INTRODUCTION
The Lifemapper project on the PRAGMA cloud addresses the challenge of understanding sensitive ecological systems and forecasting their responses to change. Lifemapper expands its computing resources onto virtual machines running at San Diego Supercomputer Center (SDSC).

The CS team at SDSC is further creating a virtual private network that enables a trust network to be established. This rich interdisciplinary application gives us the opportunity to study automated provenance capture as the biodiversity analysis is carried out and data products moved, consumed, and generated.

MOTIVATION
With the purpose to collect usable and reliable provenance information, we address the following research questions related to provenance capture and representation:

A. **Data lineage**: can our provenance approach properly capture complete mappings between output data, input data and algorithms used to enable data reuse?
B. **Data quality**: can our provenance approach capture provenance sufficient to check input biodiversity dataset quality to verify the reliability of output data?
C. **Version control**: can our provenance approach capture algorithm or program versioning?
D. **Failure trace**: Can our provenance approach determine whether output dataset is affected by some infrastructure problems such as node failure, etc.?

These challenges motivate our proposed development of the Provenance Collection Framework for Lifemapper (LMPF) based on Karma Provenance Collection Tool.

FRAMEWORK

PROVENANCE COLLECTION

- **Logical provenance** consists of the input data items and executable program or function that creates an output data item inside one job
- **Infrastructure provenance** is information about when a job is executed and what parts of the infrastructure were involved in the execution of the job

FUTURE WORK

- Extending the Karma adaptor for Lifemapper to perform more system-based gathering of provenance by improving the provenance granularity level
- Migrating from Open Provenance Model (OPM) to the W3C PROV data model for provenance representation. PROV allows richer expression of relationships, semantic annotations and semantic inferencing

ACKNOWLEDGEMENTS
My advisor and I thank other project members: Yuan Luo of IU, Aimee Marian Stewart and C J Grady of University of Kansas; Nadya Williams and Philip Papadopoulos of SDSC.