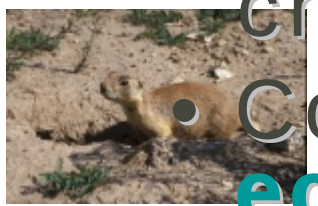
The Science Environment for Ecological Knowledge (SEEK): A Distributed Environment for Ecological Modeling and Analysis

Deana Pennington
University of New Mexico
Long Term Ecological Research Network Office

SEEK New Faculty and Postdoc Training Workshop
January 9, 2006

Grand Challenges in Ecology

- Alterations in **biodiversity**…exotic species, infectious disease
- Altered **biogeochemical cycles** at multiple spatial scales
- **Climate change** and variability, including **ecosystem response** to change
- Coupled **human-natural ecosystems**



Ecoinformatics

Understanding the environment around us will require the use of all of the information available to us

- Field data
- New data collection techniques: in situ sensor arrays
- Remotely-sensed imagery
- Large and/or widely distributed data sets
- Data heterogeneity (text, Excel, GIS, DB, image, etc.)
- Scaling issues: space, time, levels

Informatics Challenges

- Data are heterogeneous
  - Physical
  - Structural/logical
  - Semantics
    - From many disciplines
      - Biodiversity surveys, hydrology, atmospheric chemistry, spatial data, behavioral experiments,…
      - Data on economics, demographics, legal issues,…
- Data are distributed
- Data are undocumented/poorly documented
- Cultural barriers to data sharing

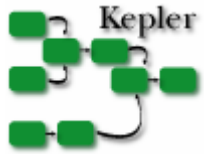
SEEK

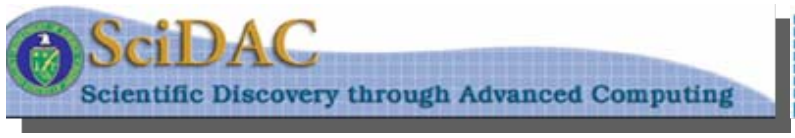
- SEEK is a research project designed to address the challenges of data integration in ecology
  - NSF CISE (Computer and Information Science and Engineering) Information Technology Research (ITR)
  - 5 years (currently in year 3)
  - Proposed by the Partnership for Biodiversity Informatics (PBI)
  - IT research: data integration
  - Broad-scale biodiversity applications: ecological niche modeling, biodiversity

Collaborators

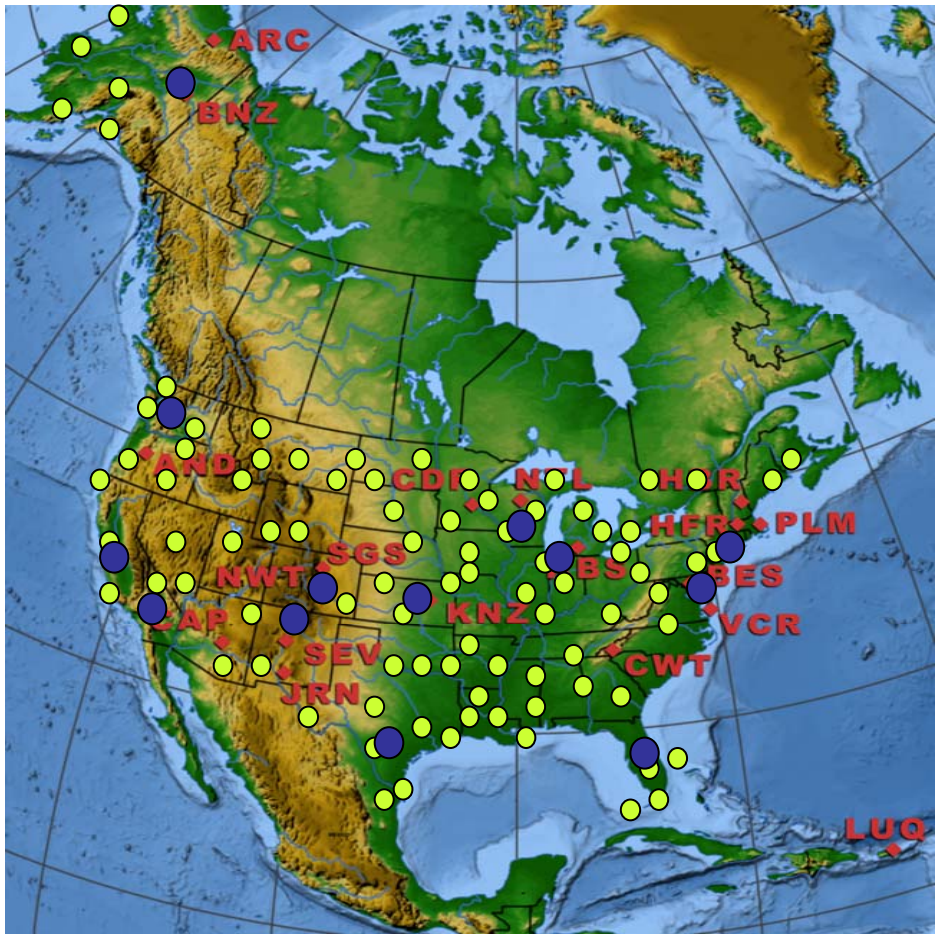
Partnership for Biodiversity Informatics (PBI)

Knowledge Network
For
Biocomplexity





Solutions



- Ecological Metadata Language (EML)

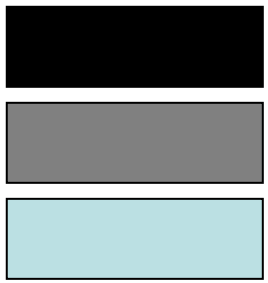
  PBI/KNB

- Morpho
  - metadata and data management software
- Metacat
  - distributed data system
  - registries: KNB, UCNRS, OBFS, NCEAS, PISCO, LTER
- EcoGrid

  PBI/SEEK
  - integrating distinct data systems and networks
- Kepler
  - grid-enabled scientific

What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

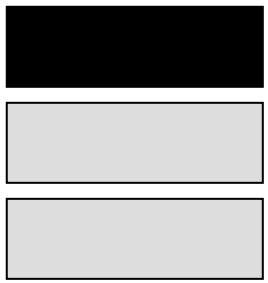
EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

What is SEEK?

System development:



Kepler analysis & modeling system

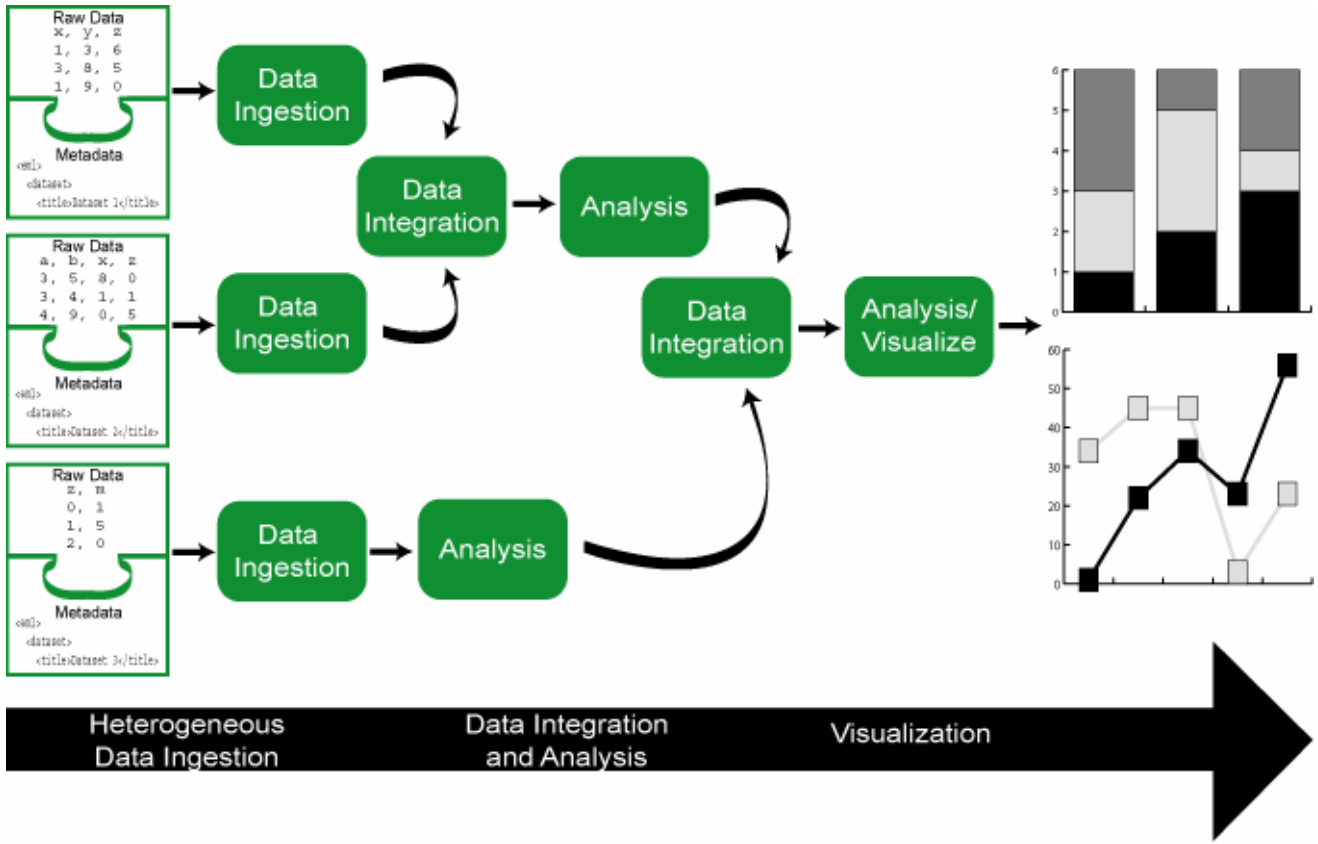
Semantic mediation system (glue)

EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

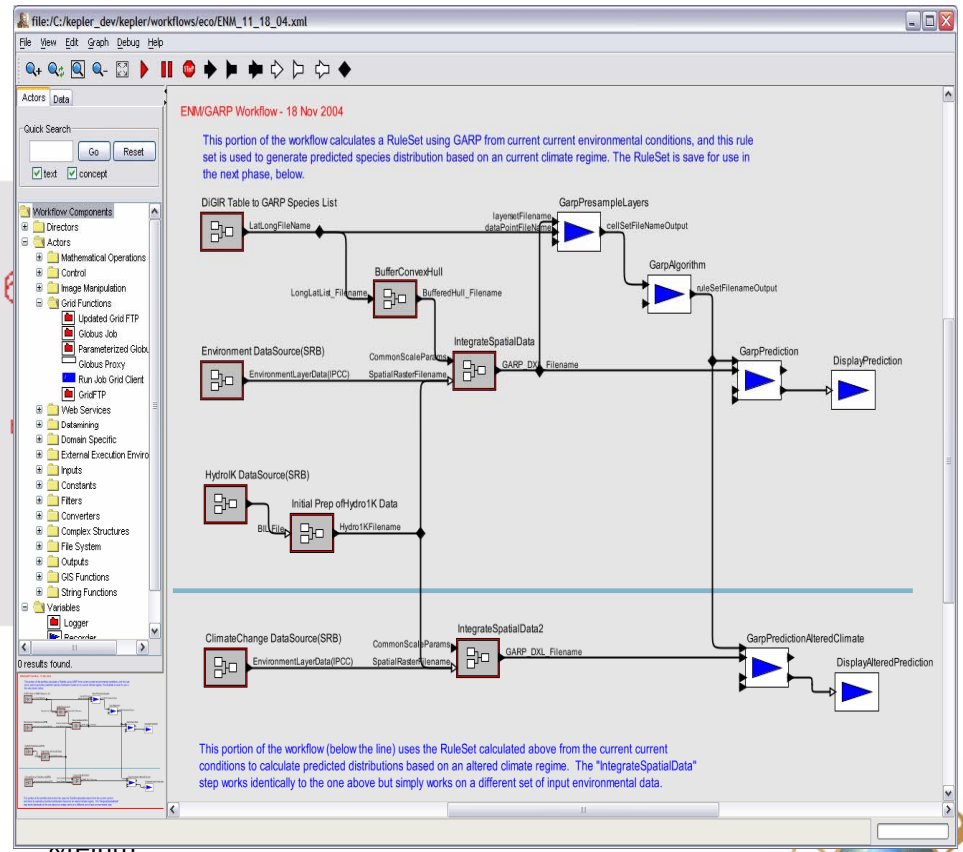
Scientific Workflows



Kepler Workflow System

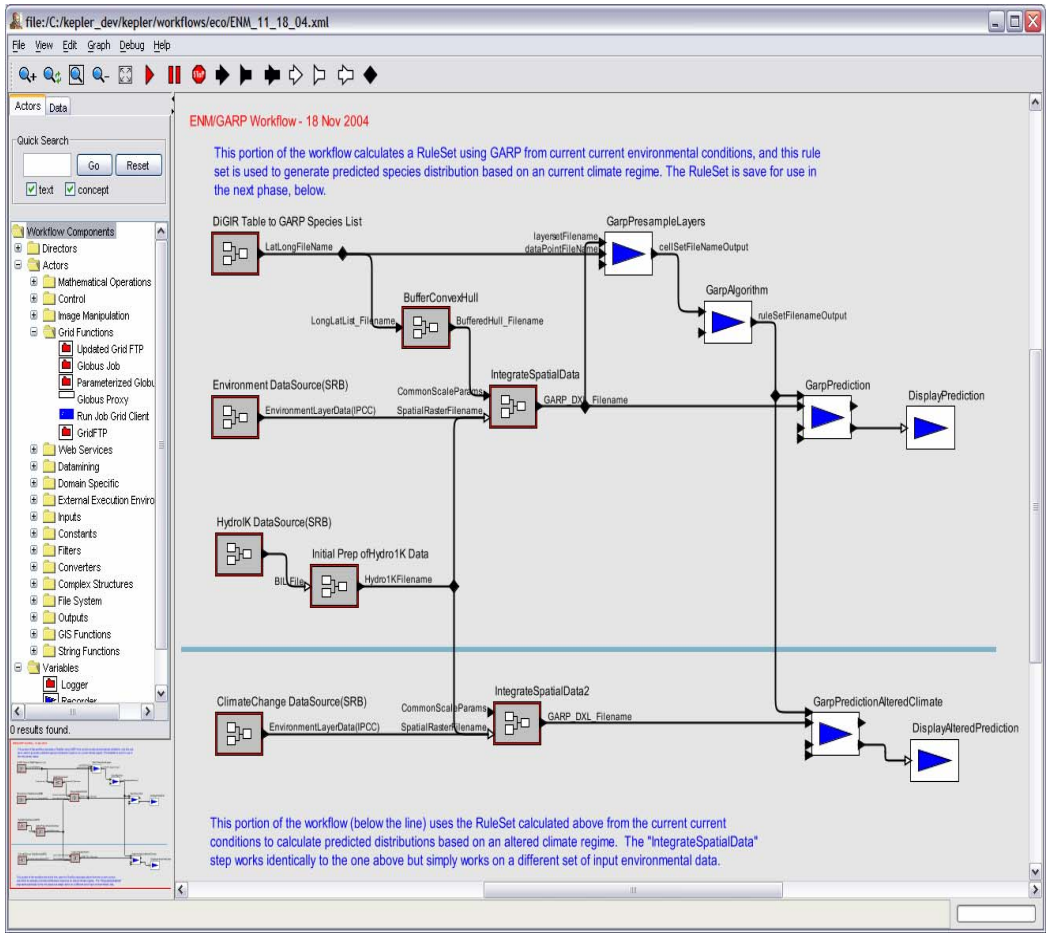
- **Scripting**
  - Single environment
  - Single platform
  - Automated procedures
- **Visual modeling (Stella)**
  - Drag & drop icons
  - Linking arrows
- **Scientific Workflows**
  - Cross-platform
  - Cross-environment
  - Distributed data & analyses

/**********Set Variables*********/



&return

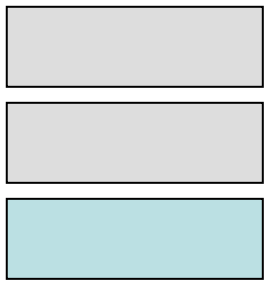
The Kepler Scientific Workflow System



- Based on Ptolemy II Visual System
  - Developed by electrical engineering community as a visual *dataflow* programming application
  - Hierarchical, *nested* workflows (actors) and
  - explicit *computation models*, e.g., continuous time, discrete event, etc.
- Kepler adds
  - Generic & domain-specific actor libraries (R, niche-modeling, phylogenetics, etc.)
  - EML-based metadata tools
  - EcoGrid access/query
  - Distributed Execution
  - Web-service support
  - Ontology and integration support

What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

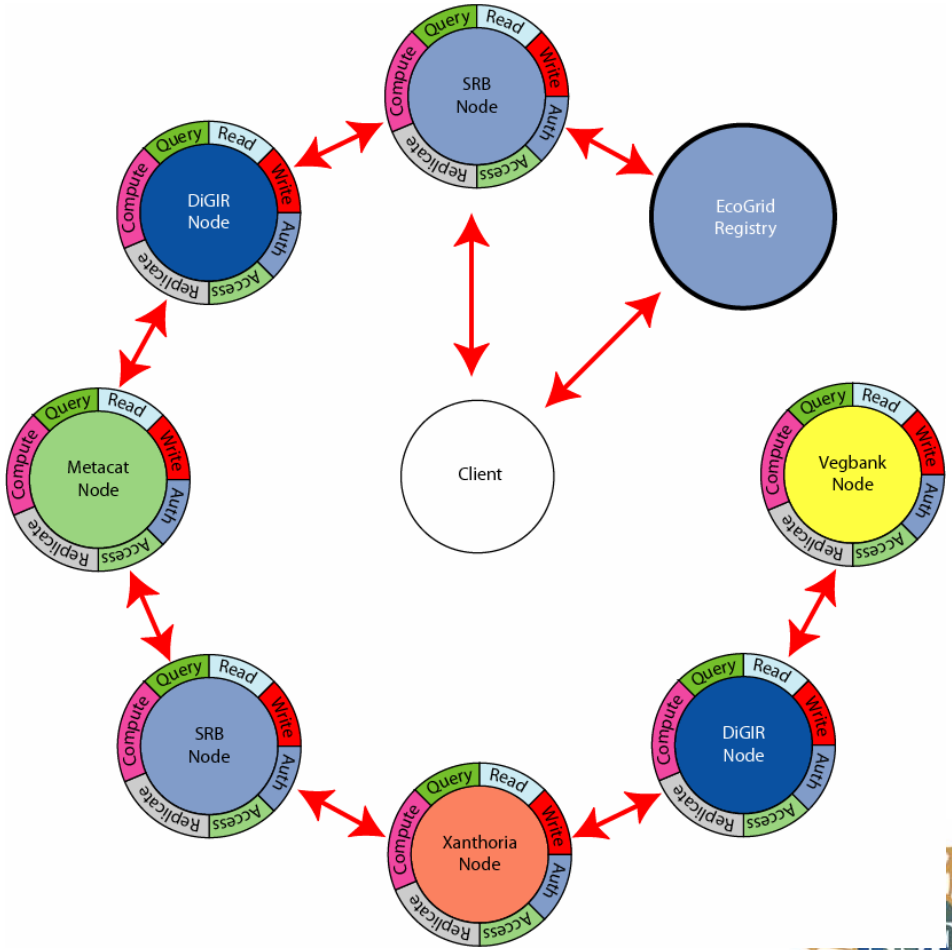
EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

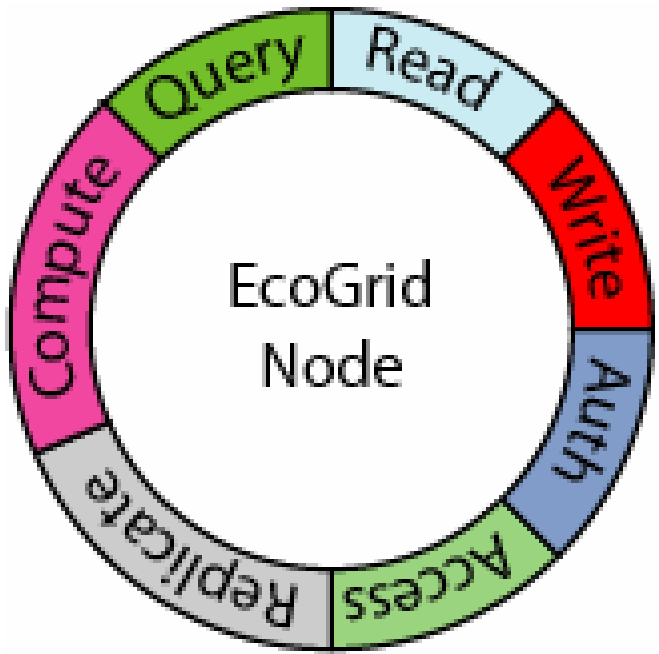
Grid: Data and Compute

- Modes of interaction
  - Client-server
  - Fully distributed
  - Peer-to-peer
- EcoGrid Registry
  - Node discovery
  - Service discovery
- Aggregation services
  - Centralized access
  - Reliability
  - Data preservation



SEEK EcoGrid

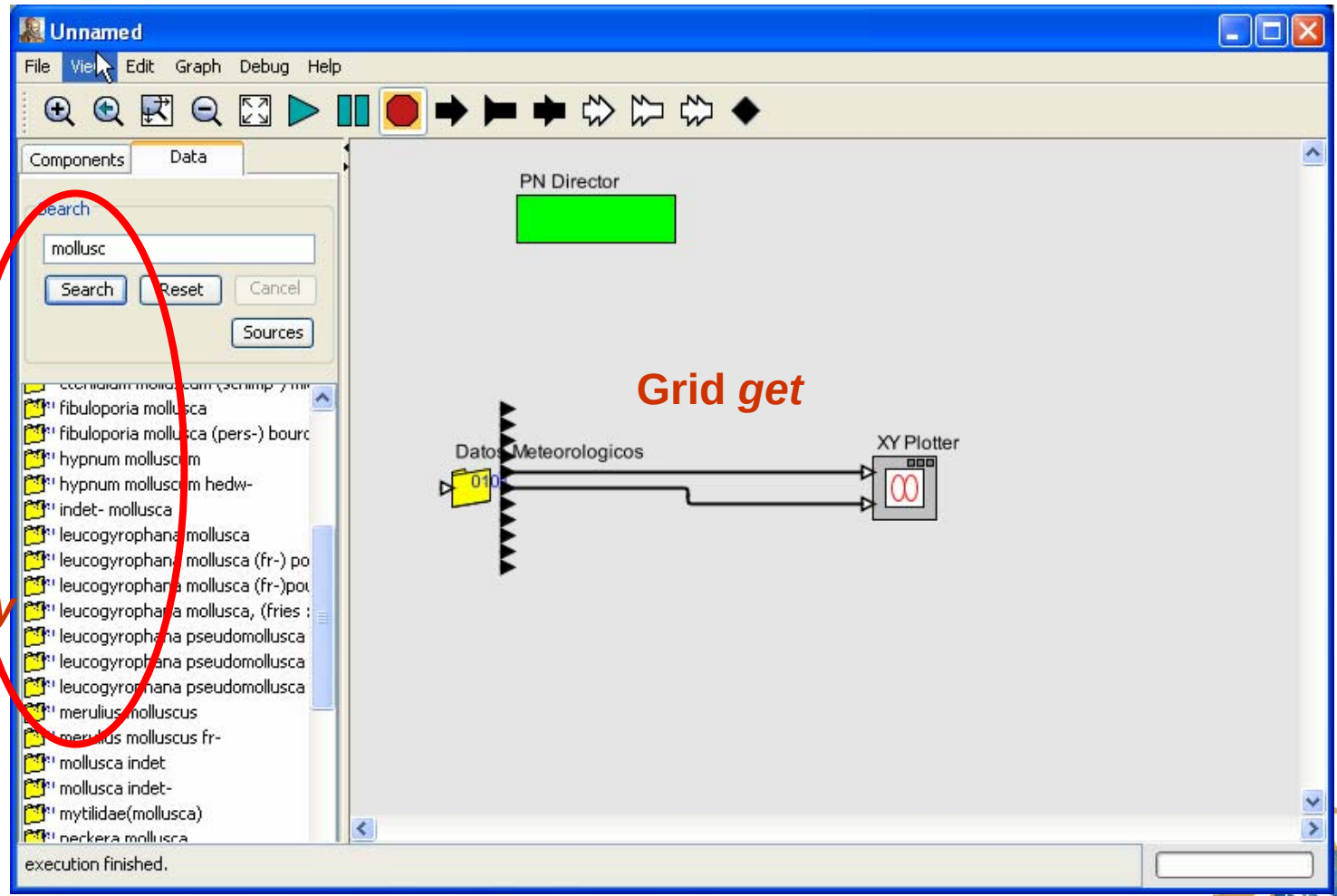
- Allow diverse environmental data systems to interoperate
  - *Integrate diverse data* networks from ecology, biodiversity, and environmental sciences
  - *Hide complexity* of underlying systems using lightweight interfaces



- Data systems
  - Systems contribute by implementing EcoGrid Grid-service interfaces
  - Prototypes exist for: Metacat, SRB, and DiGIR

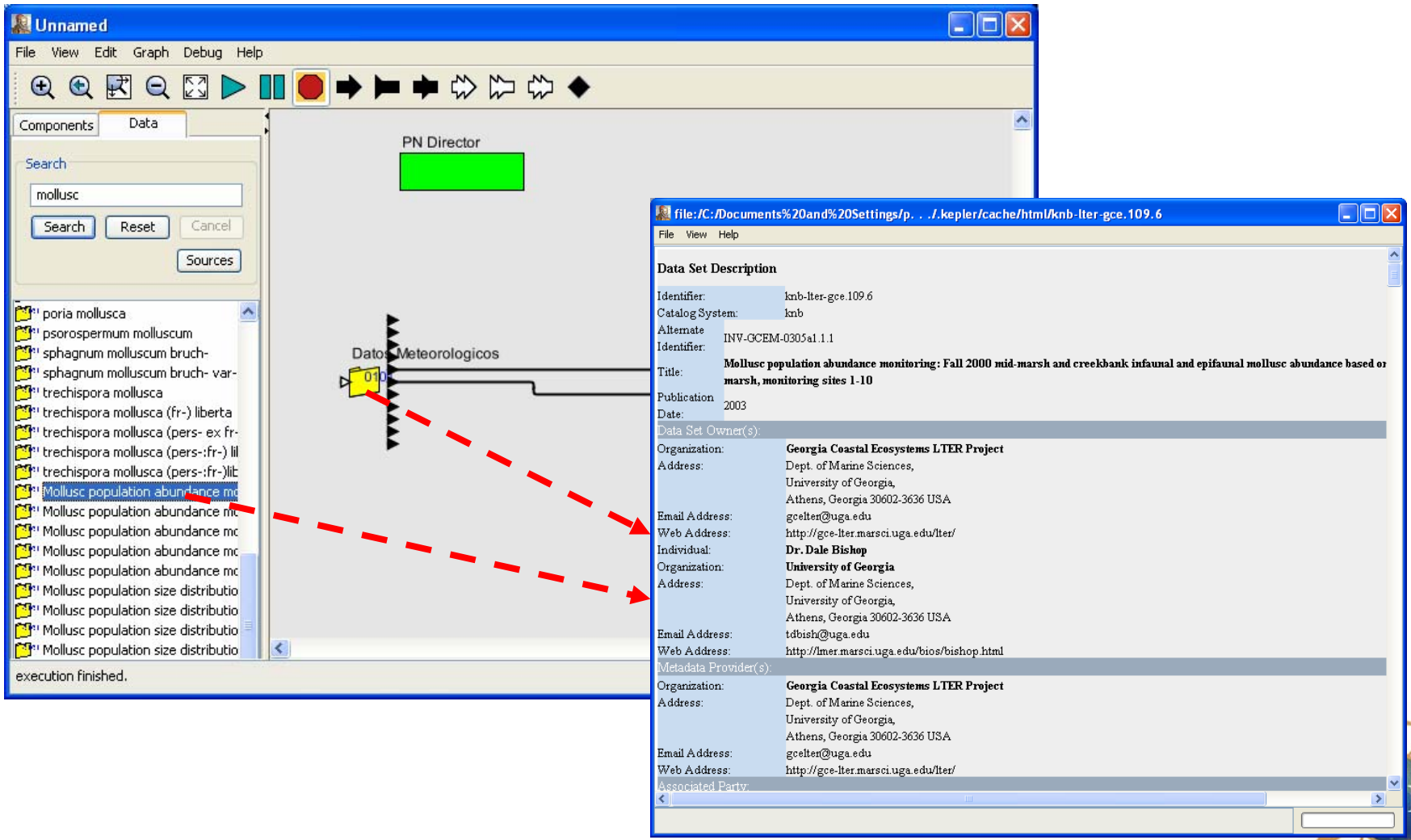
- Supports multiple XML-based metadata standards
  - EML & Darwin Core as foci

EcoGrid registered data show up in KEPLER

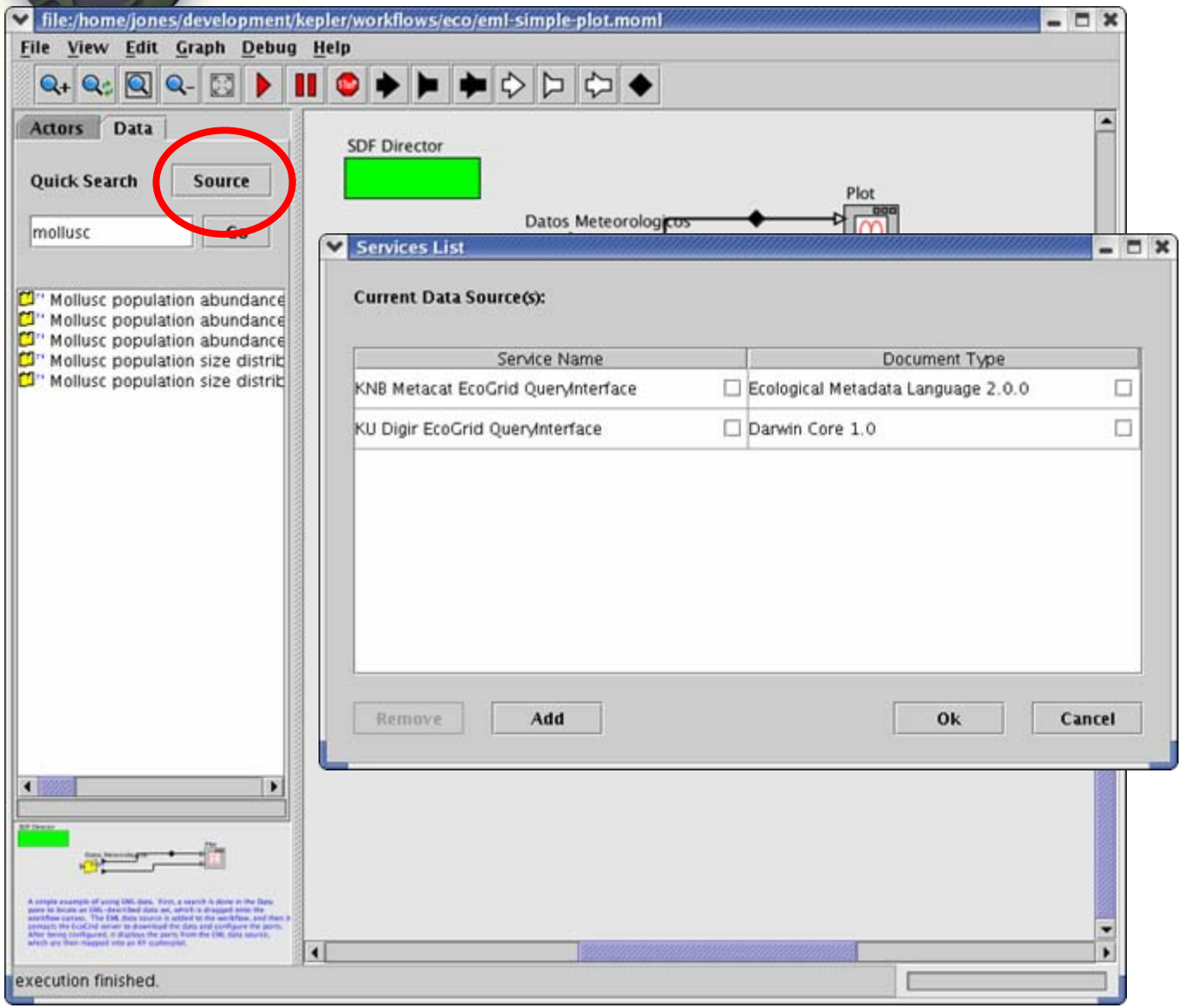


Grid *query*

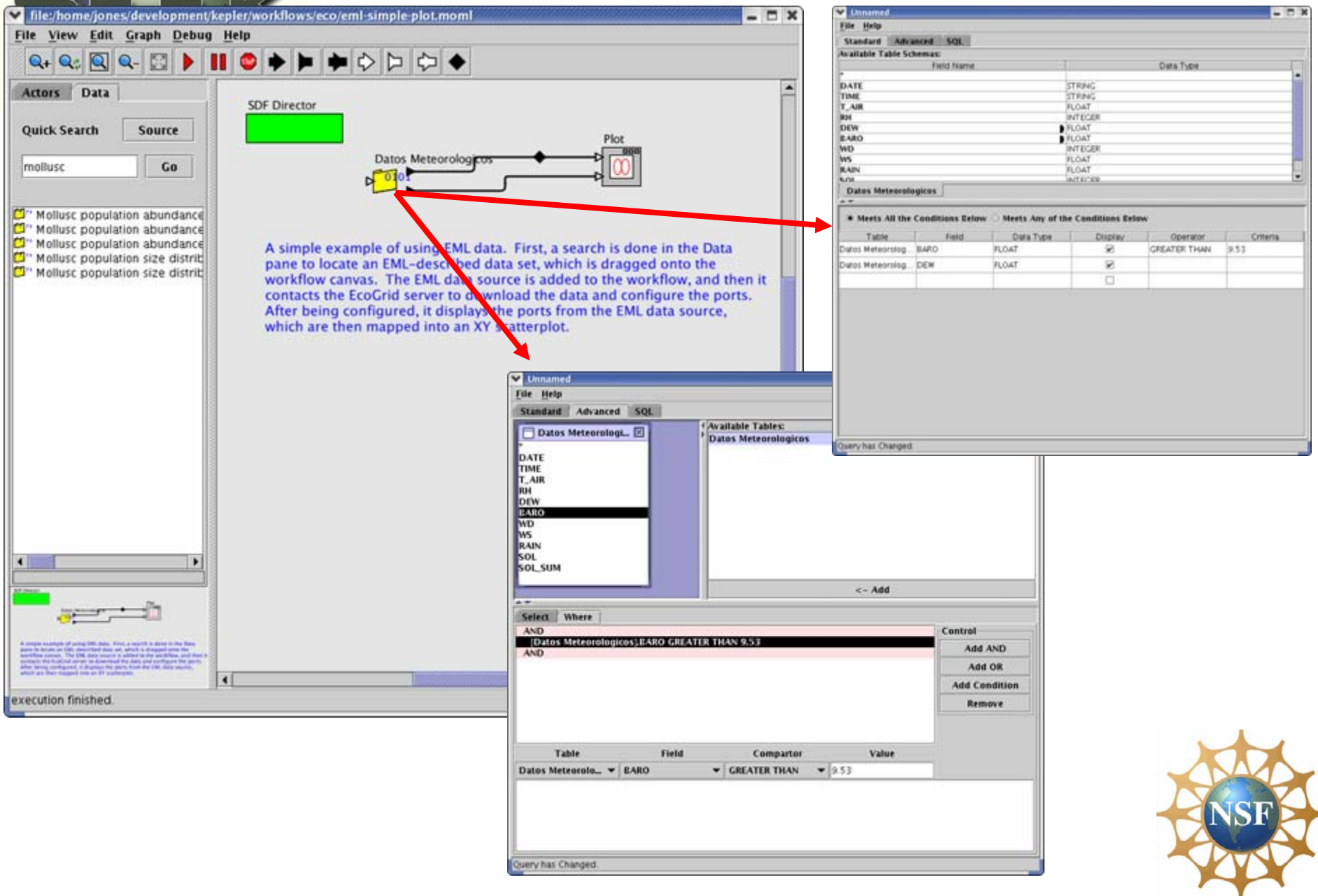
Metadata Display



Sources


file:/home/jones/development/kepler/workflows/eco/eml-simple-plot.moml
File View Edit Graph Debug Help
Actors Data
Quick Search
Source
mollusc
Mollusc population abundance
Mollusc population abundance
Mollusc population abundance
Mollusc population size distrib
Mollusc population size distrib
SDF Director
Plot
Datos Meteorologicos
Services List
Current Data Source(s):
Service Name
Document Type
KNB Metacat EcoGrid QueryInterface
Ecological Metadata Language 2.0.0
KU Digir EcoGrid QueryInterface
Darwin Core 1.0
Remove
Add
Ok
Cancel
execution finished.

Query Builder



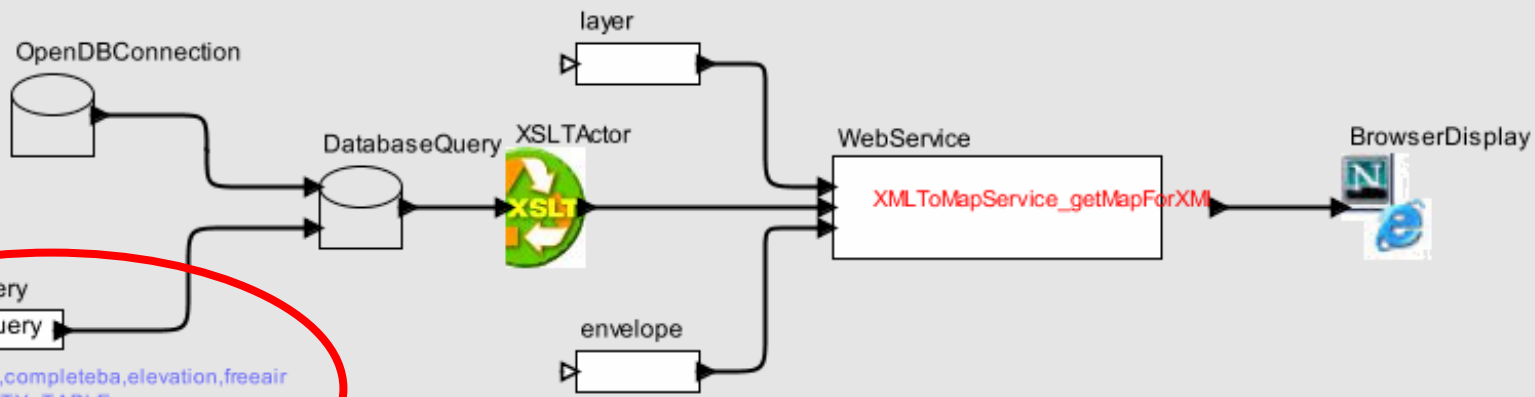
More WF Plumbing

Generating datasets on the fly.

- query: "select latdd,londd,completeba,elevation,freeair from GRAVITY_TABLE where latdd between " + latMin + " and
- latMin: "34.9"
- latMax: "35"
- longMin: "-120"
- longMax: "-119"

SDF Director

This workflow is used to extract gravity lat long point from an oracle database and generate shapefiles using a web service.



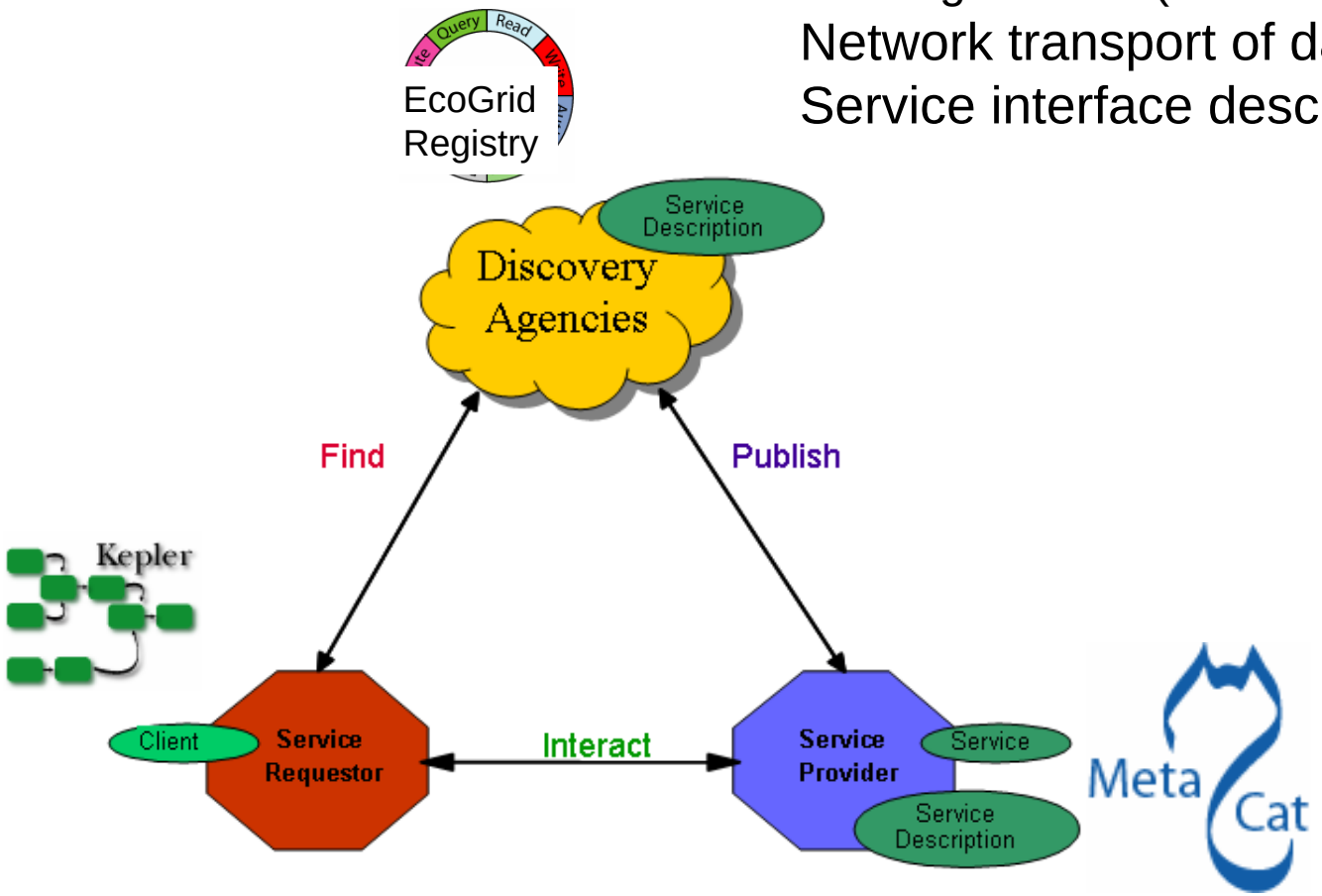
Web Service

- Web => User goes to a URL and downloads what they need
- Web application => User goes to a URL and launches an application at a remote site, then downloads the results
- Web service => Software at user's terminal finds the appropriate resource, goes to that URL, gets/does what is needed and returns it to the terminal for continued processing, seamlessly, hidden from user

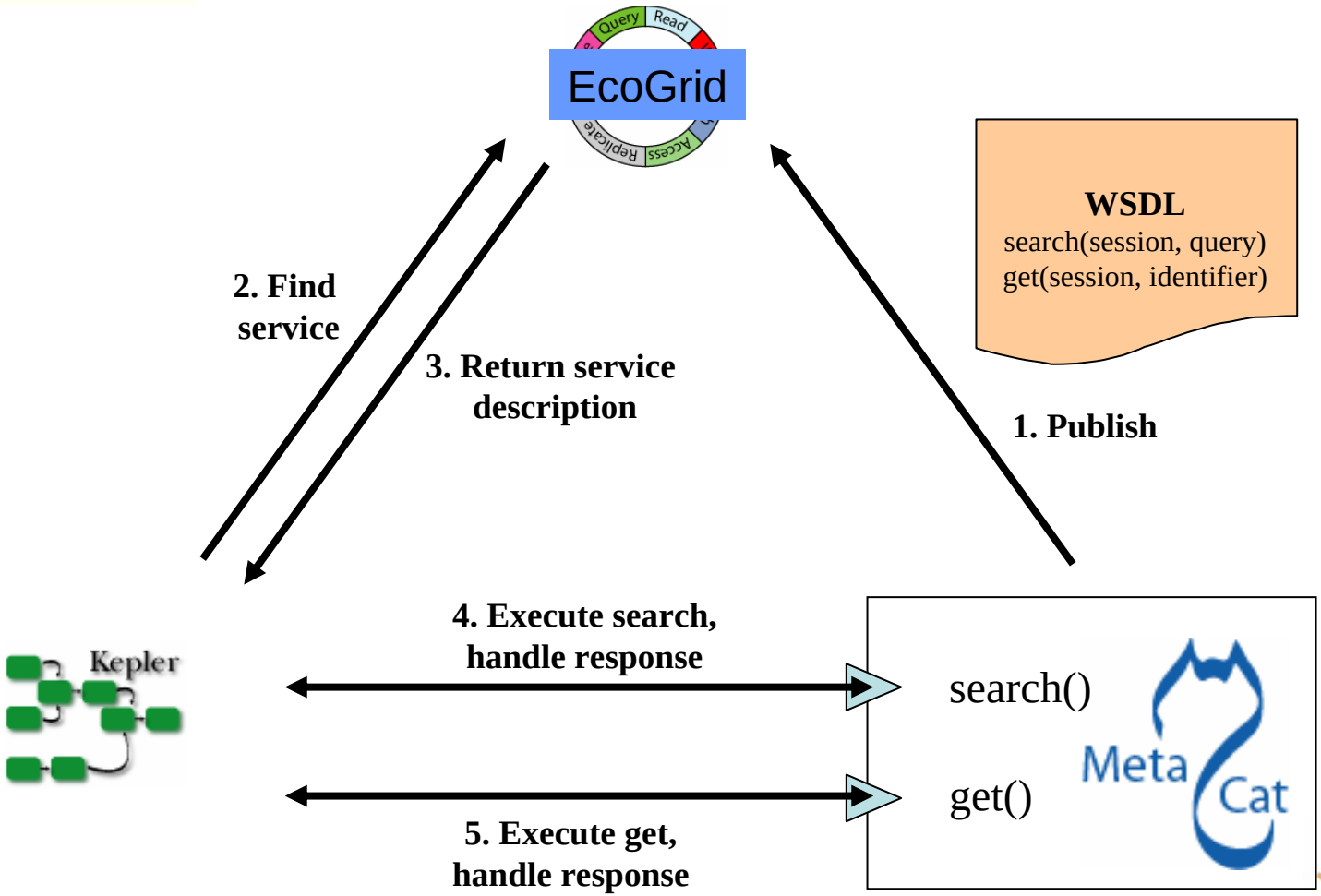
Web & grid services

Service-oriented architecture (SOA)

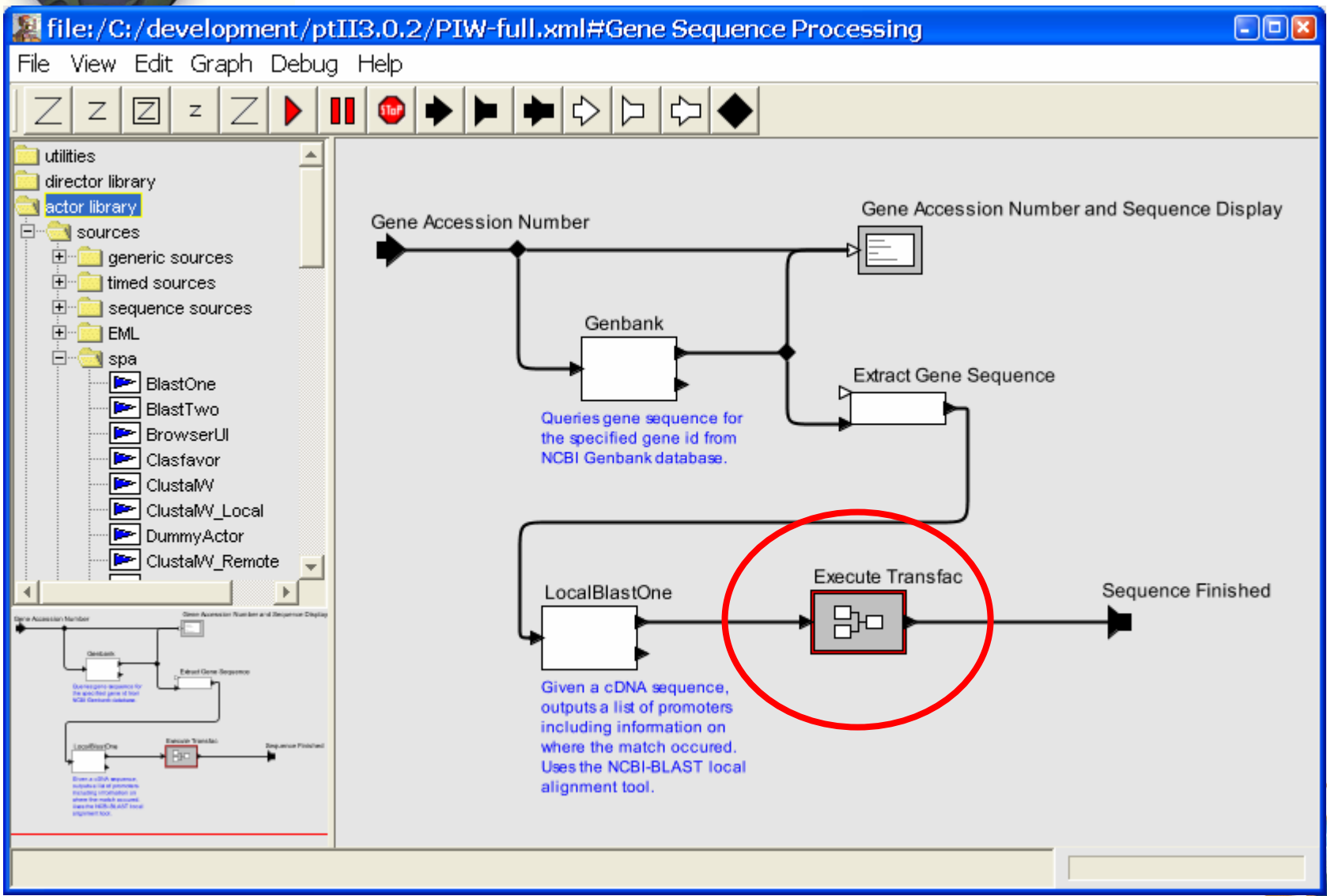
Message format (SOAP/XML)
Network transport of data (HTTP)
Service interface description (WSDL)



Example


EcoGrid
Query
Read
Access
Replicate
WSDL
search(session, query)
get(session, identifier)
2. Find service
3. Return service description
1. Publish
Kepler
4. Execute search, handle response
search()
get()
Meta Cat
5. Execute get, handle response

Kepler: web services access

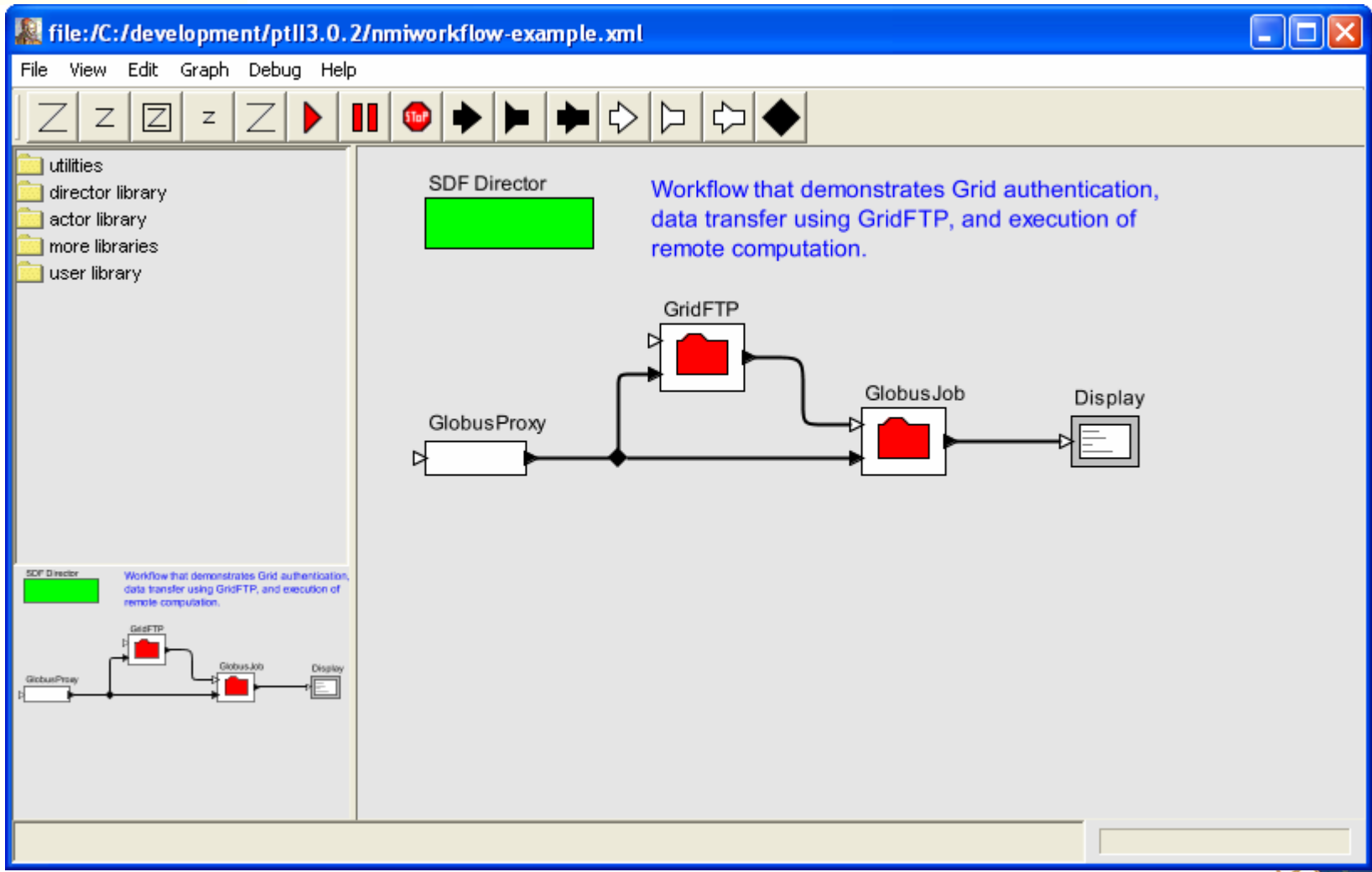


Grid services

A grid service is a web service plus

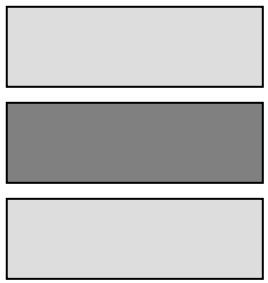
- Security
- State management (tracking sessions across multiple requests)
- Factory services (allowing many clients to connect)
- Lifecycle management (persisting the service over outages)
- And more…

Kepler: grid services access



What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

Semantic Mediation

Semantic = Of, or related to, meaning in language
Mediation = Mediation is a process in which a third-party neutral acts as a facilitator to assist in resolving a dispute between two or more parties

Semantic mediation = Process of resolving differences of meaning [in datasets or other entities] through an third-party facilitator

Dataset and tool characteristics:
•Can be described using terms (e.g. biomass, count, occurrence, beta diversity)
•Those terms are usually embedded in free text documentation (or left undocumented) and abbreviated in dataset schemas
•The terms [and abbreviations] represent concepts that must be captured and understood in order to determine how the data and/or tools may be integrated

- Identical terms mean different things
- Different terms mean the same thing (synonyms)
- Terms are related in some other way

Example

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

Data Integration Levels

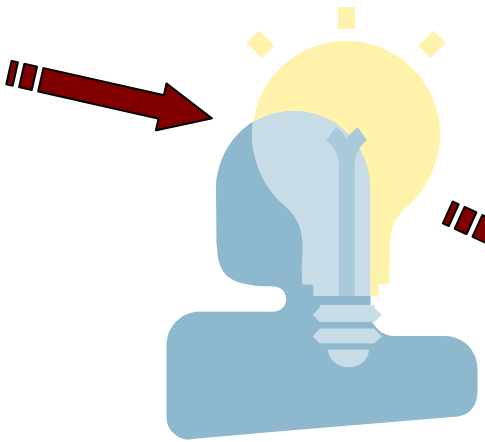
- Physical Level 
  - A dataset consists of one or more physical files that are stored in a particular format (for example, comma-delimited ASCII text, binary, etc)
- Logical Level 
  - Defines the structure of the data, how it is logically organized (order of rows/columns)
- Semantic Level

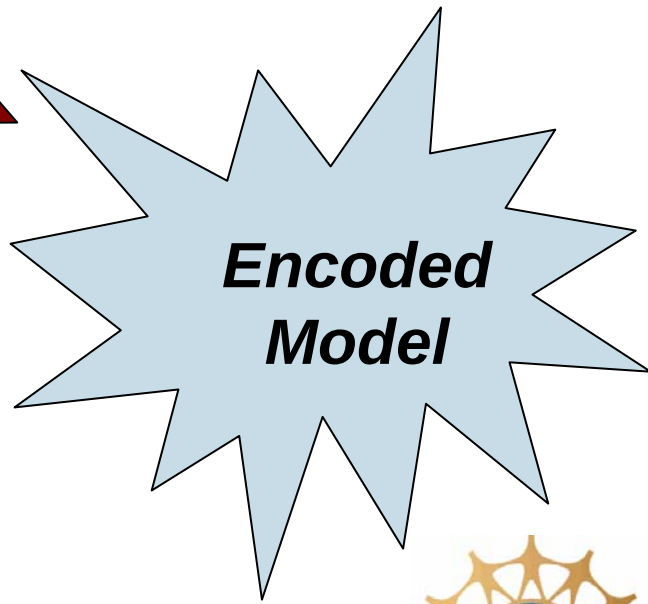
With detailed metadata

Analysis Integration Levels

- Physical Level
  - An analytic step is a particular software implementation that takes and produces physical data
- Logical Level
  - Defines the structure of input and output
- Semantic Level
  - Uses ontological information to conceptually define the analytic step

Knowledge Representation





- **An approach for encoding a conceptual model**
- **For the purpose of automated, intelligent reasoning**

Types of Encoding

Natural language

Semantic Networks

Ontologies

All encodings are necessarily imprecise and inaccurate, because the only completely precise and accurate representation of reality is reality itself.

The “best” encoding depends on the objective of the user…

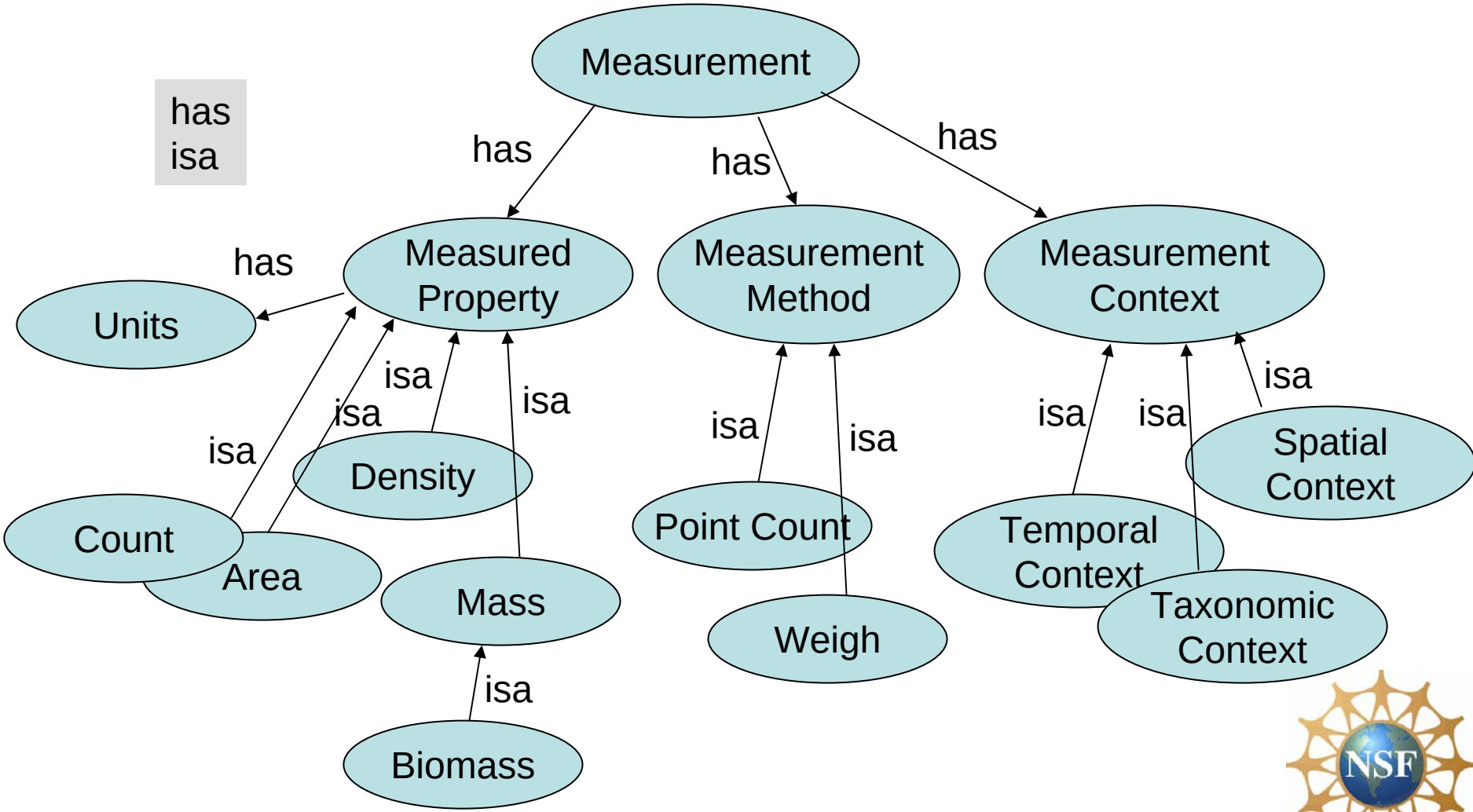
What is an Ontology?

- It depends on who you ask!
  - We focus on the data-management view

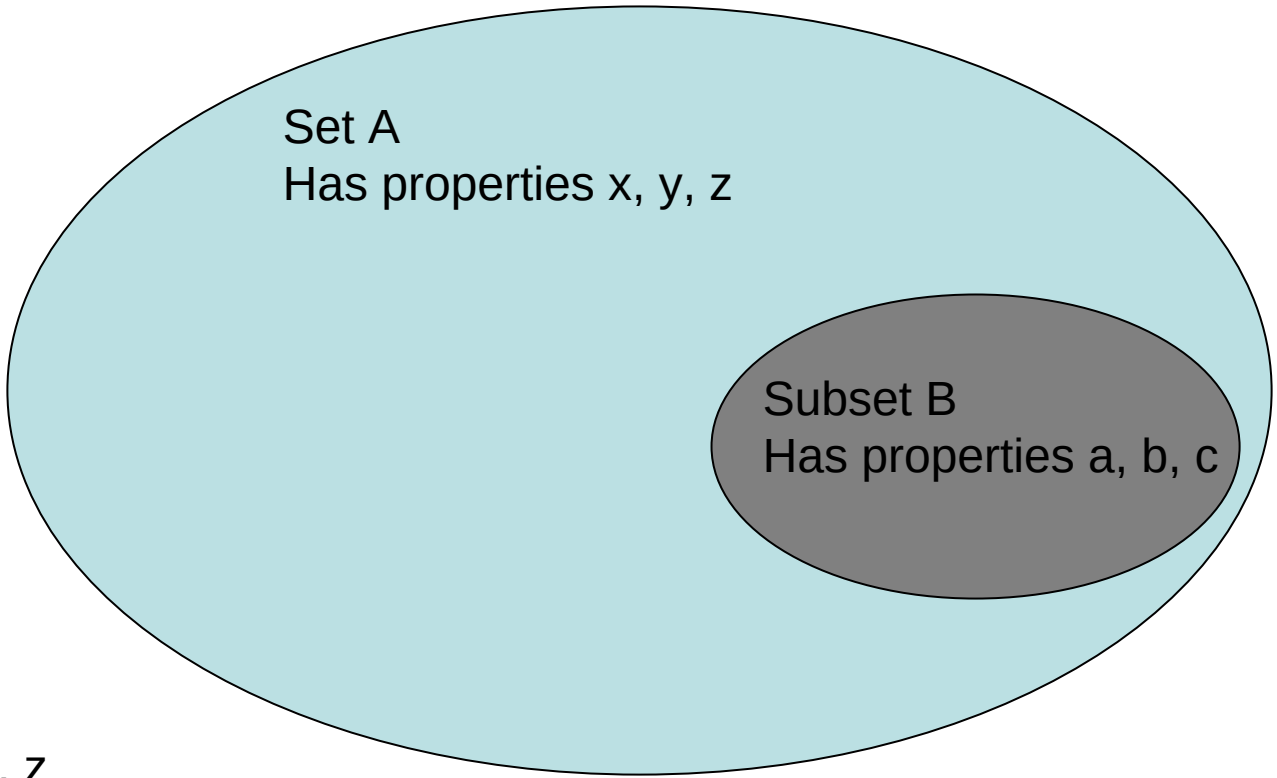
Generally speaking, an ontology specifies a conceptual model by defining and relating generic concepts representing features of the real or abstract world (within a domain of interest)

- Specified in a formal way, in a computer-readable language
- Using certain allowed constructs from set theory
- So that logic can be used to automate integration

Ontology example

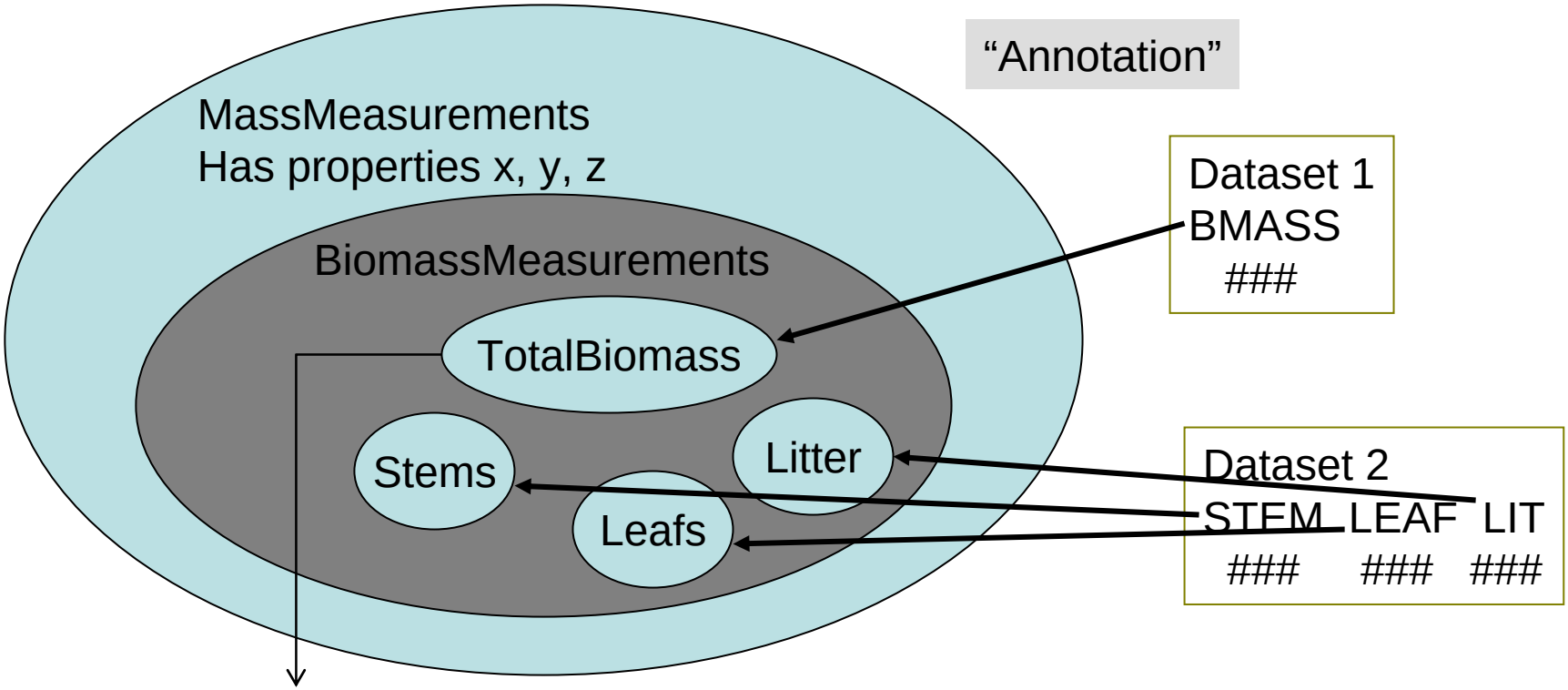


Sets



A has x, y, z
B isa A
=> B has x, y, z

Sets

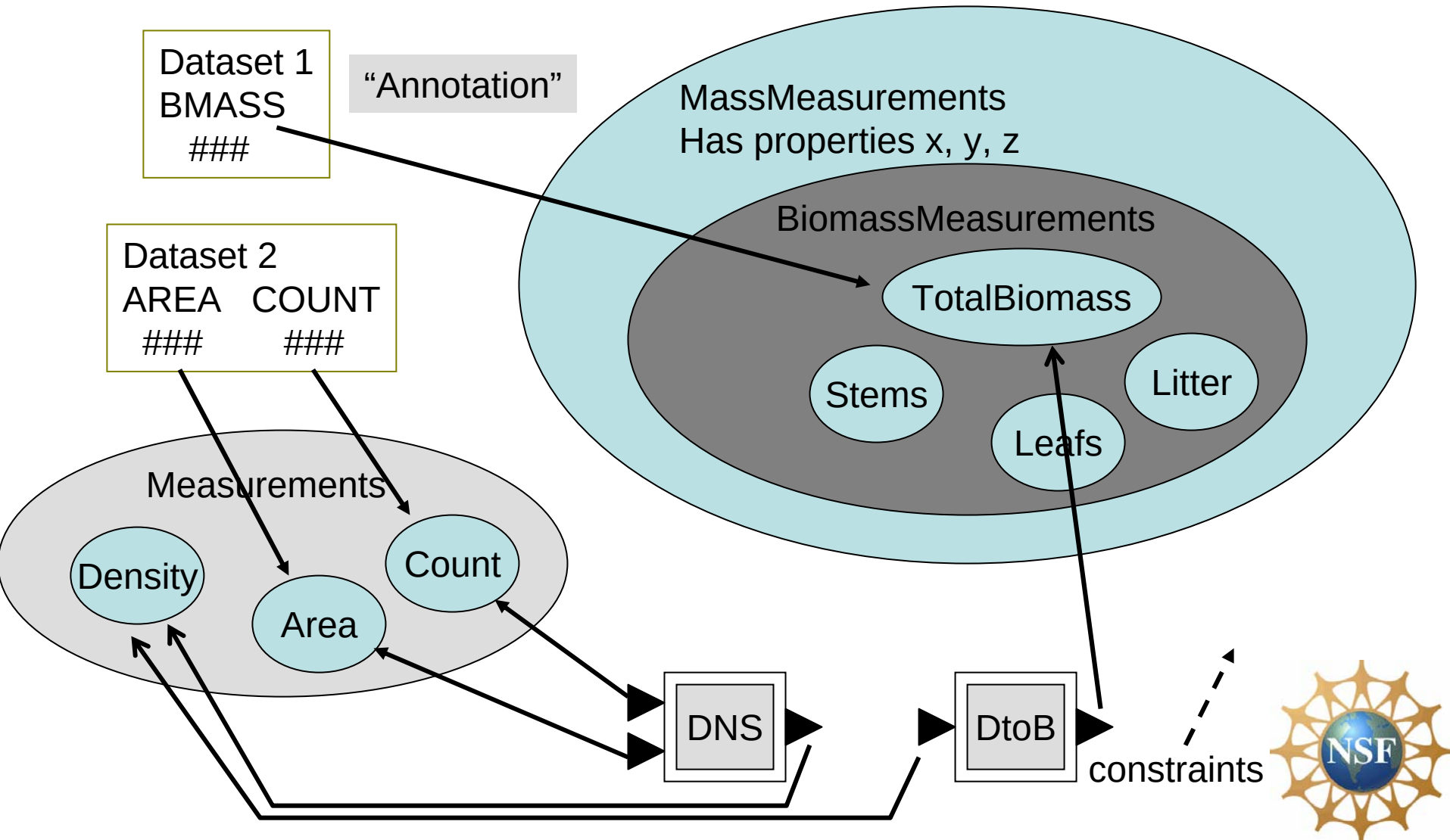


Necessary & Sufficient: Stems + Leafs + Litter

- Time and effort to build the ontology
- Requires domain & IT collaborative work
- Time and effort to annotate the datasets
- Payoff is in reusability***

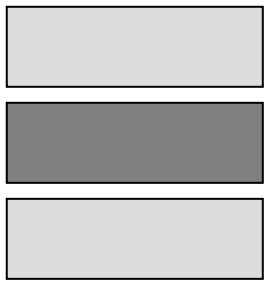
Given Area & Count => TotalBiomass
Under given constraints

Sets



What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

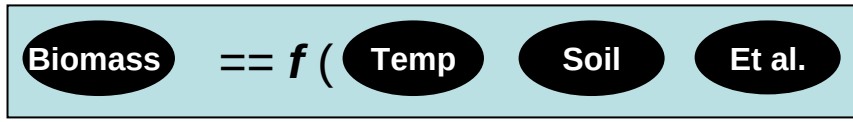
EcoGrid distributed resource system

Working groups:

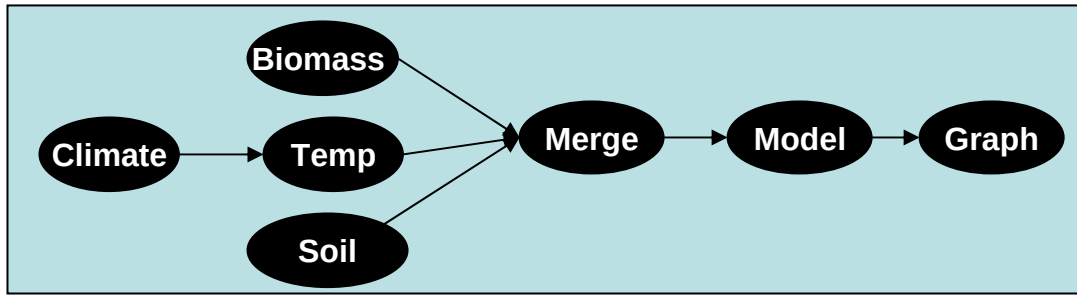
Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

Productivity Example

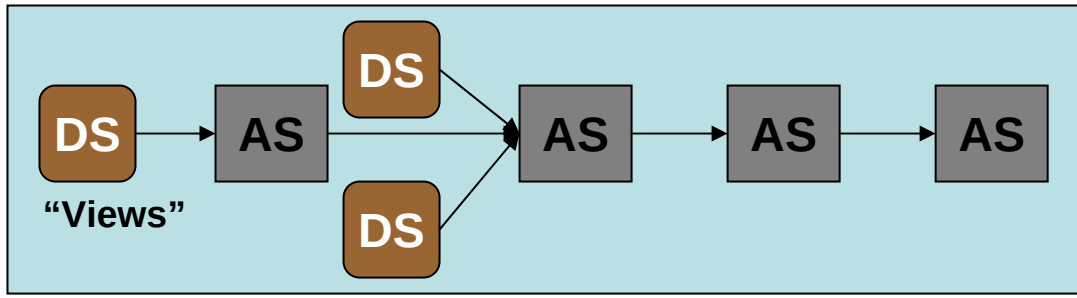
Mental Model



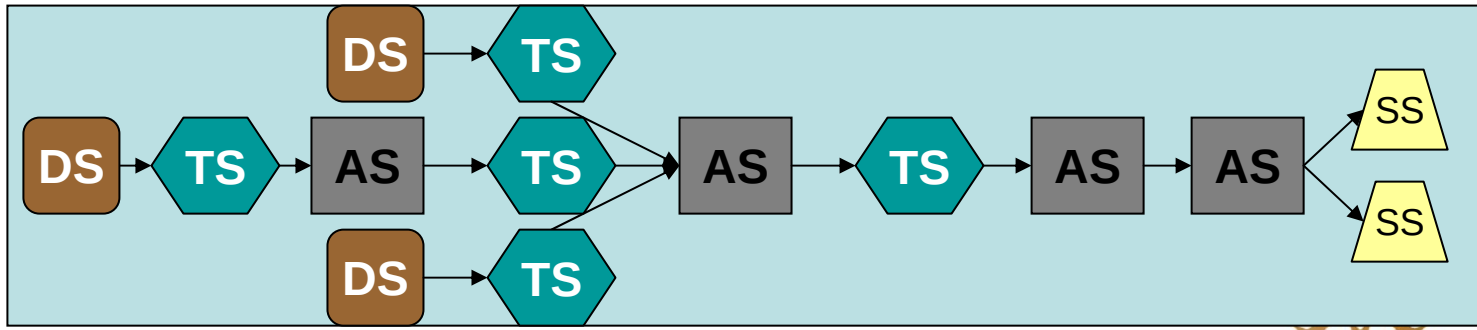
Conceptual Workflow



Abstract Workflow



Executable Workflow



C — Concept

DS — Data Step

AS — Analysis Step

TS — Transformation Step

SS — Sharing Step

Semantic Mediation System
Kepler Workflow System

Technology-enabled

Mental Mode

Ontologies

Biomass == *f* (Temp Soil Et al.

Conceptual Workflow

Biomass
Climate
Temp
Soil
Merge
Model
Graph

Abstract Workfl

Semantic Annotation

Data Discovery

Analysis Discovery

Executable Workflow

DS
AS
DS
AS
AS
AS

Workflow design

C Concept

DS Data Step

AS Analysis Step

TS Transformation Step

SS Sharing Step

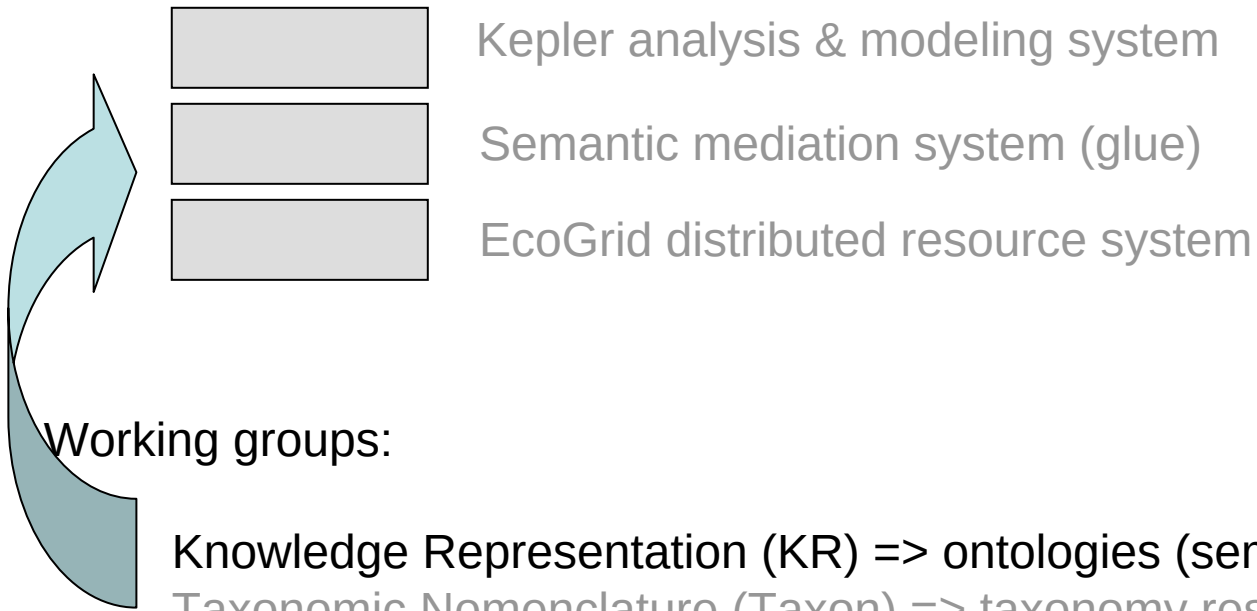
DS
TS
AS
TS
AS
AS
SS
SS

Semi-automatic data integration

Automate executable TS

What is SEEK?

System development:

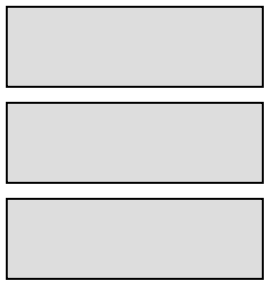


Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

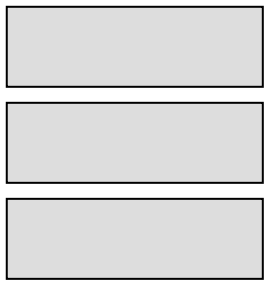
EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

Prototype Projects

Broad-scale analyses:

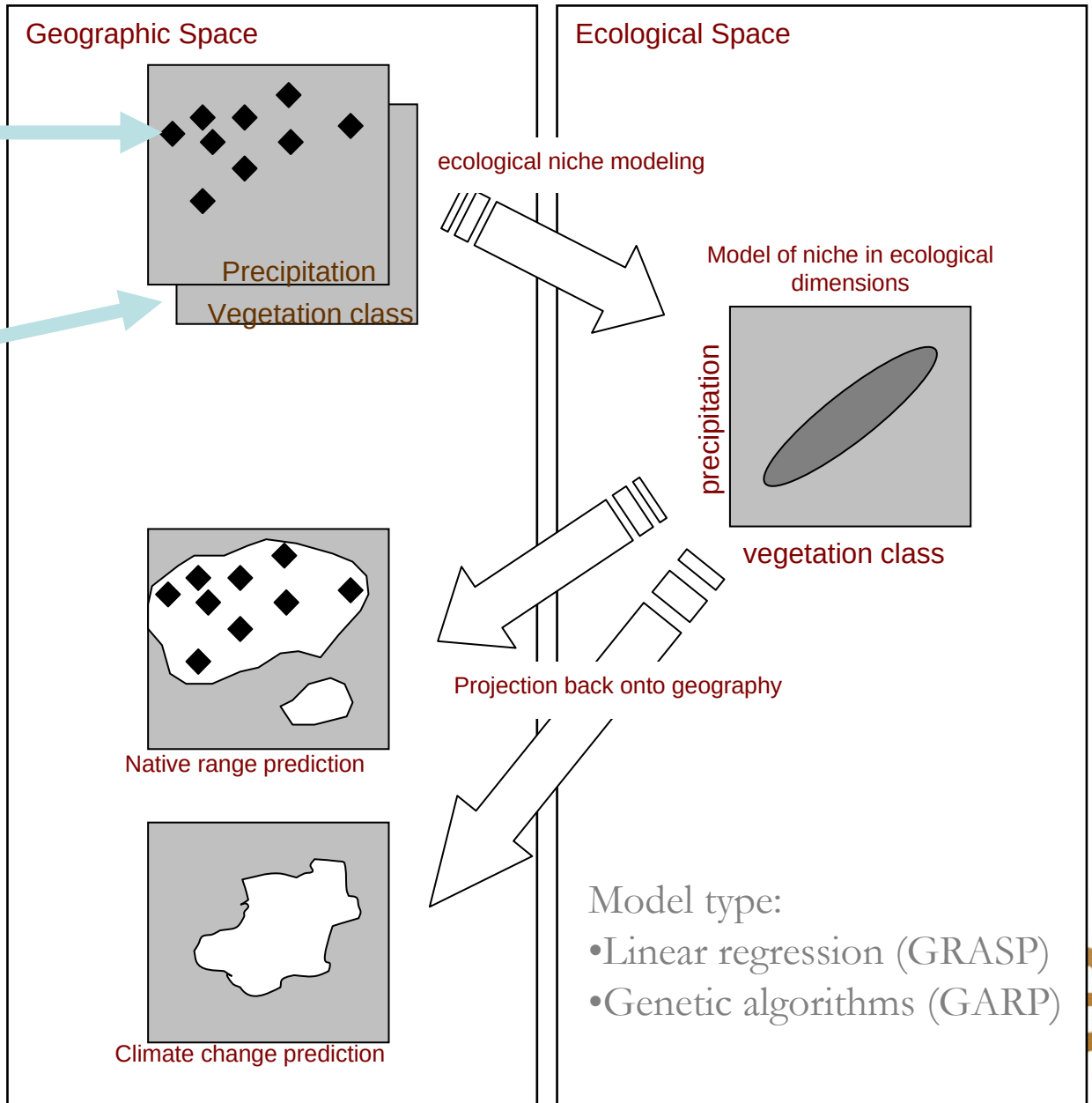
- Ecological niche modeling
  - Complex workflow using multiple software environments
  - Genetic algorithm uses many, diverse data streams
  - Computation intensive
- Biodiversity/productivity
  - Integration of complex, heterogeneous field data

Ecological Niche Modeling

Biodiversity information … e.g., data from museum specimens

Geospatial and remotely sensed imagery

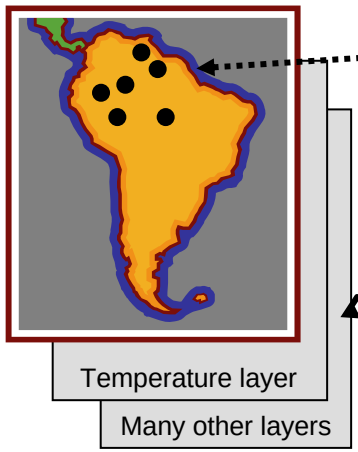
Results used for integration with other data realms (e.g., human populations, public health, etc.)



Model type:
- Linear regression (GRASP)
- Genetic algorithms (GARP)

Modified from B. Michener

Ecological Niche Modeling



Known Species Locations

Environmental Characteristics from gridded GIS layers

Develop Model

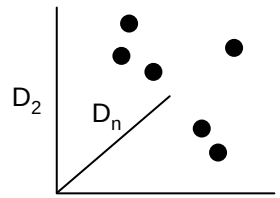
Environmental Characteristics Of Surrounding Geographic Area

Native Distribution Prediction

Environmental Characteristics Of Different Geographic Area

Invasion Area Prediction

Multidimensional Ecological Space



D_1 = Temperature

Future Scenarios Of Environmental Characteristics

Environmental Change Prediction

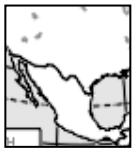
Ecological Niche Modeling

Excel File
Access File

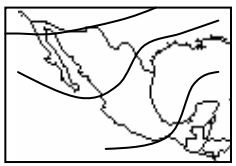
Sample 1, lat, long, presence
Sample 3, lat, long, absence
Sample 2, lat, long, presence

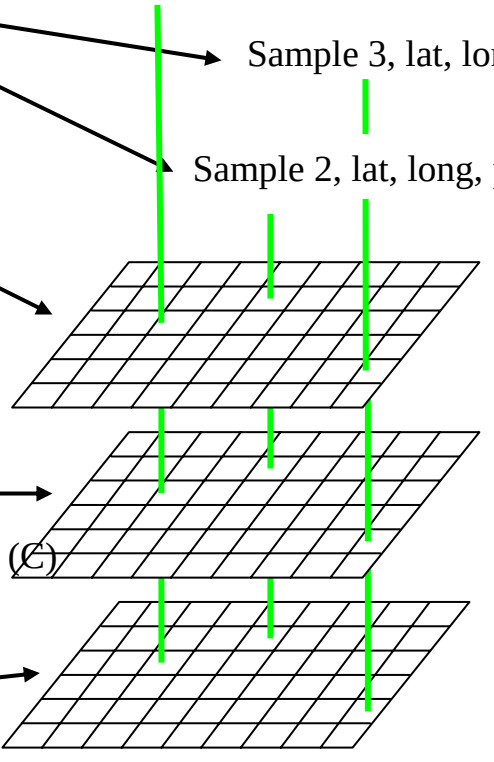
Vegetation cover type

Elevation (m)



Mean annual temperature (C)

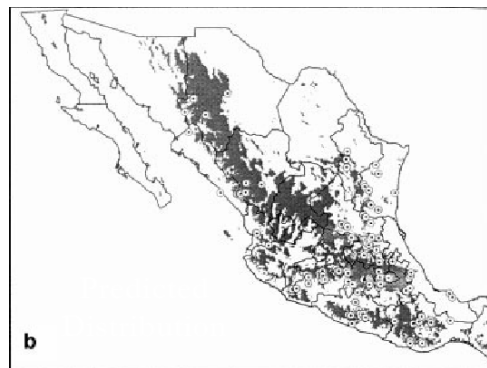




Integrated data:

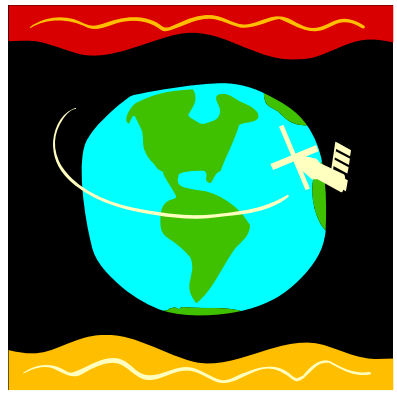
2200, 16…13
2320, 14…11
1535, 22…15
…
1823, 18…12

Genetic Algorithm (GARP) Model



…and potentially many more environmental layers

Mammal Project



Climate Change Analysis

2-3 dispersal scenarios

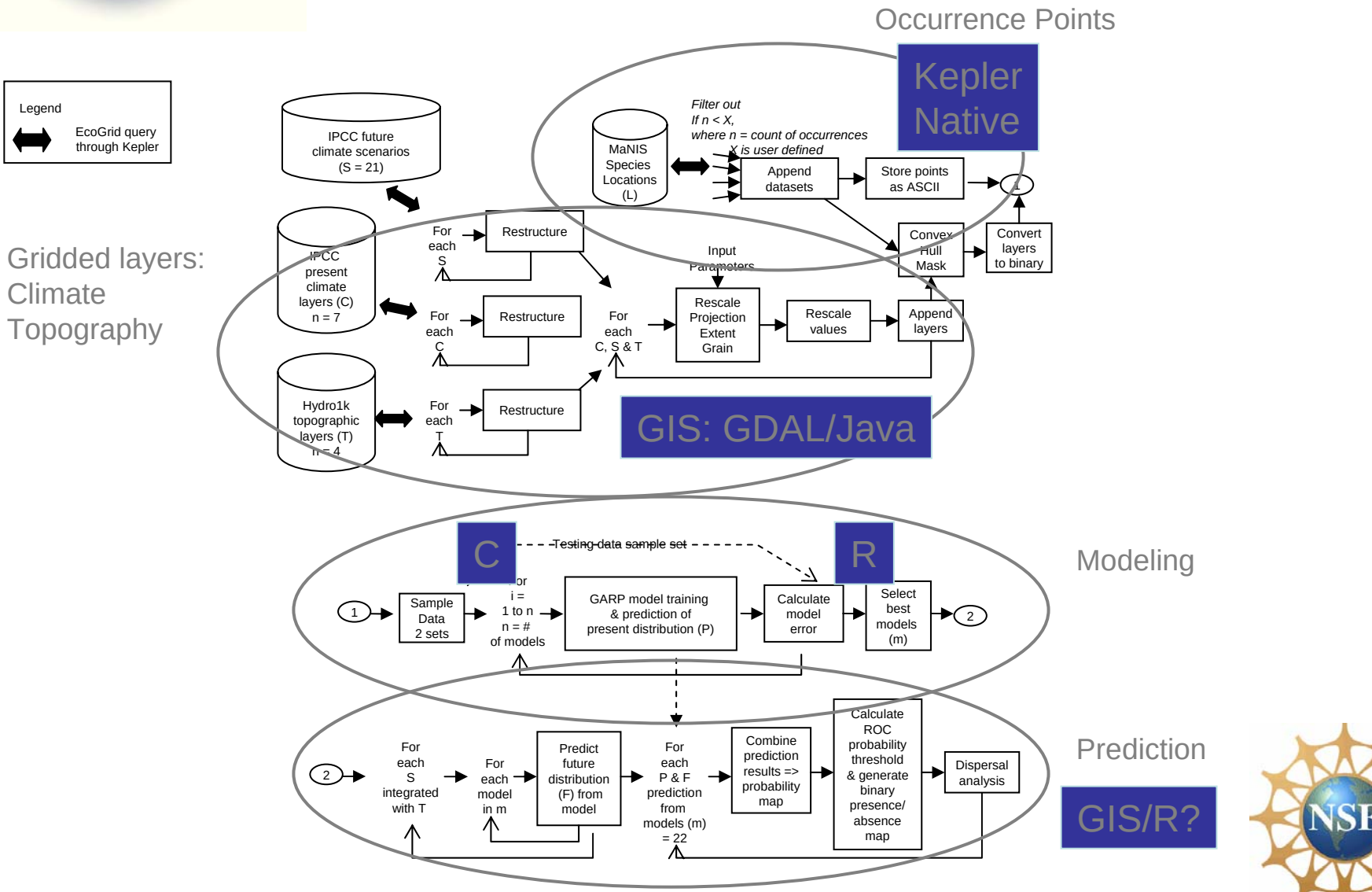
2 major evolutionary-computing algorithms (GA and NN)

21 GCM scenarios, including all IPCC scenarios

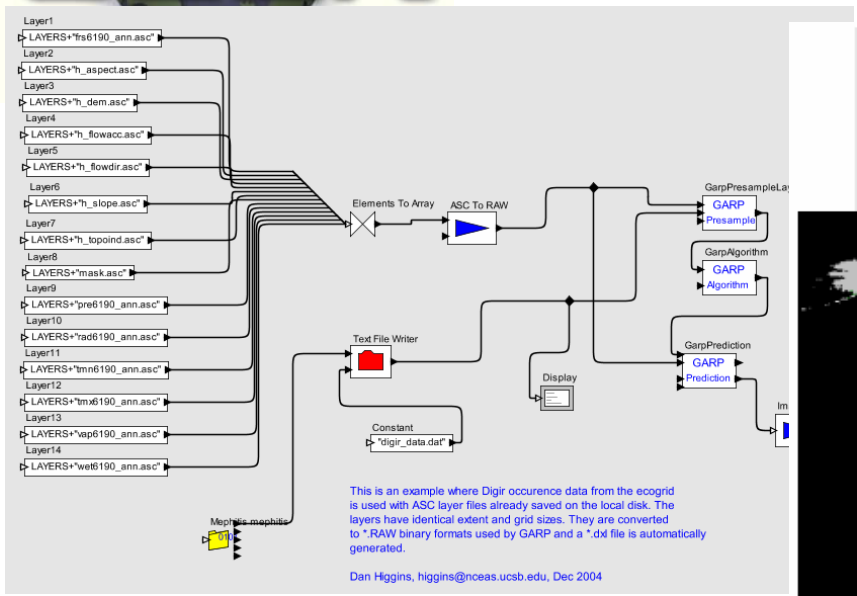
- 100 models/species
- 2 algorithms: GARP/NN
- 200,000 – 400,000 models
- 20 sec/model = 1500+ hours

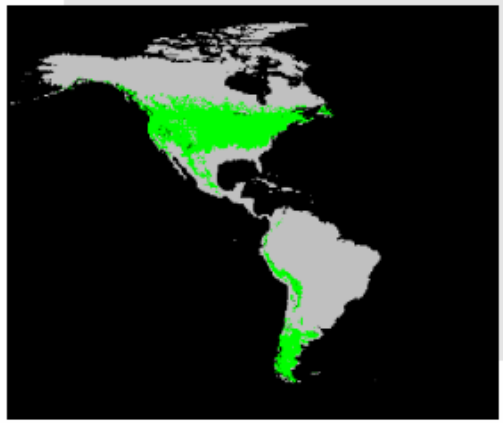
- Test large-scale implementation of Kepler

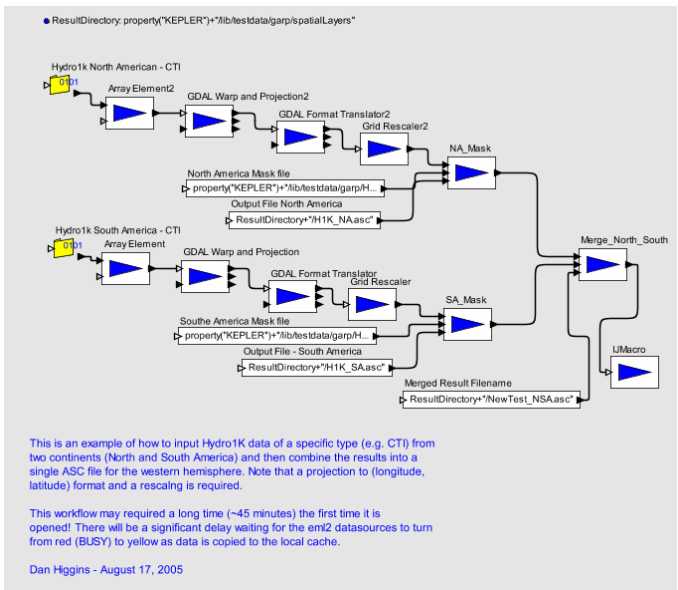
Abstract Workflow

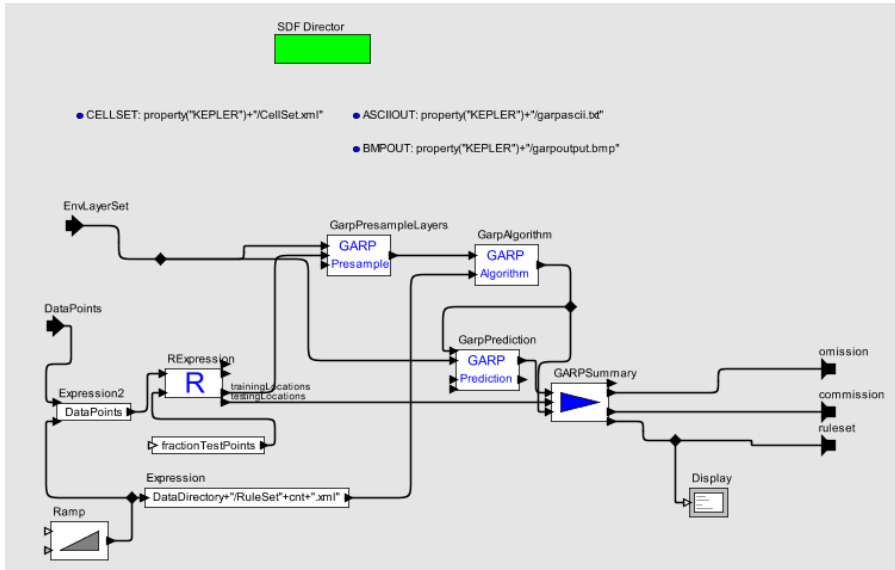


Executable Workflow

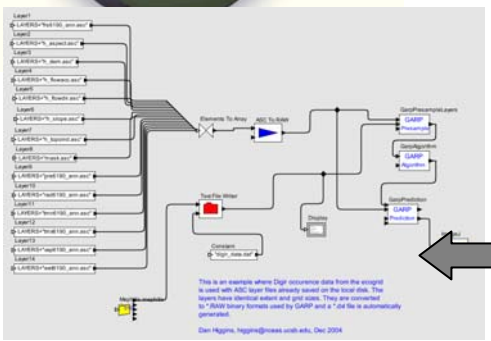


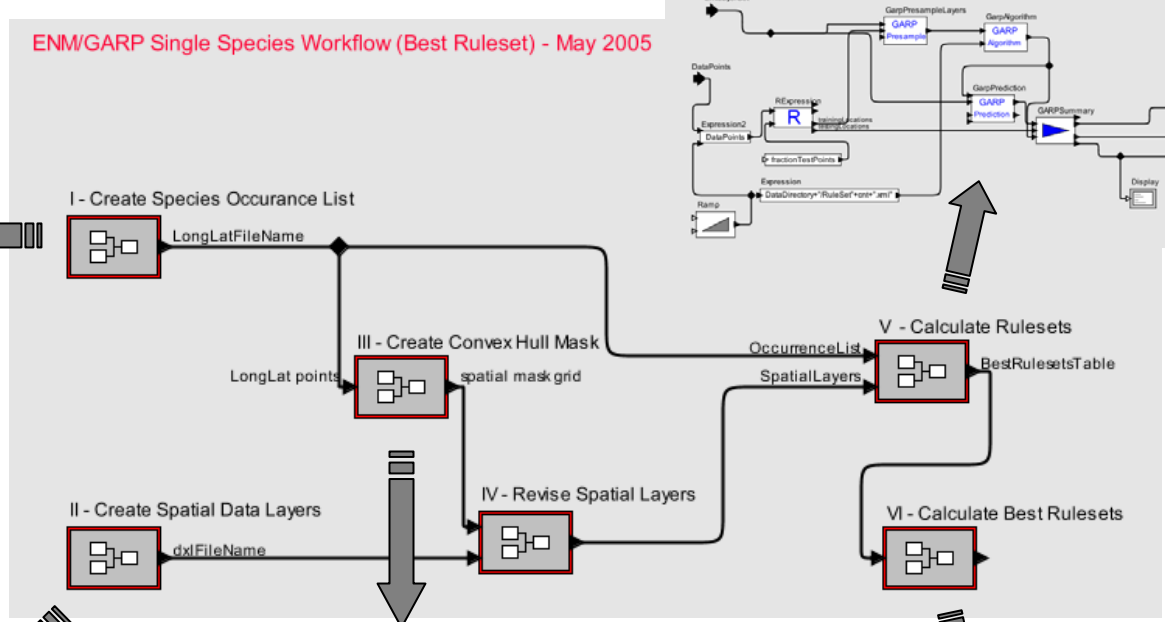


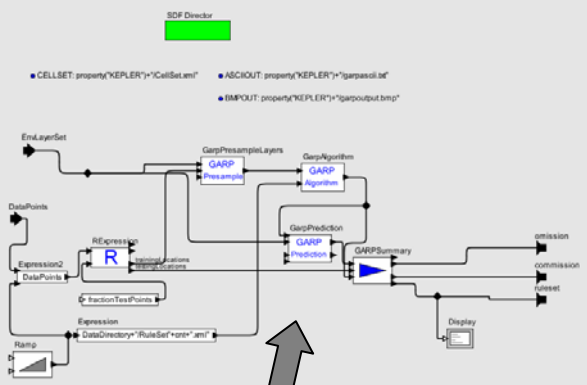


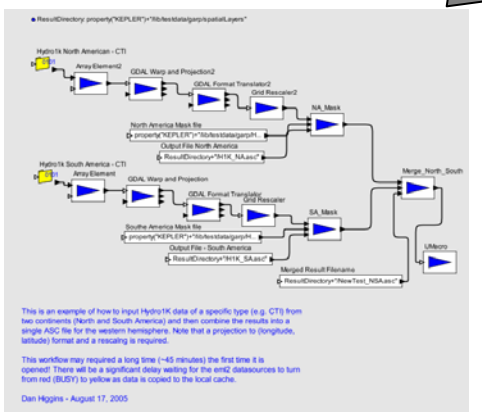


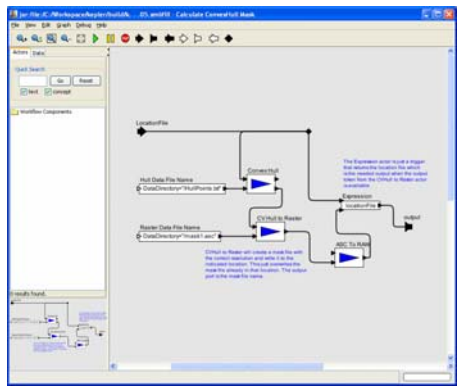
Executable Workflow

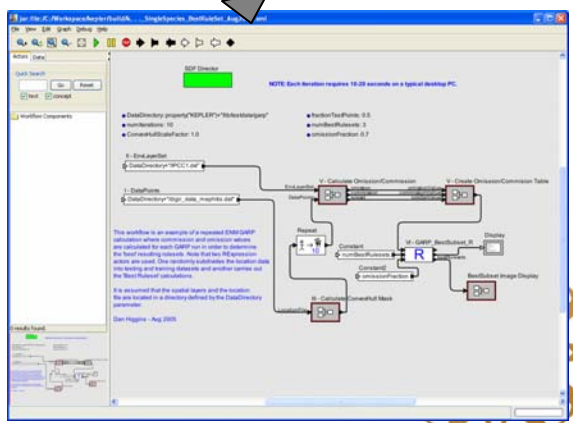




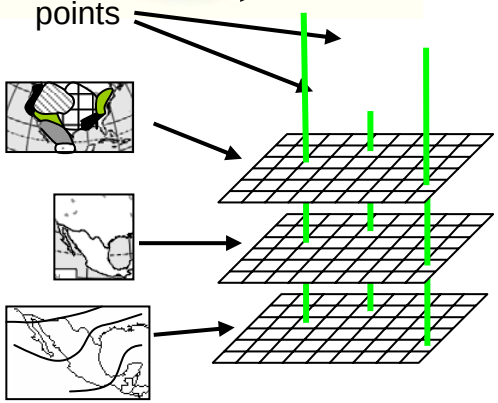


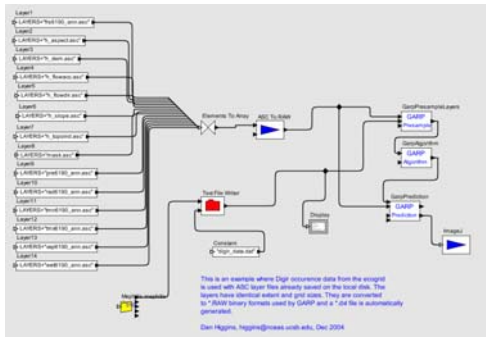


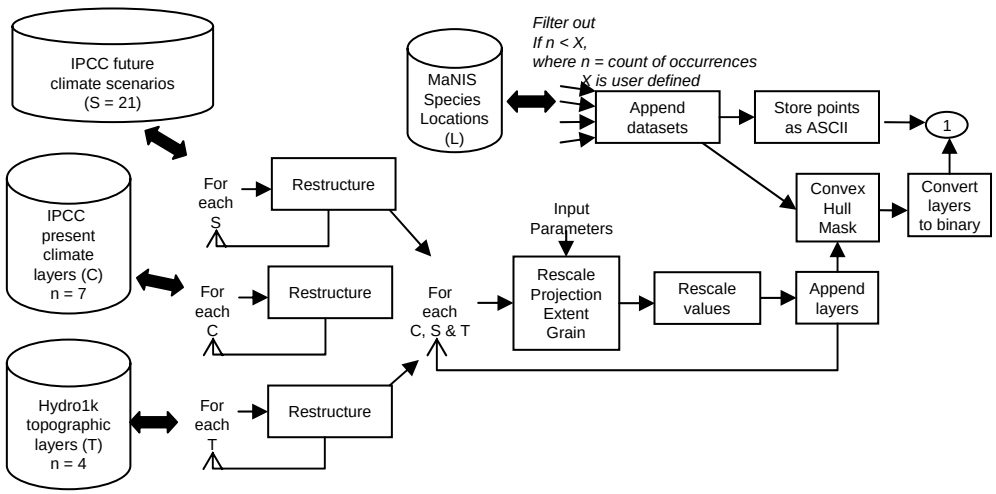


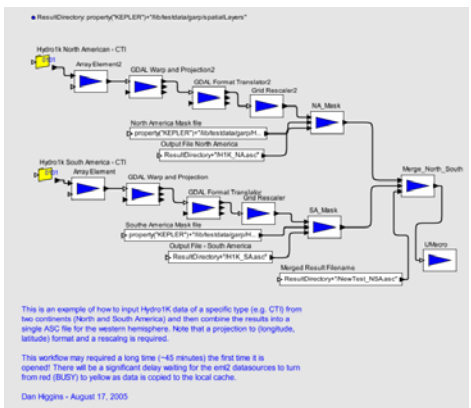


Data Preprocessing









Given:

Formally annotated input data

Formally annotated actor input ports

Semi-auto/auto data/actor integration?

Data & actor integration

Syntactic Level

Data as comma-delimited, plain text file

106.789098, 33.454606;
106.789097, 33.454606;
…

Data as binary

56.25 57.23 54.66 55.56 53.25
56.37 52.29 53.52 53.55 56.21
…

parameters

Sample Data

Structural Level

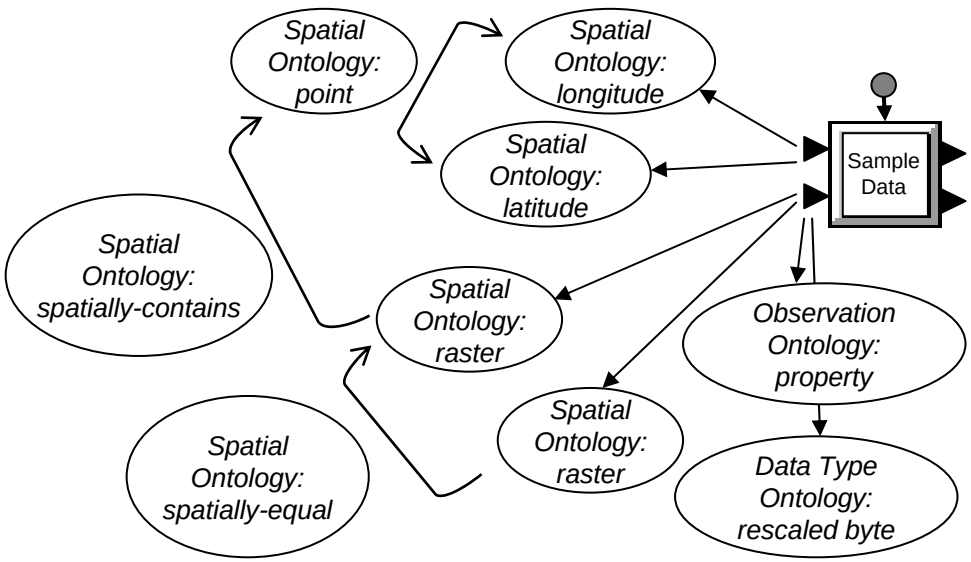
2-dimensional matrix — *matrix*[x, y]

List of 1-dimensional vectors, one vectpr per environmental layer
x & y inferred from position in vector — *list*(*vector*[z])

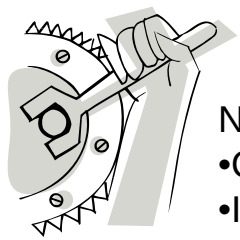
parameters

Sample Data

Semantic Level



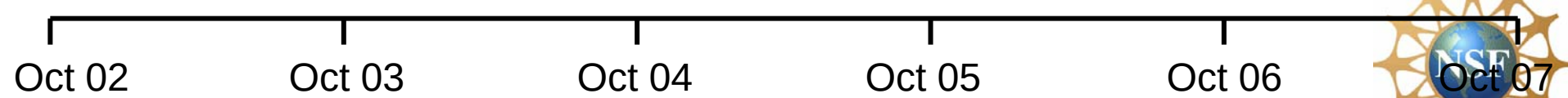
Issues:
- Ontology development
- KR expertise of person doing the annotating
- Scalability?



Nuts & bolts for:
- Constructing the ontologies
- Implementing the SMS

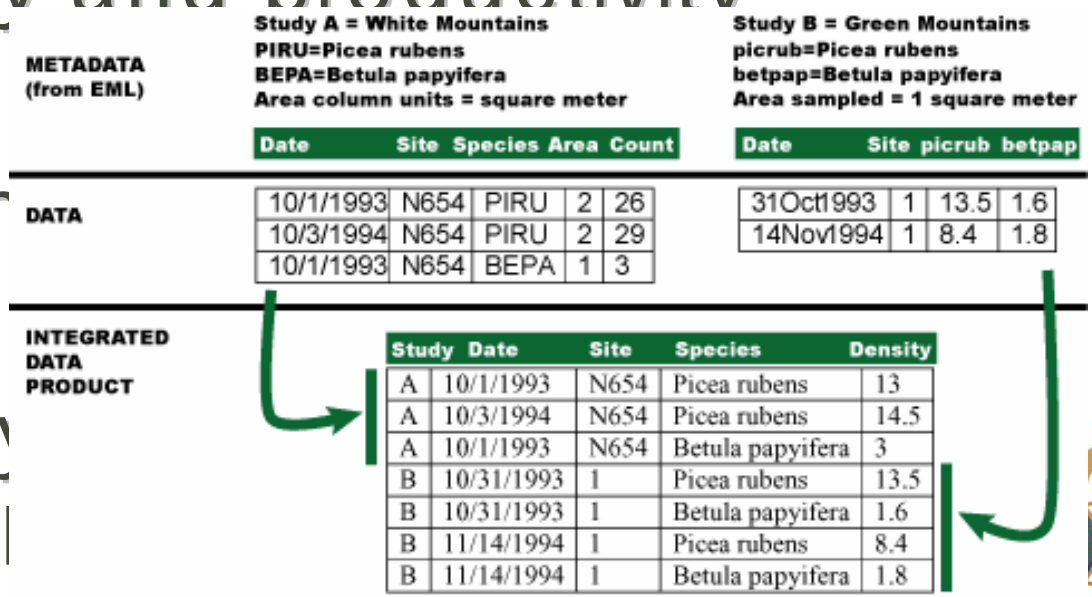
Timeline

- Timeline
  - Initial outreach (Apr 03)
  - Task analysis (Aug 03)
  - Workflow specification (Feb 04)
  - Training, plan functionality (Dec 04)
  - Mammal project results (Spring 06?)
  - Design follow up analyses (Summer 06?)



2nd Project Selection Criteria

- Syntax and schema transformations, semantic conversion
- Use biodiversity and productivity field data
- Conceptual lir
project (niche
- Use biodiversit
(demonstrate
- Expand the group of domain



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

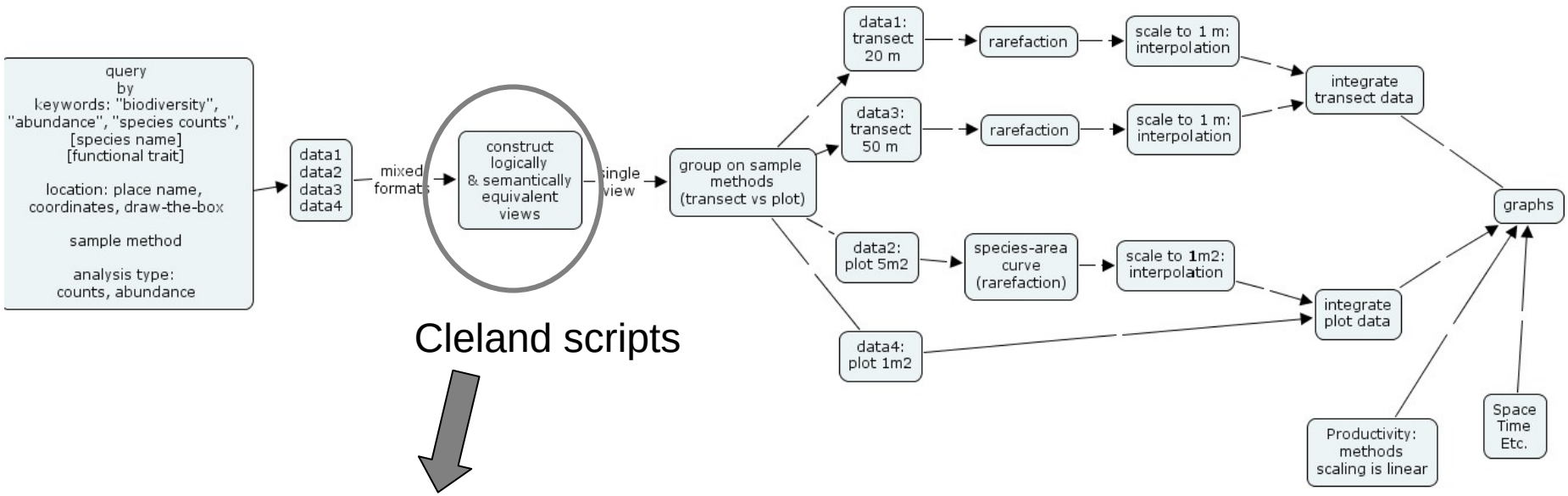
Date	Site	Species	Area	Count
10/1/1993	N654	PIRU	2	26
10/3/1994	N654	PIRU	2	29
10/1/1993	N654	BEPA	1	3

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

Abstract Workflows

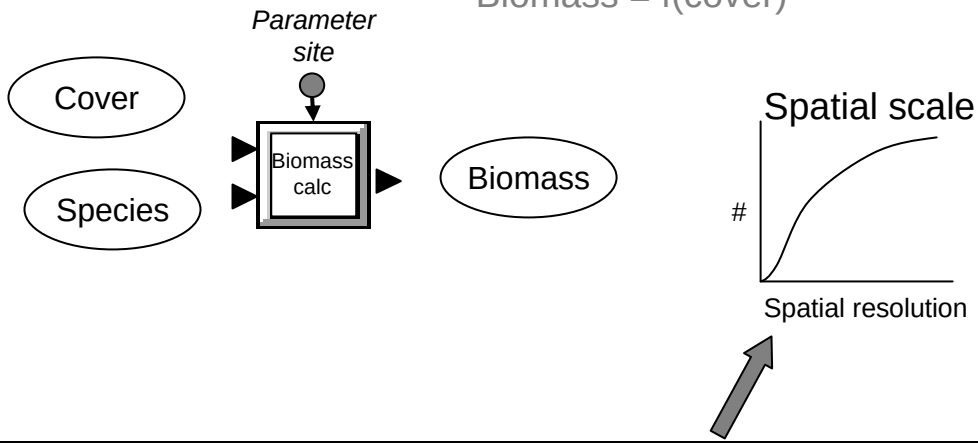


Cleland scripts

Data integration tool
- Data annotation
- User-guided mapping
- Archive final

Issues

Biomass = f(cover)

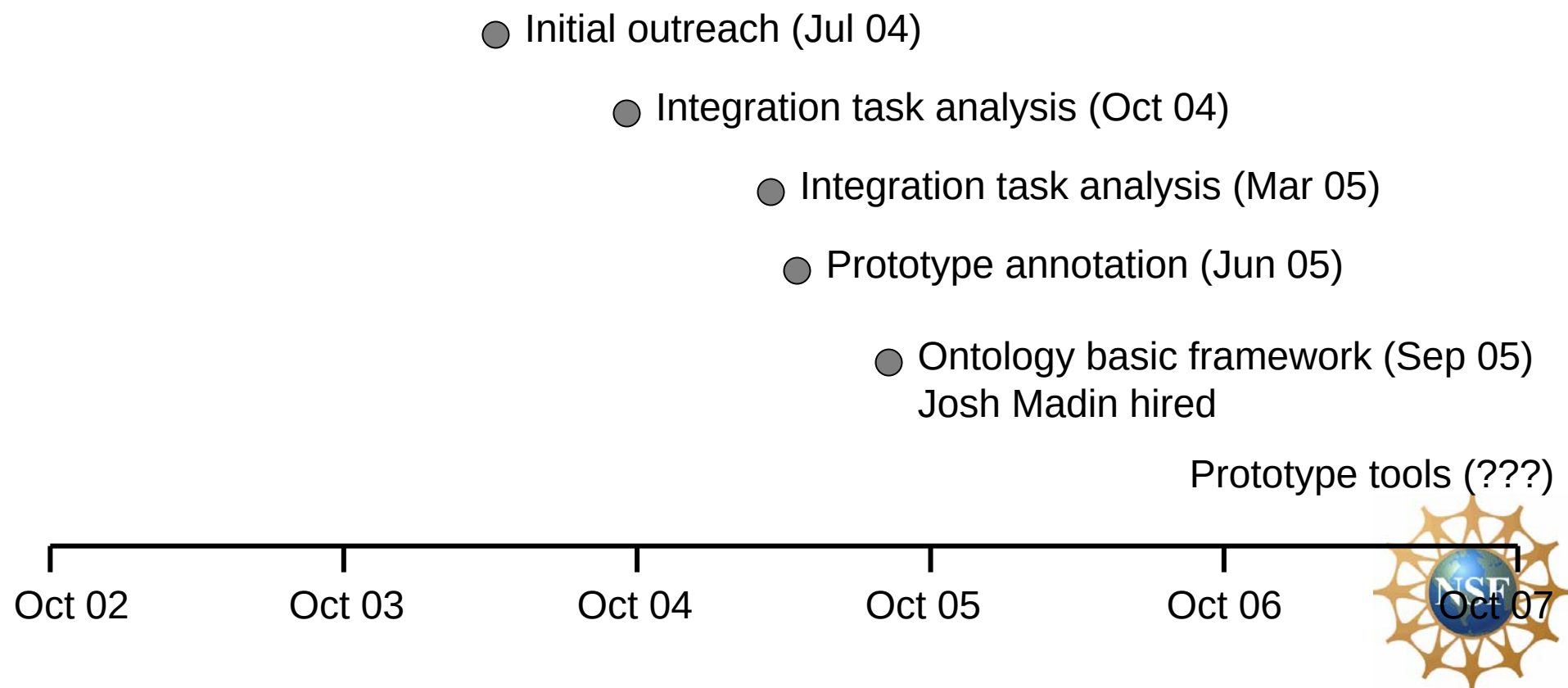


Treatment Codes

Site code	System Type and Location	Experiments (Community Types)	N amt (g /m^2/yr)	Form of N	Time N added	Trt plot (m^2)	Rep #	Resp Meas	Smpl area (m^2)	Duration	Year used
ARC	Arctic tundra, Toolik Lake, AK	5 types: dry heath (2 locations), moist acidic tussock, moist nonacidic tussock, moist tussock tundra	10	NH4-NO3 pellets	June	20, 100, 100, 100, 20	16 (2 true reps), 3, 4, 3, 16 (2 true)	Cover (visual)	1	1985, 89, 89, 97, 85 - present	14, 10, 10, 5, 14
CRP	Coastal salt marsh, Carperteria, CA	5 zones dominated by different species	420 (y1), 840 (y2-3) *	Urea and NH4 pellets	April and Nov	0.25	10	Cover	0.25	1999-present	3
CDR	Sand prairie/old field, Cedar Creek Natural History Area, MN	2 sites abandoned from agriculture 1957 (last crop=soybean) or 1934 (last crop=corn)	9.52	NH4-NO3 pellets	Mid-May and mid-June	16	6	Bio-mass	0.3	1982-present	Avg 17-19†

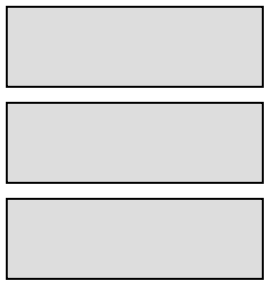
Timeline

- Timeline



What is SEEK?

System development:



Kepler analysis & modeling system

Semantic mediation system (glue)

EcoGrid distributed resource system

Working groups:

Knowledge Representation (KR) => ontologies (semantics)
Taxonomic Nomenclature (Taxon) => taxonomy resolution
Biodiversity and Ecologic Analysis and Modeling (BEAM)
Education, Outreach and Training (EOT)

Ecoinformatics Training

- October 20 – November 2, 2002 – Sevilleta
- October 27 – November 7, 2003 – Sevilleta
- January 4-9,2004 – Sevilleta
- September 28-30, 2004 – University of Santa Barbara
- October 17 - 30, 2004 – UNM
- January 3-7,2005 – UNM
- February 2-4, 2005– UNM
- October 31 – November 11, 2005 – UNM
- *January 9-13, 2006 – UNM*
- October 16-27, 2006 – La Selva Biological Station, CR
- January xx, 2007