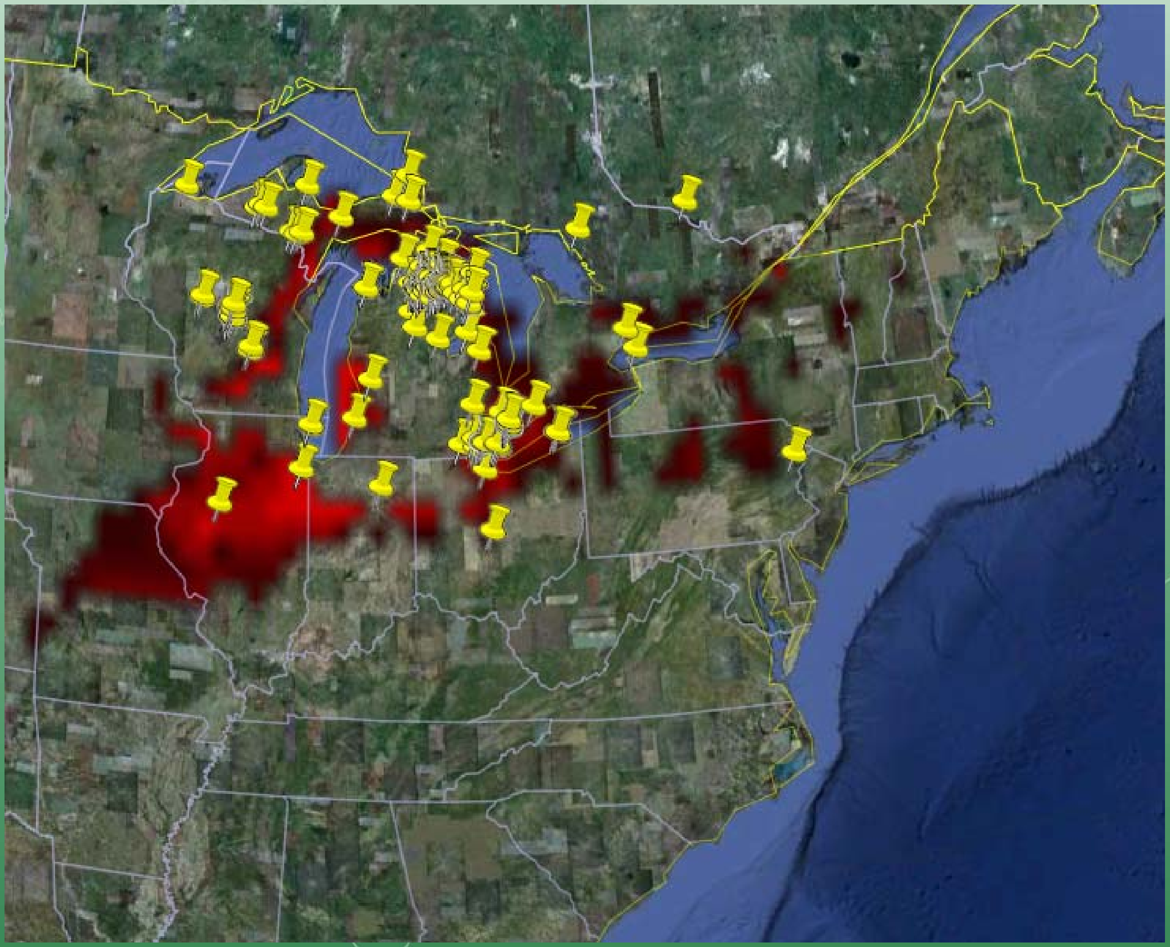
Enabling Biodiversity Research with Open Source Workflow, GIS and Metadata Tools

CJ Grady,
Jim Beach, Aimee Stewart, Jeff Cavner
University of Kansas Biodiversity Institute
http://lifemapper.org

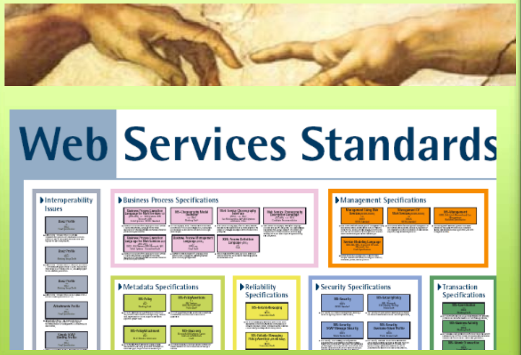
What is Lifemapper?



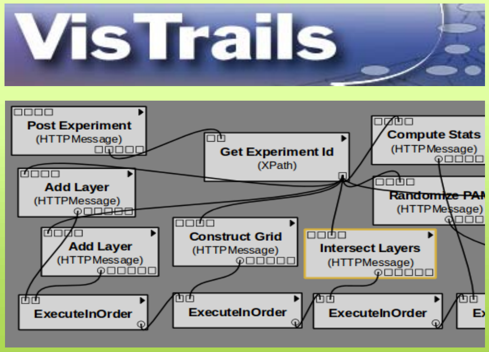
Current museum (GBIF) vouchered occurrences for Dendroica kirtlandii

What is Lifemapper?

Integration



Work Flow Tools



Archiving

DataONE

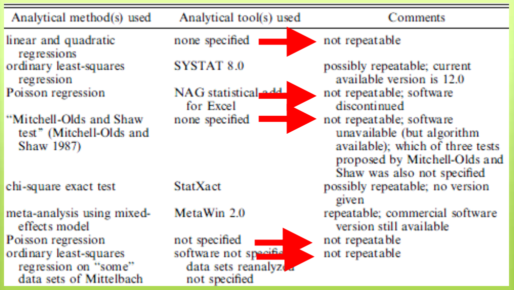
Meta Cat

```
<dataset id="ds.1">
  <title>Sample Dataset Description</title>
  <creator id="23445" scope="document">
   <individualName>
    <surName>Smith</surName>
   </individualName>
  </creator>
 </dataset>
</eml:eml>
```

EEML

Repeatable Transparent Science

Ellison, A. 2010. J. Ecology



Analytical method(s) used	Analytical tool(s) used	Comments
linear and quadratic regressions	none specified	not repeatable
ordinary least-squares regression	SYSTAT 8.0	possibly repeatable; current available version is 12.0
Poisson regression	NAG statistical add-for Excel	not repeatable; software discontinued
"Mitchell-Olds and Shaw test" (Mitchell-Olds and Shaw 1987)	none specified	not repeatable; software unavailable (but algorithm available); which of three tests proposed by Mitchell-Olds and Shaw was also not specified
chi-square exact test	StatXact	possibly repeatable; no version given
meta-analysis using mixed-effects model	MetaWin 2.0	repeatable; commercial software version still available
Poisson regression	not specified	not repeatable
ordinary least-squares regression on "some" data sets of Mittelbach	software not specified data sets reanalyzed not specified	not repeatable

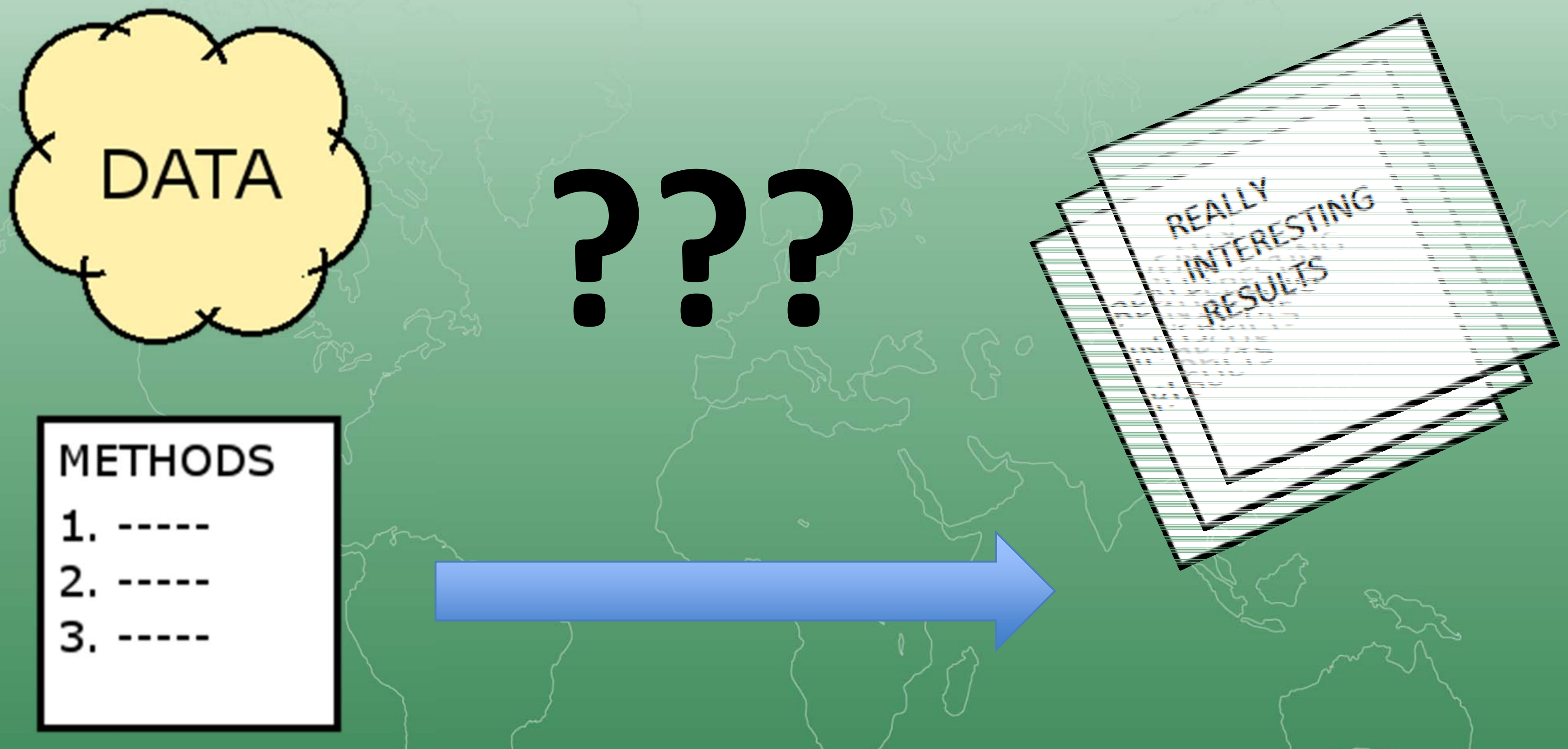
Background

- Workflows
  - A series of connected steps describing a process
- Metadata
  - Information that describes a data set or process.
  - Data about data
- Open Source Software
  - Software that includes access to the source code used to create it. This is provided to encourage study, contribution and sharing.

So What’s the Problem?

- Difficult and time consuming to assemble biodiversity experiments by hand
- Scientists often don’t have adequate computing resources
- Experiments can be difficult or impossible to reproduce

Study of Experiment Reproducibility



DATA

METHODS

1. -----
2. -----
3. -----

???

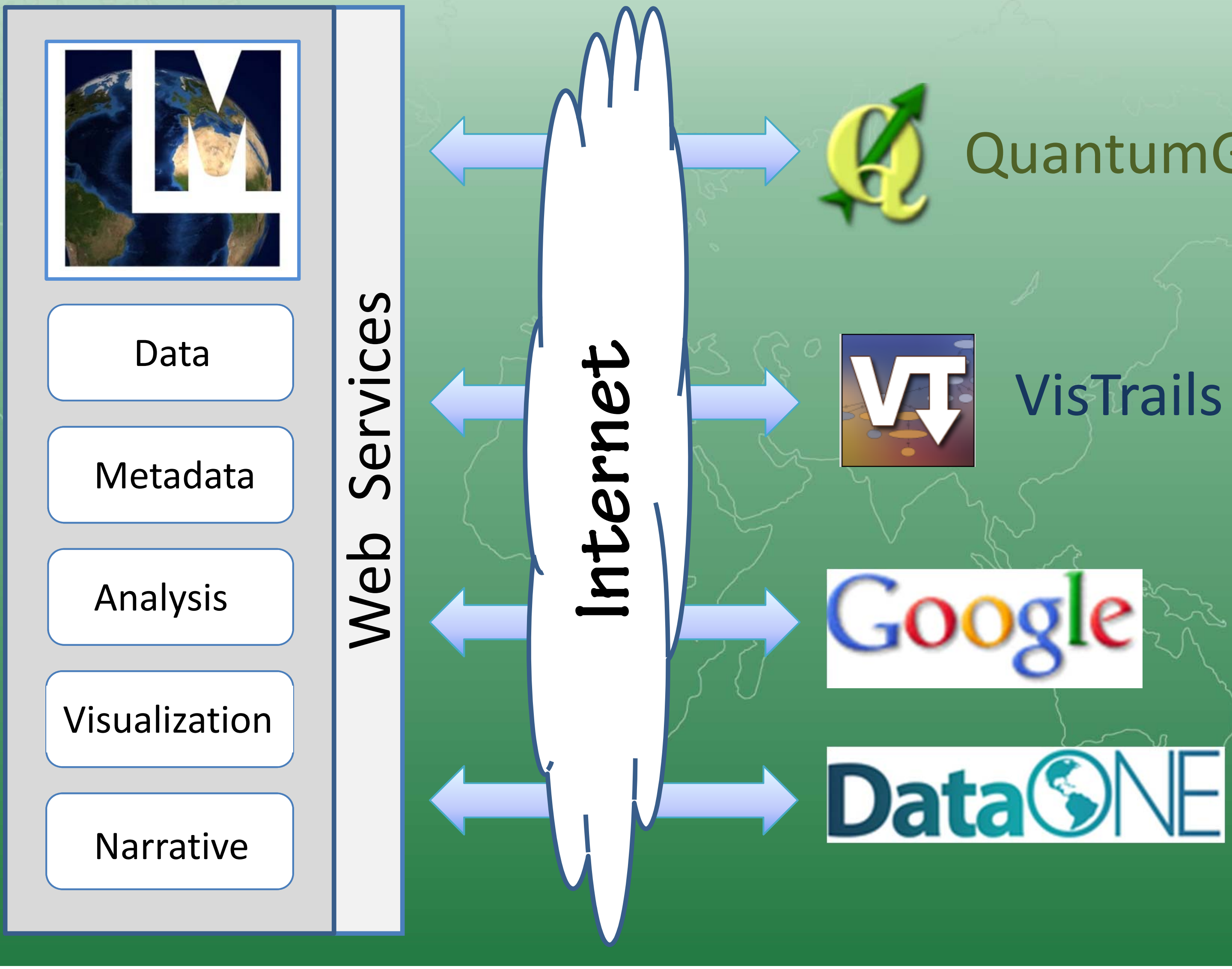
REALLY INTERESTING RESULTS

Ellison, Aaron. 2010. Repeatability and transparency in ecological research. Ecology 90.

What we have done

- Metadata for all of our Species Distribution Modeling services
- Simple process metadata
  - Documents how an experiment is ran through our cluster including what versions of software
  - Also describes what web services would be called to execute the experiment again

Web Services



Data

Metadata

Analysis

Visualization

Narrative

Web Services

Internet

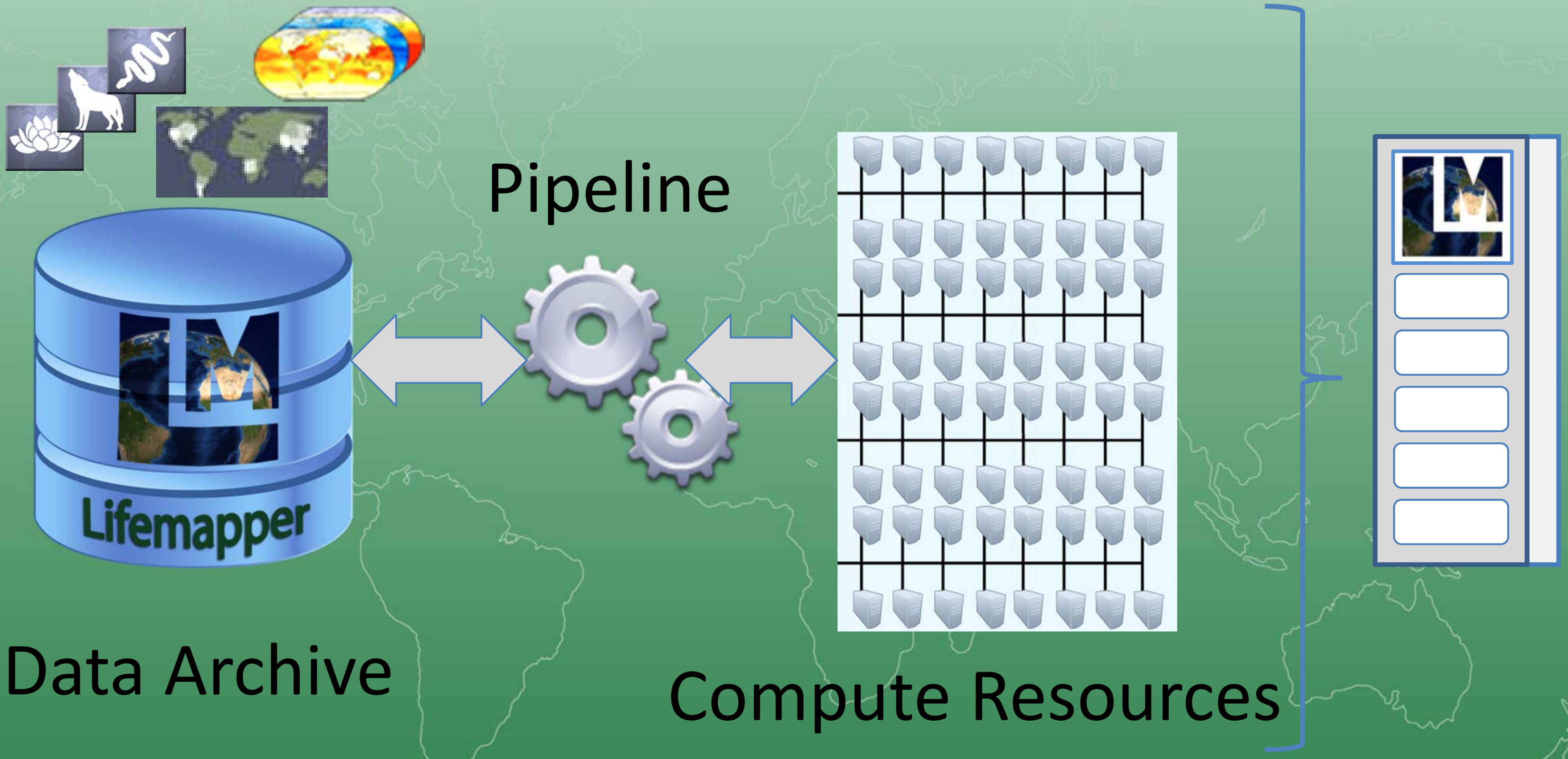
QuantumGIS

VisTrails

Google

DataONE

Lifemapper Backend

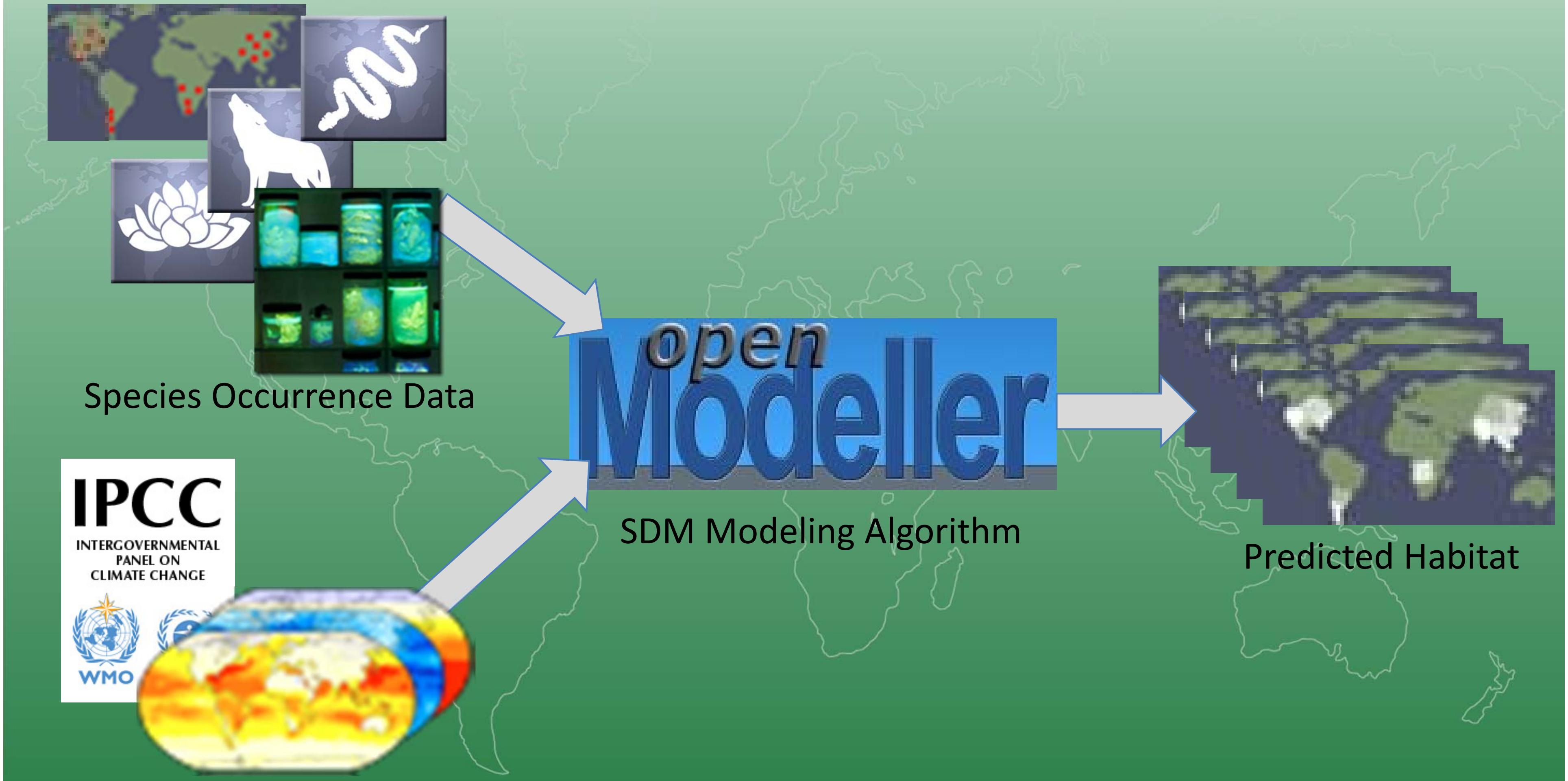


Pipeline

Data Archive

Compute Resources

Species Distribution Modeling



Species Occurrence Data

SDM Modeling Algorithm

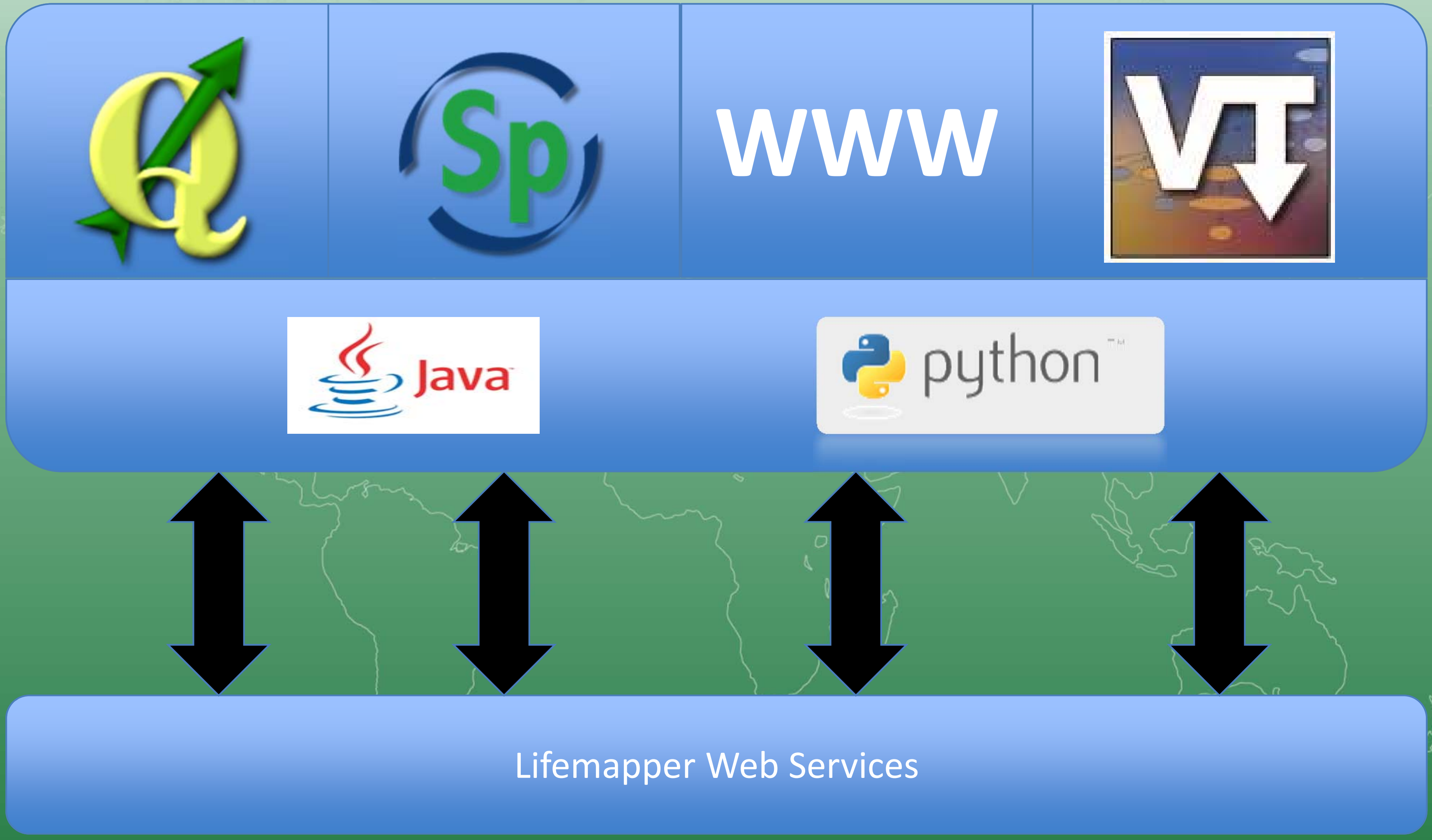
Predicted Habitat

Environmental Data

What we are doing

- Publishing metadata to a public repository
- Client extensions
- Lifemapper Range and Diversity

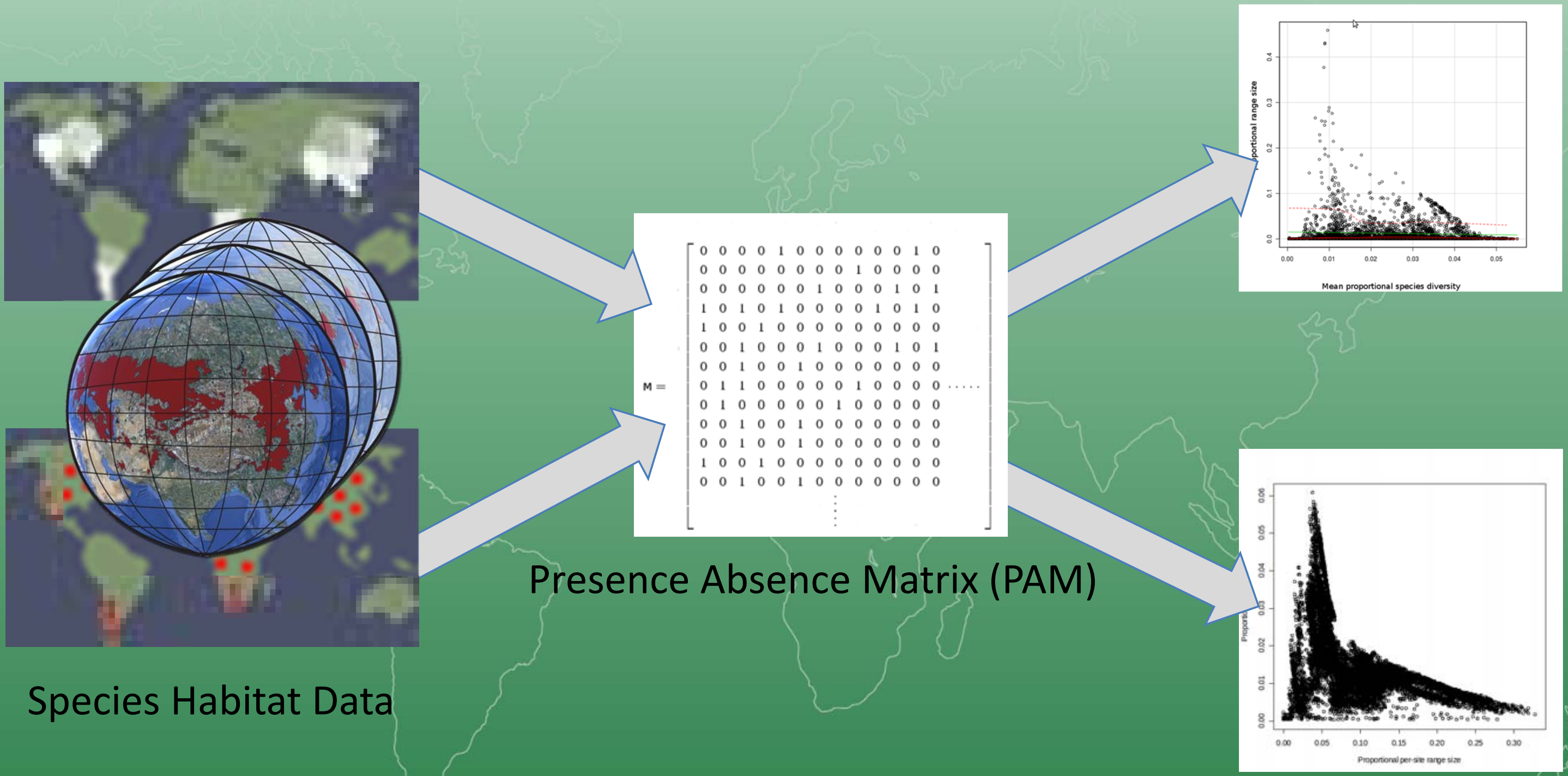
Clients



www

Lifemapper Web Services

LmRAD: Range and Diversity



Species Habitat Data

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & \vdots \end{bmatrix}$$

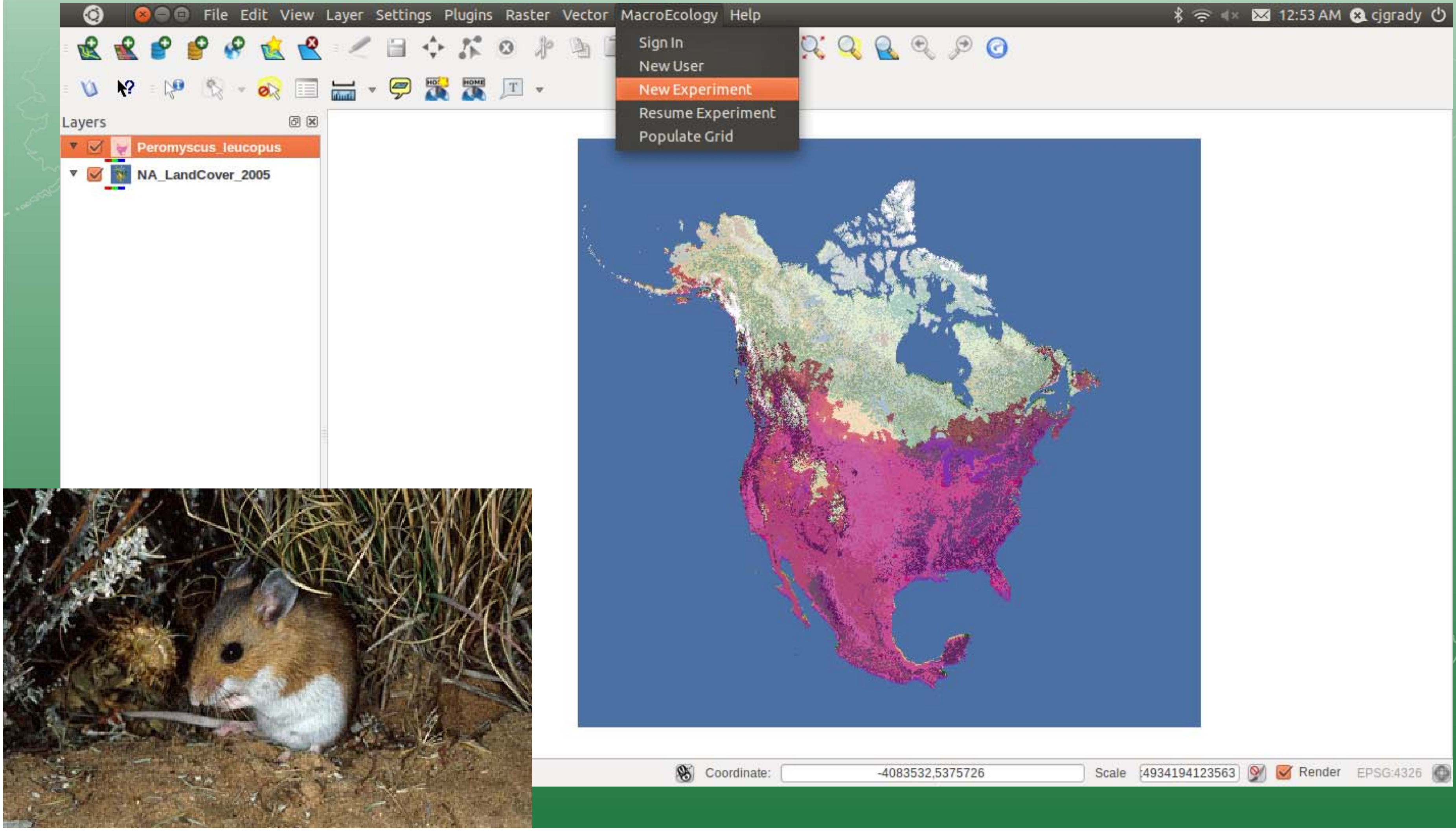
Presence Absence Matrix (PAM)

Range and Diversity Quantifications

Species
Distribution and Diversity

- Biodiversity patterns
  - Species abundance, distribution and diversity
  - Multiple scales and extents
  - Used for conservation and management decisions
- Challenges
  - Large extents (> 10,000 km2)
  - Fine resolution (< 1000 m2 ≈ 30m x 30m)
  - Many species (10,000 +)

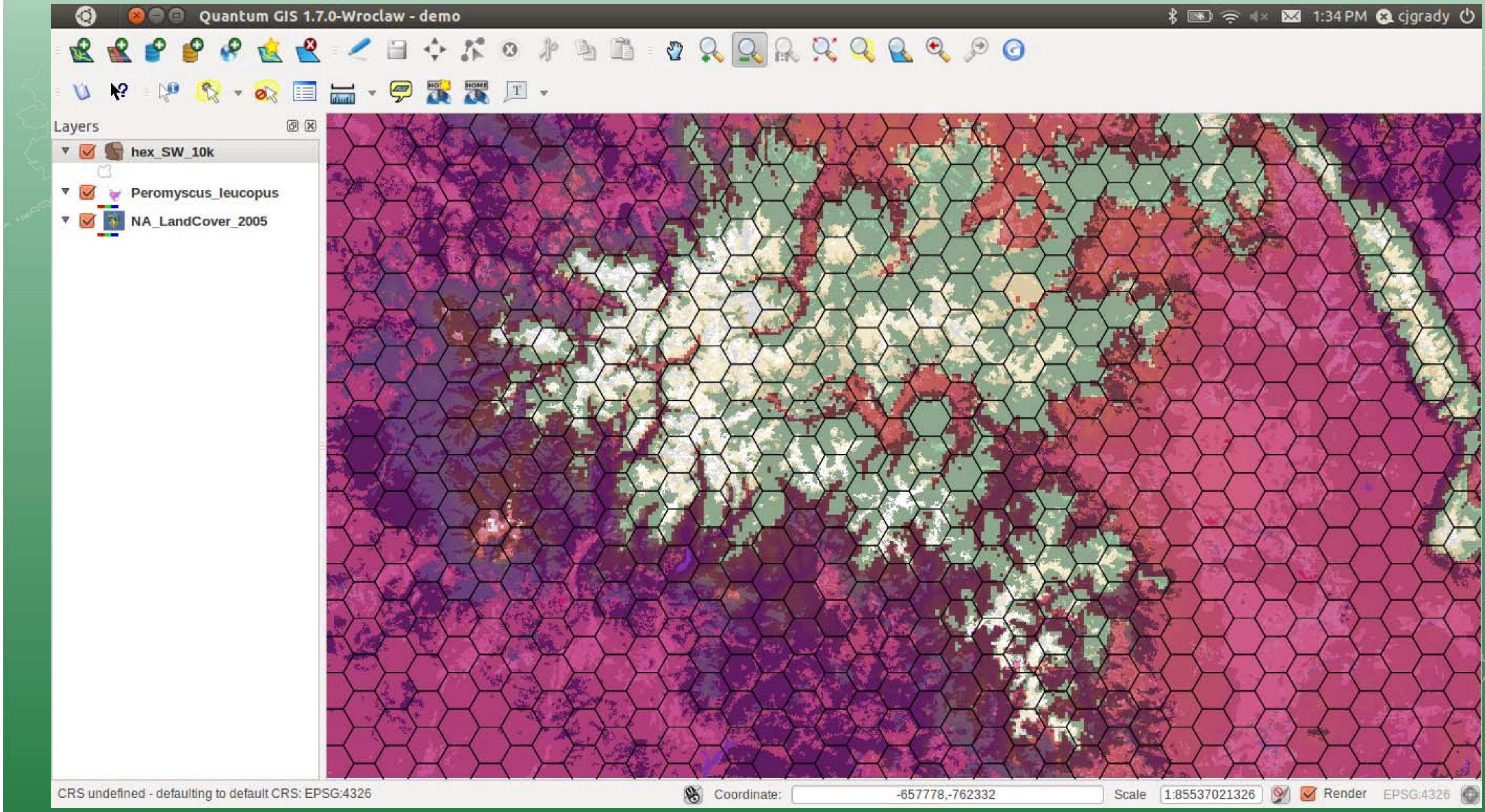
QGIS with Lifemapper MacroEcology plug-in



Coordinate: -4083532,5375726

Scale :4934194123563

Hexagonal Grid



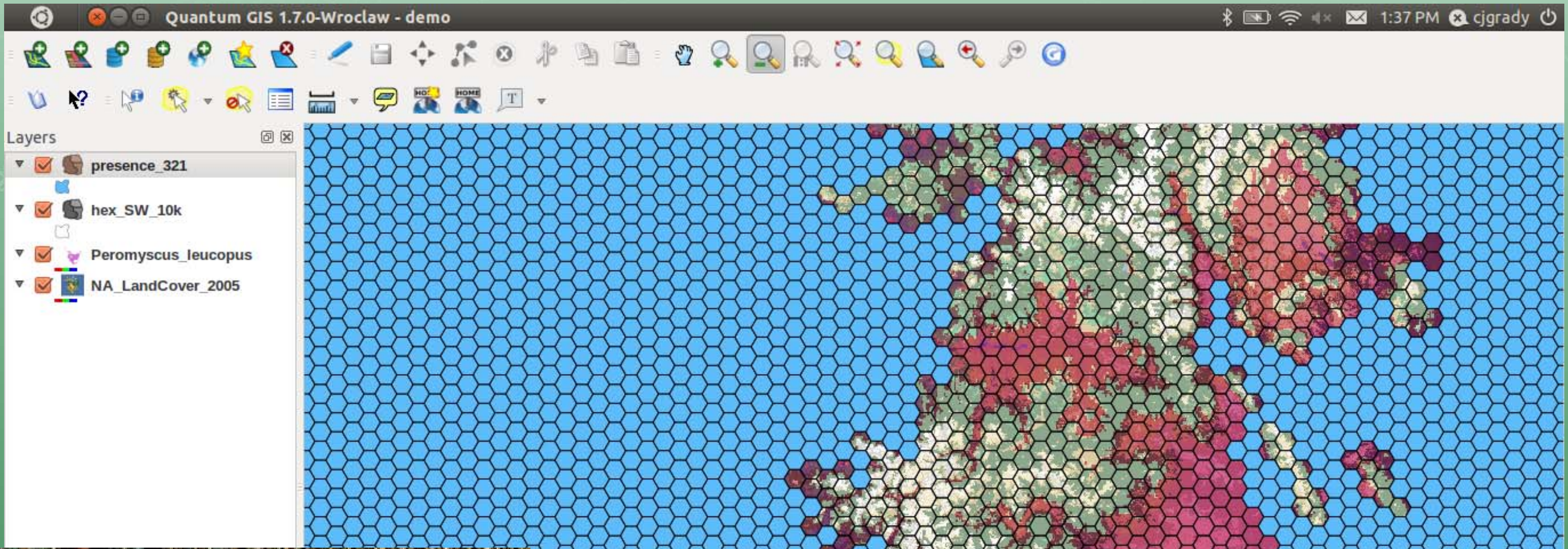
CRS undefined - defaulting to default CRS: EPSG:4326

Coordinate: -657778,-762332

Scale 1:85537021326

EPSG:4326

Mouse ‘Presence’



Quantum GIS 1.7.0-Wroclaw - demo

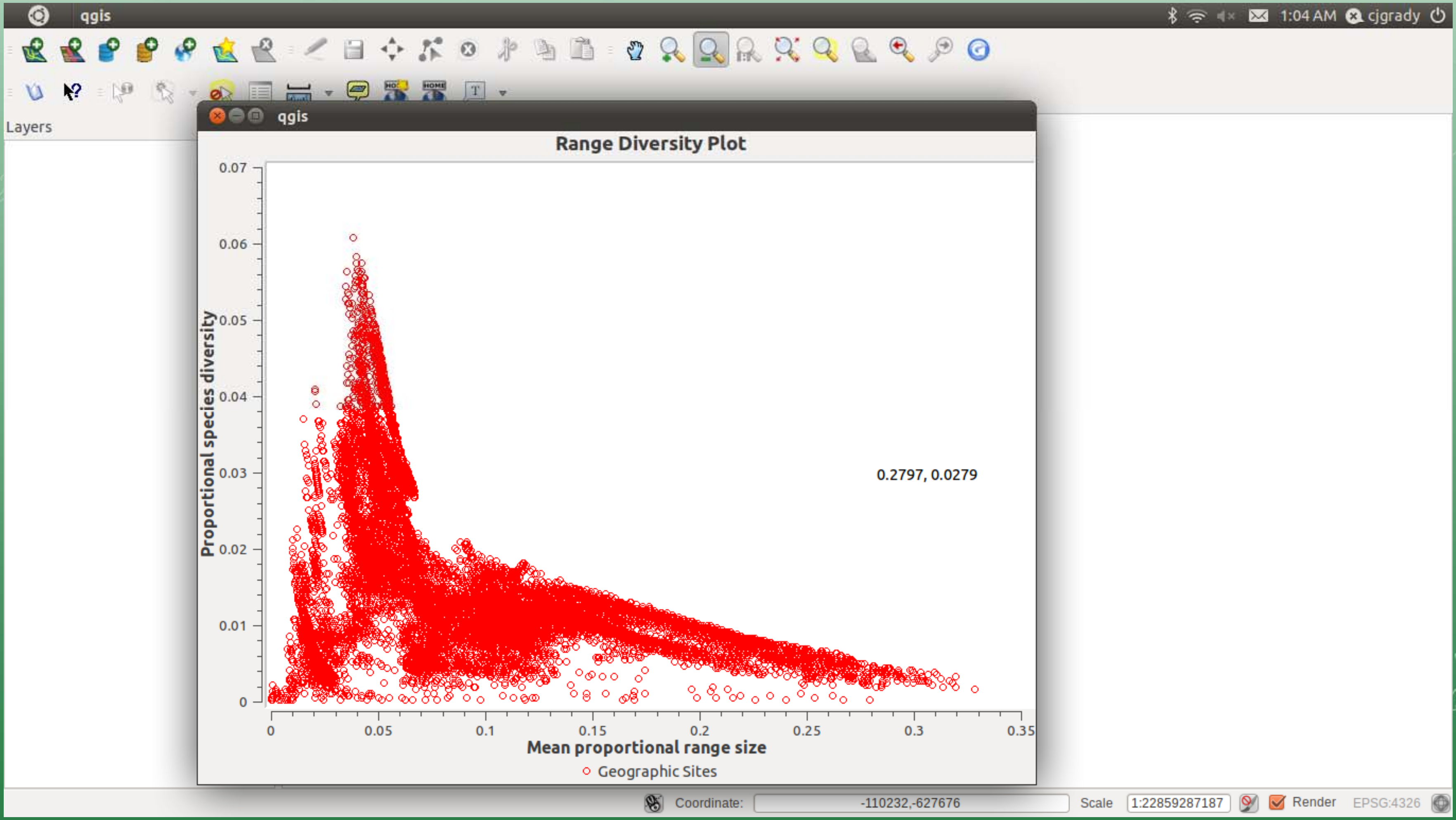
1:37 PM cjgrady

Layers

- ☑ presence_321
- ☑ hex_SW_10k
- ☑ Peromyscus_leucopus
- ☑ NA_LandCover_2005

Coordinate: -642592,-814435 Scale 1:141866887374 ☑ Render EPSG:4326

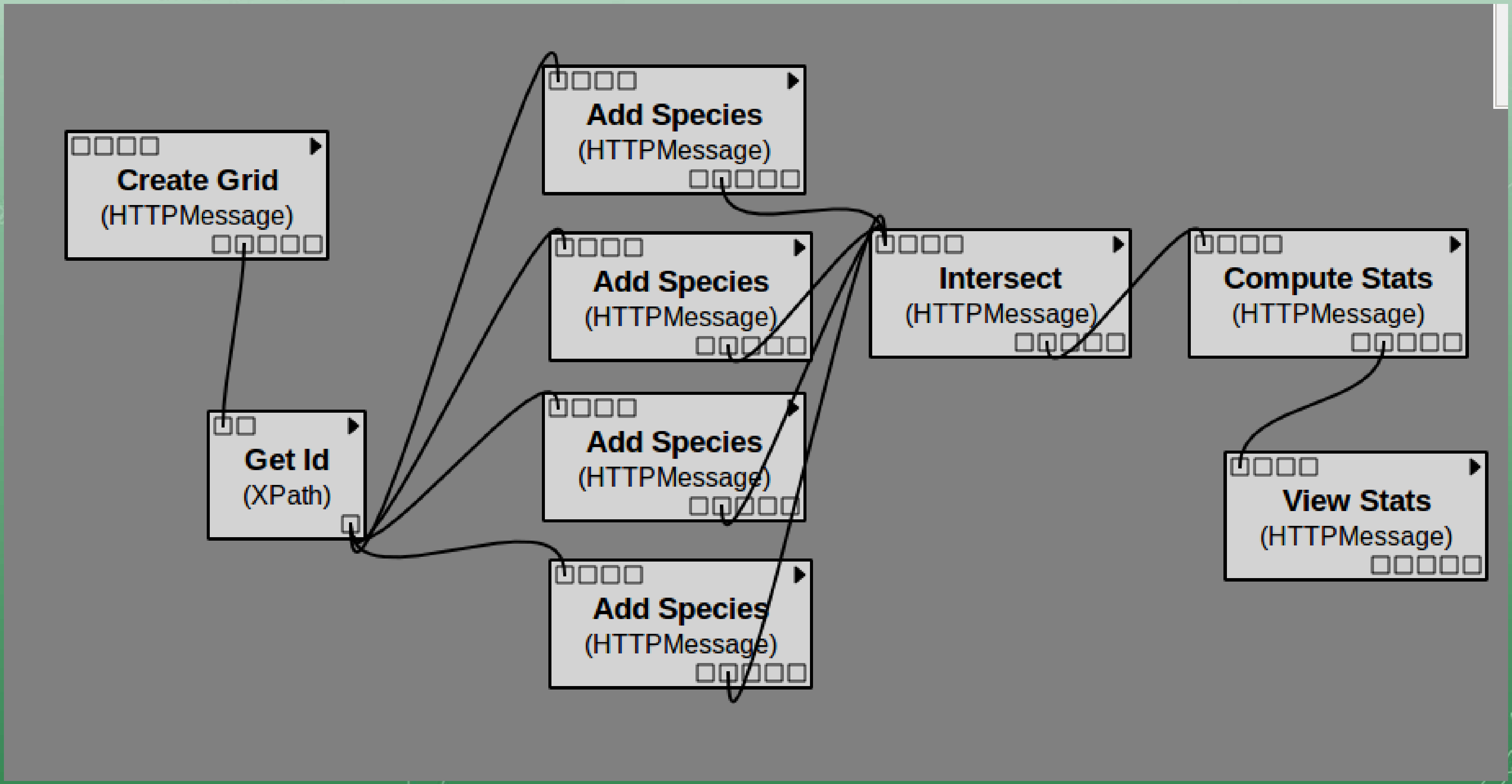
Sites / Species Plot



Coordinate: -110232,-627676

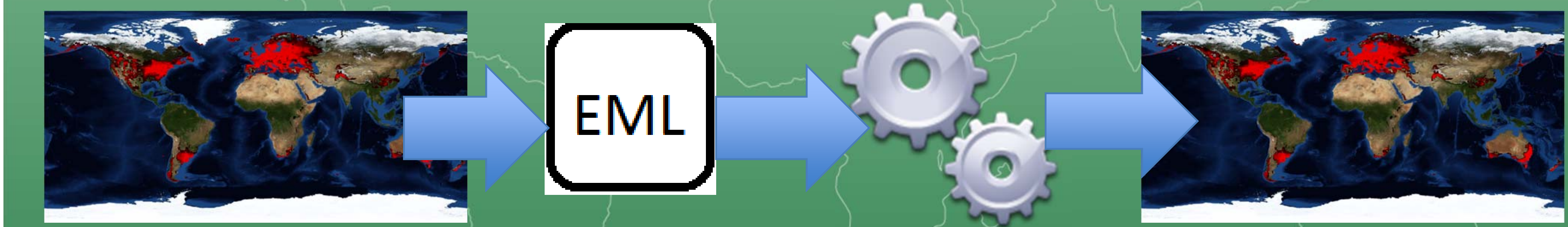
Scale 1:22859287187

VisTrails


Create Grid
(HTTPMessage)
Get Id
(XPath)
Add Species
(HTTPMessage)
Add Species
(HTTPMessage)
Add Species
(HTTPMessage)
Add Species
(HTTPMessage)
Intersect
(HTTPMessage)
Compute Stats
(HTTPMessage)
View Stats
(HTTPMessage)

Reproducibility

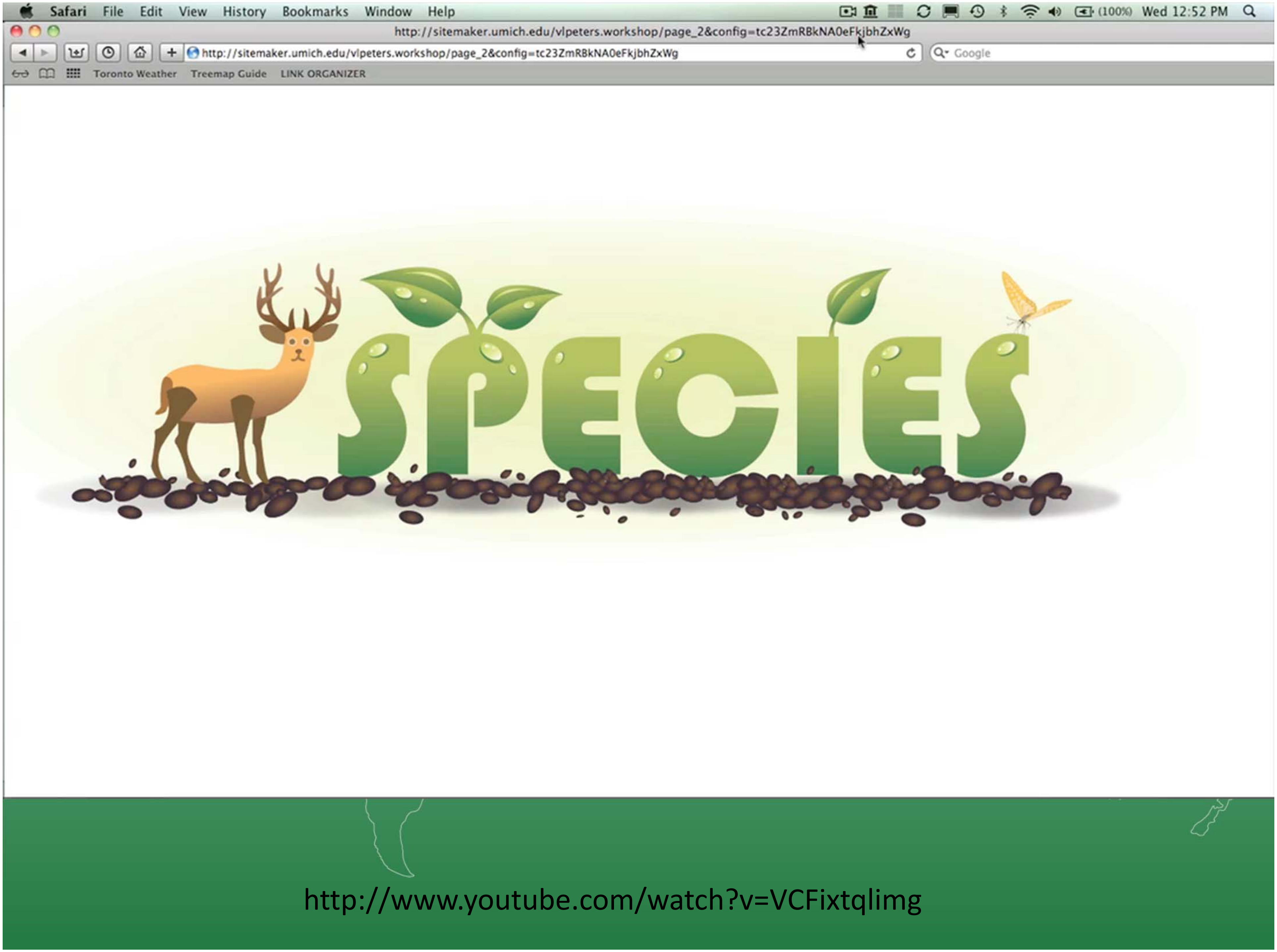
- Simple process metadata
- Process metadata extensions
- Lifemapper client metadata reader



EML

Collaborations

- KU Biodiversity Institute
- NSF Cyber-Commons
- Change Thinking
- CI Team

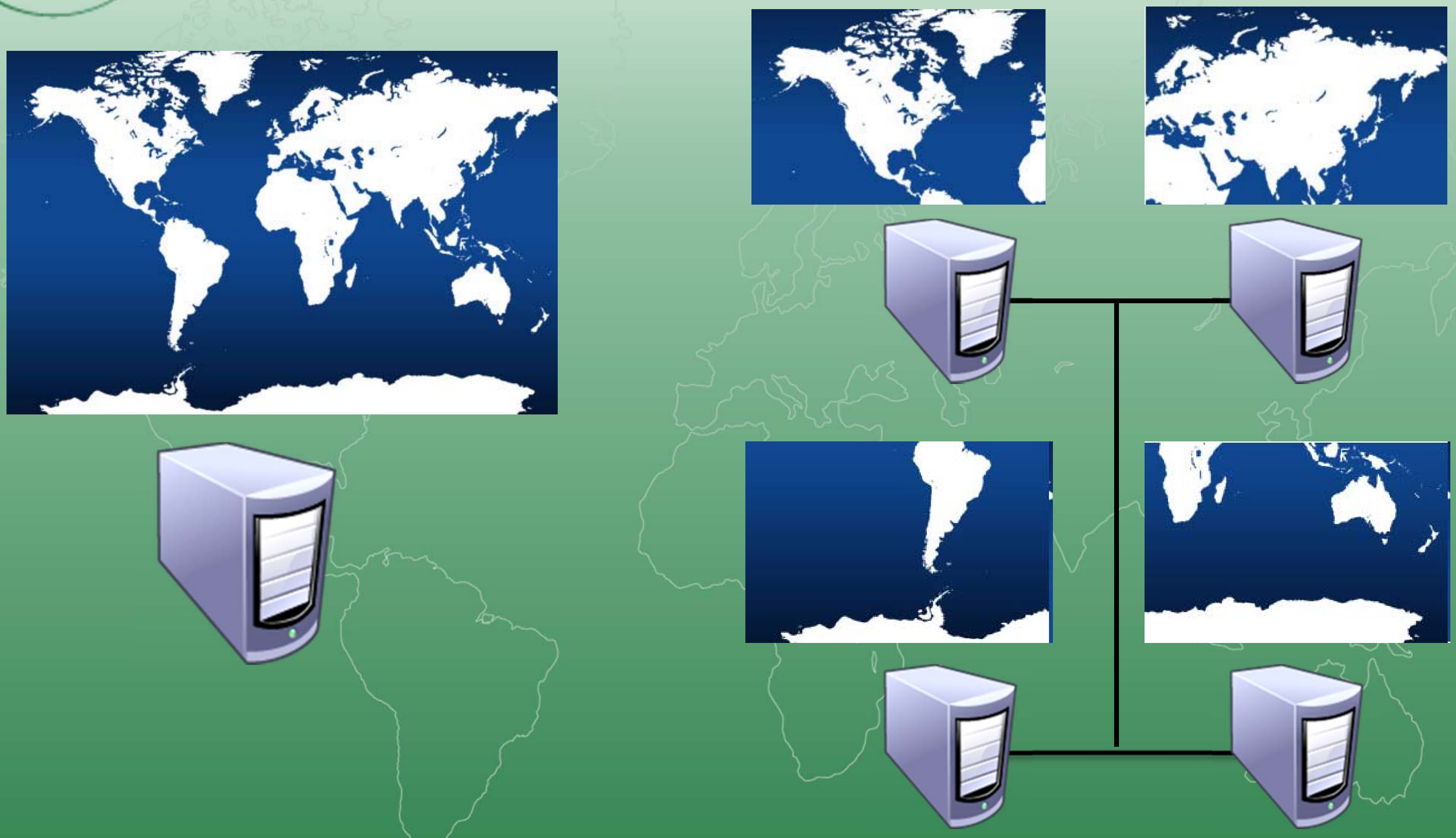

Safari File Edit View History Bookmarks Window Help
(100%) Wed 12:52 PM
http://sitemaker.umich.edu/vlpeters.workshop/page_2&config=tc23ZmRBkNA0eFkjbhZxWg
Google
Toronto Weather
Treemap Guide
LINK ORGANIZER
SPECIES

http://www.youtube.com/watch?v=VCFixtqlimg

Future Directions

- Publish metadata through standard APIs
- Contribute process metadata extensions to community
- Gesture based interface
- Explore extensions into cloud and other grid computing environments

Parallel Processing



Summary

- Provide end-users with clients to assemble and manage biodiversity modeling experiments
- Allow users to harness the computing power available through our cluster to perform computationally intensive tasks
- Include process metadata to document how an experiment was performed

Funding

U.S. National Science Foundation

NSF EPSCoR 0553722

NSF EPSCoR 0919443

EHR/DRL 0918590

BIO/DBI 0851290

OCI/CI-TEAM 0753336

Questions

- cjgrady@ku.edu
- http://lifemapper.org