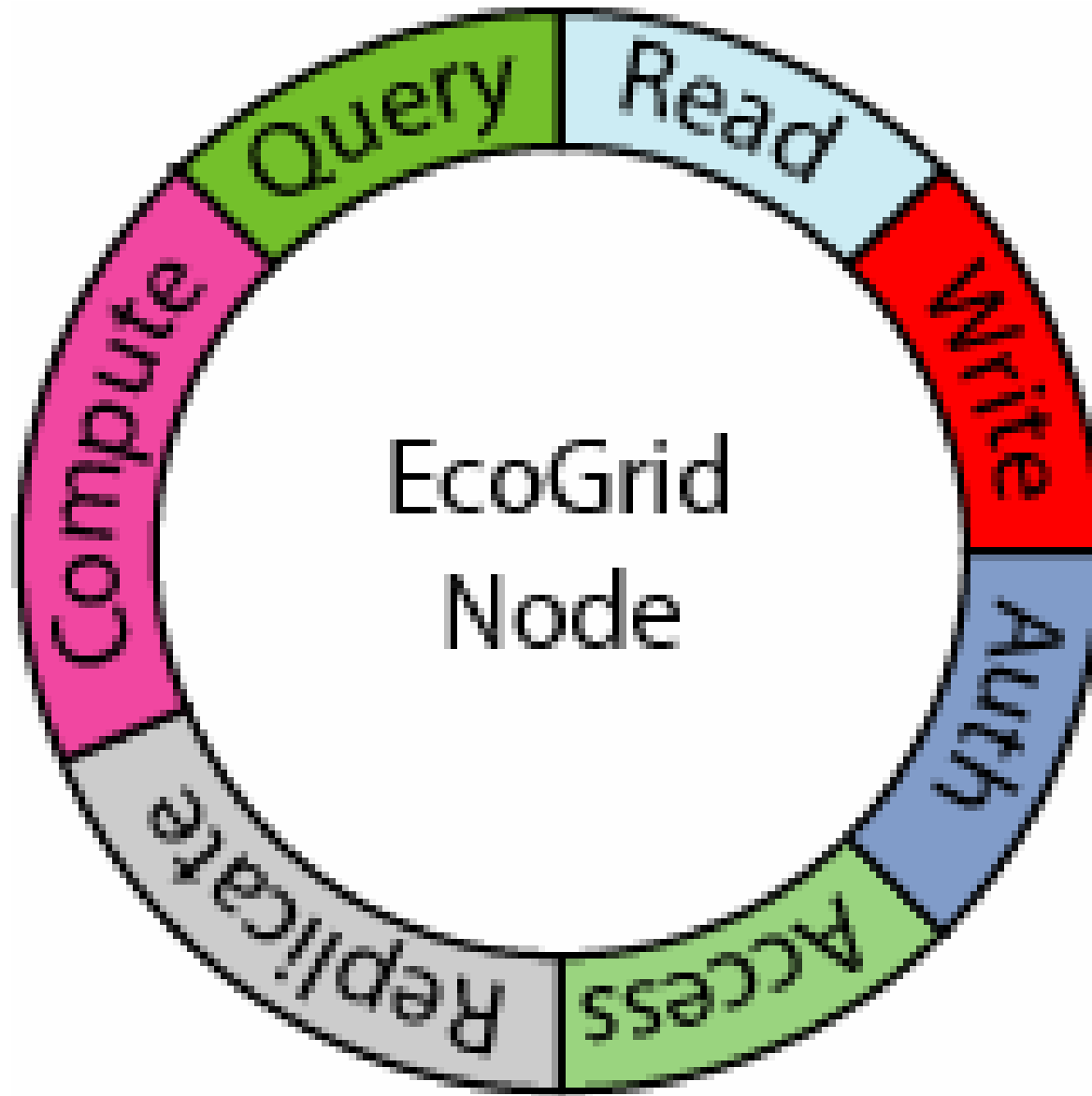


- More about this in Bertram's talk

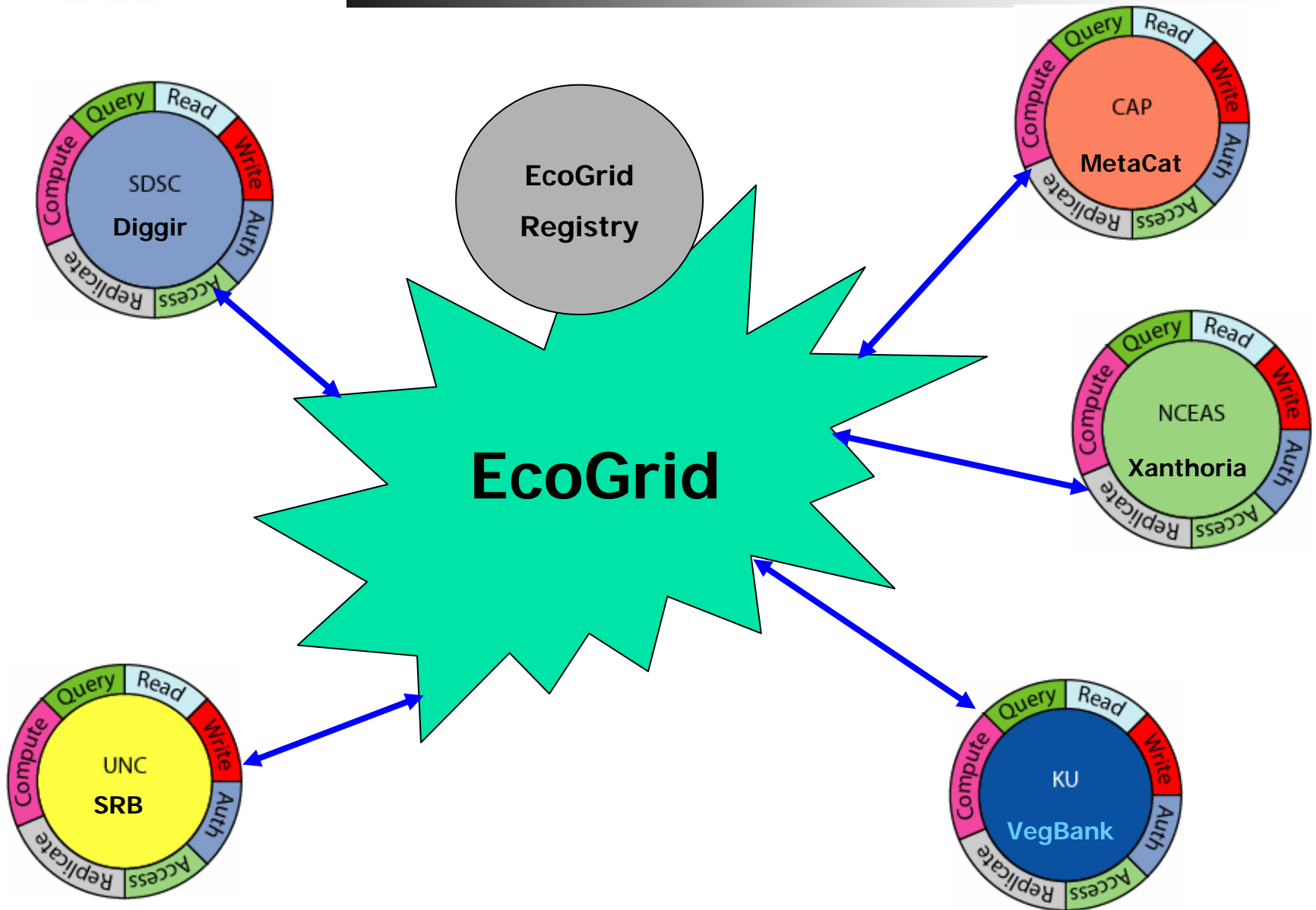
Layers in EcoGrid



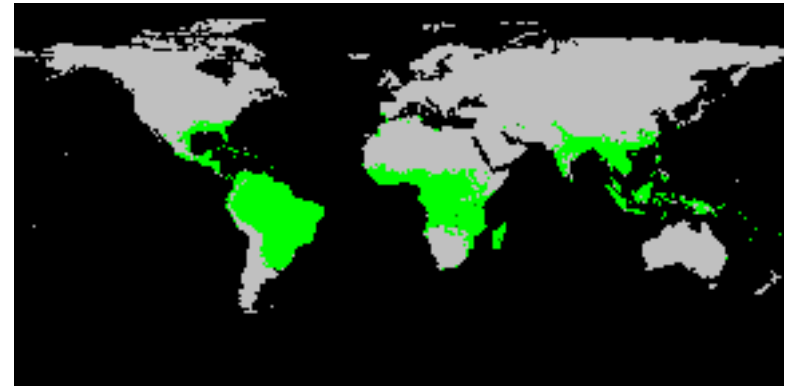
EcoGrid Node



EcoGrid Resources



- Read, Query & Register Completed
- Simple Registry Operational
- EcoGrid Wrappers completed for:
 - MetaCat
 - SRB
 - DiGGiR
 - Xanthoria
- Available Interfaces
 - WSDL
 - Simple Web Interactivity
 - Kepler



Acknowledgements

This material is based upon work supported by:

The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

PBI Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research)



Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON

