Scalable and Provenance-Enabled Scientific Workflows for Species Distribution Modeling

Luiz Gadelha\textsuperscript{1}, Guilherme Gall\textsuperscript{1}, Andrea Sanchez-Tapia\textsuperscript{2}, Marinez Ferreira de Siqueira\textsuperscript{2}, Jorge Velásquez\textsuperscript{3}, Daniel Lopez\textsuperscript{3}

LNCC\textsuperscript{1} (Brazil), JBRJ\textsuperscript{2} (Brazil), IAvH\textsuperscript{3} (Colombia)

TDWG 2014 Annual Conference, Jönköping, Sweden
October 28, 2014
In addition to: empirical, analytical, computational (simulation).

Enabled by simulations and scientific sensors that generate large amounts of data.

The fourth paradigm is based on the analysis and exploration of this data. Commonly called e-Science.

Life Cycle of Computational Scientific Experiments

Requirements of Computational Scientific Experiments

- **Abstraction.** High level of abstraction, avoiding details of computational environment.

- **Reproducibility.** Third-party verification of the experiment.
  

- **Reusability.** Using specifications of previously executed computational scientific experiments.
  
  Example: MyExperiment (http://www.myexperiment.org).

- **Scalability.** Resilience to increased number of computational tasks and amount of data.
A **scientific workflow** is given by the specification of a set of activities and their data dependencies.

A **scientific workflow management system** (SWMS) supports the management of this life-cycle.

Provenance describes the history of conception (prospective) and execution (retrospective) of a computational scientific experiment.

- consumption and production relationships between activities and artifacts;
- agents that control these activities;

---


The Biomodelos portal, from the IAvH in Colombia, allow for validation species distributions by taxonomists.

ENM specialists publish the distributions produced with DISMO along with metadata.

Taxonomists validate, edit, and annotate the distributions published in the portal.

With support from the GBIF Mentoring Program (2013), SiBBr and IAvH are collaborating on reusing and scaling Biomodelos.

http://biomodelos.humboldt.org.co.
Source: Biomodelos (http://biomodelos.humboldt.org.co)
Proposal for SDM portal in SiBBr
Swift

- Focus on scalable execution on parallel and distributed environments.
- Proven scalability to 116,000 CPUs.
- Given by:
  - High-level scripting language with implicit parallelism.
  - Execution engine supporting:
    - local execution,
    - parallel execution (e.g. PBS, SGE),
    - distributed execution (e.g. Condor, Globus).
- Provenance-enabled.

Swift: A language for distributed parallel scripting. 

Parallel Scientific Workflow for ENM

- Implemented in Swift, a scalable scientific workflow management system.
- Workflow activities given by pre-processing scripts and ENM algorithms (from the openModeller library).
- Parallelism opportunities explored:
  - per species,
  - per climatological scenario,
  - per ENM algorithm (Maxent, SVM, ...).
- Support for capturing provenance.
- Github repository:
  - https://github.com/sibbr/sdm-workflows
Parallel Scientific Workflow for ENM: Performance

Performance of parallel scientific workflow for ENM with 24 (left) and 900 (right) species.
Concluding Remarks and Future Work

- Scientific workflow showed good scalability on parallel computers.
- Provenance tool allows for querying workflow, performance, and domain-specific information about the workflow execution.
- Future work:
  - Inclusion of R scripts (from IAvH and JBRJ) based on DISMO with Snowfall, Rmpi in the scientific workflow.
    - https://github.com/AndreaSanchezTapia/sdm_dismo
    - https://github.com/LBAB-Humboldt/parallelMaxent
  - Simple web portal for running the scientific the workflow using tools from SINAPAD.
  - Improved web portal that allows for ENM specialists and taxonomists to interact.

Acknowledgements
Thank you!

lgadelha@lncc.br