Biodiversity Data Exchange Using PRAGMA Cloud

U. Pavalanathan and B. Plale
Indiana Univ

A. Stewart and C.J. Grady
Univ of Kansas

R. Beaman
Univ of Florida

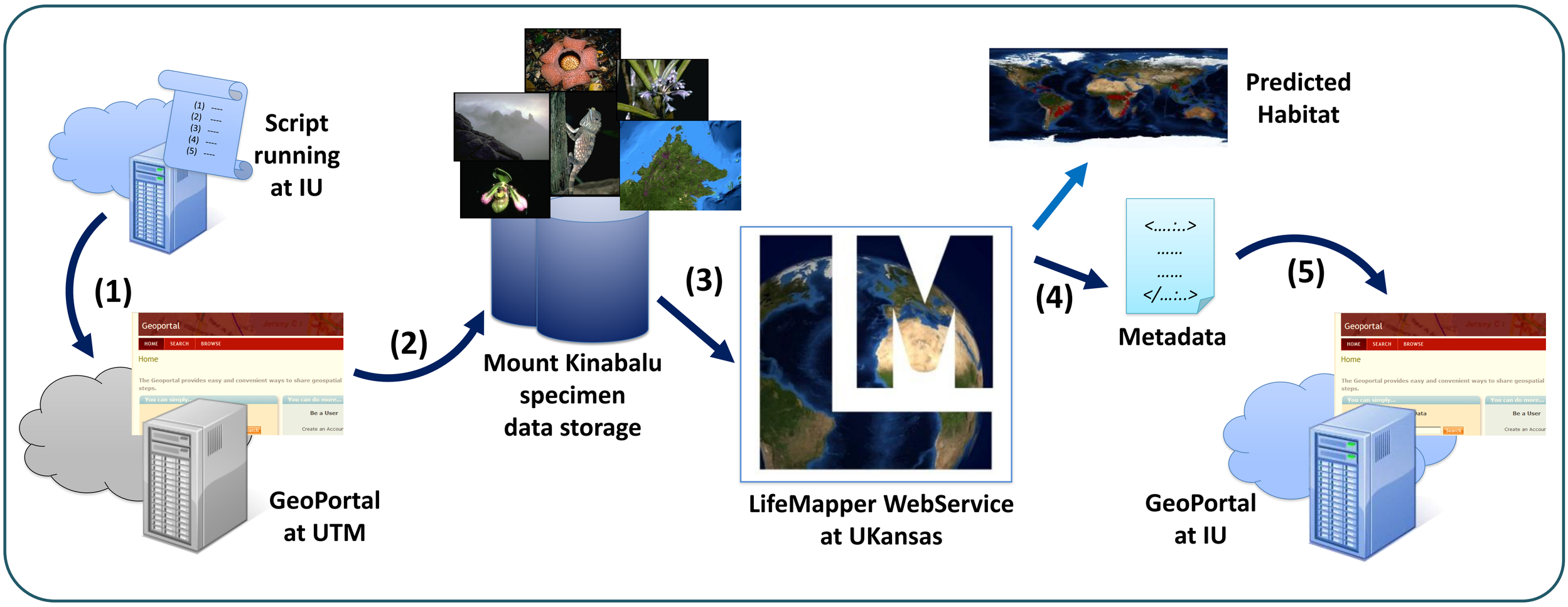
S. Shamsir, S.N. Azmy and C.T. Han
Univ of Tech, Malaysia

Introduction

- Biodiversity applications are data driven by nature. Distribution patterns can be revealed through analysis of large volumes of species occurrence data using techniques such as species distribution modeling. Analysis tools, data discovery methods, and cloud computing all contribute to the solution.
- Lifemapper – a biogeographical analysis workbench that brings together open source geospatial and biodiversity informatics tools.
- ESRI GeoPortal – open source metadata catalog that enables discovery and use of geospatial resources including datasets, raster data and Web services.
- Computing resources in the cloud are increasingly useful for the processing of complex ecological modeling algorithms. Data resources in the cloud enable provenance capture and sharing of produced data outputs.

Experiment

- **Experiment:**
  Proof of concept biodiversity application utilizing distributed data and doing useful in the PRAGMA cloud.
- **Biodiversity analysis :**
  Basic application of species distribution modeling using Lifemapper
- **Data:**
  Specimen collection records compiled from herbarium resources worldwide and maintained at the University of Florida. Records illustrate plant diversity on Mount Kinabalu, notable for its high diversity and endemism of species and ultramafic environments
- **Metadata:**
  Metadata files describing nine species distribution data sets are uploaded to a GeoPortal server running at Universiti Teknologi Malaysia (UTM).



Workflow

- **Step 1**:
  A script running at the Indiana University (IU) PRAGMA node queries the GeoPortal server running at UTM. Retrieves metadata describing raster datasets of Mount Kinabalu, environmental and satellite data, and point occurrences of specimen collections
- **Step 2**:
  By analyzing the resulting metadata records, the script downloads the datasets stored in a Dropbox location
- **Step 3**:
  Script posts the data sets as occurrence sets to the Lifemapper instance running at University of Kansas. Using the returned occurrence set identifiers, the script submits post experiment requests to Lifemapper
- **Step 4**:
  Script retrieves metadata results returned for each Lifemapper experiment
- **Step 5**:
  Script catalogs provenance metadata describing each experiment, along with the URL for Lifemapper prediction outputs, in a GeoPortal server instance running at IU

Open Problems

- **PRAGMA Cloud Security**:
  Data are sensitive in that they reveal ecologically sensitive information . What are the cloud security measures to be taken for controlled access of sensitive data?
- **Agreements on Core Metadata**:
  Discovery and reuse of scientific outcomes from these applications depend on automated or manual extraction of rich metadata about the datasets and prediction outputs. For this to happen, some agreement must exist on core metadata.
- **Provenance and Metadata Capture:**
  For automated metadata extraction, proper tools must be in place at point of creation (e.g., extracting metadata automatically from the result set from Lifemapper.) This can be accomplished through provenance capture
- **Ownership of Results:**
  When analysis is carried out on PRAGMA cloud, the resulting dataset can contribute to enriching the data of the cloud. How is ownership and sharing tracked?
- **Metadata Catalog Federation:**
  We demonstrated use of two GeoPortal instances. What is the PRAGMA-wide solution for metadata catalog federation?