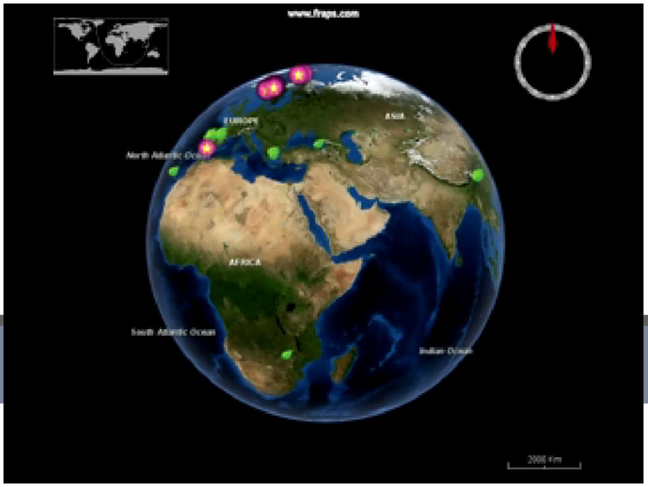
SizeUp: A Tool for Interactive Comparative Collection Analysis for Very Large Species Collections

Andrew Ozor



Wide Ranging Biological Data

- Global repository of species collections
  - Patchwork of specimen collecting programs
  - Diverse research interests
  - Paper documentation
- Online Databases
  - Global Biodiversity Information Facility (GBIF)
  - A global cache of museum data
- Inefficient data access
- How do we compare and analyze large data sets, and visualize the result in a user friendly tool?

PHYCOLOGICAL HERBARIUM
A 67496
UNIVERSITY OF BRITISH COLUMBIA
THE UNIVERSITY OF BRITISH COLUMBIA HERBARIUM C/C
ALGAE OF ALASKA
NAME: Opuntiella denticulatum Lindstrom 1987 (cystocarpic) HOLOTYPE
LOCATION: Kelp Transect Side, Coghlan Island, Alaska (31 ft., on rock)
COLL. S.C.Lindstrom & M.Calvin DATE 13/June/1974
IDENT. S.C.Lindstrom DATE No. ABFL-278

Multiple Problems

- No formal definition for 'quality'
  - Inherently subjective
  - Changes based on domain
- Time consuming computation required to analyze common attributes among large sets of data
- Distance calculation has traditionally been a very time consuming among geospatial points
- User interface to display the spatial distribution of specimen data and provide tools to select relevant comparison criteria

Fast Geospatial Calculation

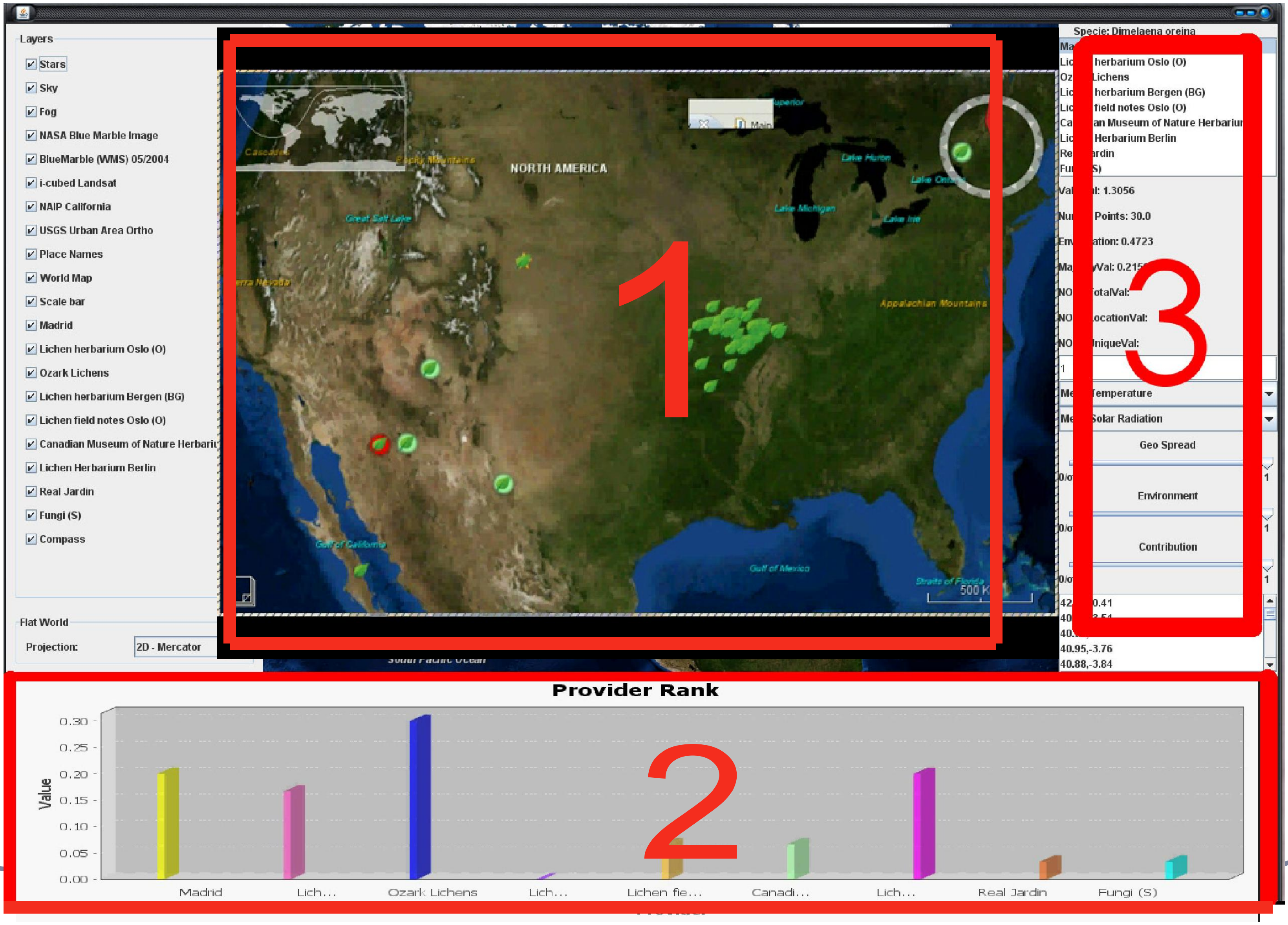
- QuadTrees
  - Create a spatial hierarchy based on the geospatial location
  - *Shown to be desirable when working with geospatial data
  - Hierarchical aggregation alleviates the need for *n* by *n* comparison
  - Efficient with many types of queries
- Branch bypassing
  - Significant reduction of nodes on highly clustered data sets
  - Speeds up computation time
- Approximate distance
- Overcome the problem of slow comparison amongst large quantities of geospatial data

**Samet, et al.* *Processing geographic data with quadtrees*

Value Measure

- Location
  - Analyze the geospatial spread of the specimen localities
- Environment
  - Provide a measure of environmental diversity
  - Uses environmental layers
    - Example: temperature, precipitation, solar radiation, etc…
- Contribution
  - A ratio of unique information a collection contributes
- Applicable to any biological collection
- Ability to include more attributes for specific domains

User Interface



Layers

- ☑ Stars
- ☑ Sky
- ☑ Fog
- ☑ NASA Blue Marble Image
- ☑ BlueMarble (WMS) 05/2004
- ☑ i-cubed Landsat
- ☑ NAIP California
- ☑ USGS Urban Area Ortho
- ☑ Place Names
- ☑ World Map
- ☑ Scale bar
- ☑ Madrid
- ☑ Lichen herbarium Oslo (O)
- ☑ Ozark Lichens
- ☑ Lichen herbarium Bergen (BG)
- ☑ Lichen field notes Oslo (O)
- ☑ Canadian Museum of Nature Herbariu
- ☑ Lichen Herbarium Berlin
- ☑ Real Jardin
- ☑ Fungi (S)
- ☑ Compass

Flat World

Projection: 2D - Mercator

Specie: Dimelaena oreina

Ma
Lic herbarium Oslo (O)
Oz Lichens
Lic herbarium Bergen (BG)
Lic field notes Oslo (O)
Ca an Museum of Nature Herbariu
Lic Herbarium Berlin
Re ardin
Fu S)

Va l: 1.3056

Nu Points: 30.0

En ation: 0.4723

Ma Val: 0.215

NO otalVal:

NO ocationVal:

NO niqueVal:

1

Me Temperature

Me Solar Radiation

42 0.41
40
40.
40.95,-3.76
40.88,-3.84

Evaluation

- Five test subjects from the University of Kansas Biodiversity Institute.
- Results
  - Subjects understood how and why collections were ranked
  - Subjects foresaw many uses for comparative collection analysis
  - Subject mentioned no “input lag” or “slow response” from the application

Conclusion

- Collaboration
- More easily find research resources
- Aide in the evaluation of:
  - Staff
  - Building resources for biological collection repositories
  - Collection roadmaps
- Help users assess the quality of their data
- Incentive for museums to make their data available online
- Applicable to any geo-referenced data set