PROVENANCE Contract Number:
Enabling and Supporting Provenance in Grids for Complex Problems 511085

Title:	BioDiversity Informatics
Authors:	Ian Wootten, Shrija Rajbhandari and Omer Rana
Editor:	
Reviewers:	
Type:	
Version:	1.0
Date:	September 2006
Status:	Final
Class:	Public

Summary

This document provides a case study of Provenance use in the BioDiversity Informatics. The document is based on work undertaken in the BDWorld project.

The *PROVENANCE* Project receives research funding from the European Community's Sixth Framework Programme

Members of the PROVENANCE Consortium

IBM United Kingdom Limited	United Kingdom
University of Southampton	United Kingdom
University of Wales, Cardiff	United Kingdom
Deutsches Zentrum für Luft- und Raumfahrt e.V.	Germany
Universitat Politecnica de Catalunya	Spain
Magyar Tudomanyos Akademia Szamitastechnikai es Automatizalasi Kutato Intezet	Hungary

Contents

The *PROVENANCE* Project receives research funding from the European Community's Sixth Framework Programme

Provenance Use Case:
BioDiversity Informatics

Ian Wootten, Shrija Rajbhandari and Omer Rana
o.f.rana@cs.cardiff.ac.uk
Cardiff School of Computer Science/Welsh eScience Centre

1 Introduction

This document describes a particular scenario developed as part of the BDWORLD project and provenance issues that are related to issues being considered in the project.

2 Problem Description and Provenance Questions

The BDWORLD project is investigating the impact of climate change on the distribution of particular species across the world. The project is aimed at running various "what-if" scenarios to investigate how: (i) change in climate is likely to impact a given species of plants (primarily) and animals; and (ii) which species are likely to be under threat as a result of a rise/fall in temperatures or increase/decrease in rainfall. A computational model of the distribution of species has been developed in the project, and integrated with various global databases that contain details about particular traits of these species.

The main BioDiversityWorld scenario (bioclimatic modelling) shown in fig. 1 begins with the generation of a taxonomy for the particular species' of interest. This is then queried against the Global Biodiversity Information Facility (GBIF) to obtain the locality information for the species. In parallel with this, climate layers containing estimations of such attributes as temperature and rainfall are obtained and selected to produce a 'climate envelope'. This is then used with a specific selected Open Modeller (OM) algorithm by interpolating the climatic data at the points of locality of specimens producing a bioclimatic model. This model is then able to be projected upon a map of the world in order to make predictions of the anticipated effects of climate change upon biodiversity.

2.1 Use Case 1: Workflow and Result Accuracy

A bioinformatitian runs the bioclimatic modelling experiment presented in fig 1. Later, another bioinformatition B wants to use the result of this experiment to do comparative studies. B determines whether the resultant projection image is one which is accurate and can be relied upon.

In this case, B could simply view the order in which the experiment was carried out and make a judgement on the process to place a degree of trust on the result. It can be speculated that after viewing the services/algorithms

The *PROVENANCE* Project receives research funding from the European Community's Sixth Framework Programme

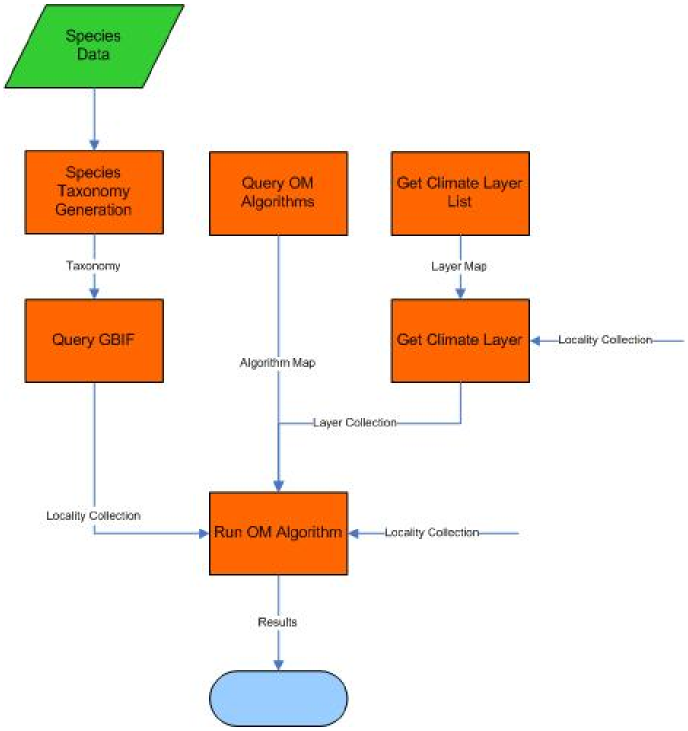

Figure 1: Bioclimatic Modelling within BDWorld

involved and the way the process was constructed, B can also ascertain a certain level of result accuracy.

Thus, given that each actor state provenance data contains the accuracy to which an actor produces a given result, B is able to determine what the overall accuracy of the projection image shall be depending on how the bioclimatic process is constructed. In a similar manner B is also able to calculate the overall reliability that can be placed in the process which produced a particular projection image through inspection of all the involved actors reliability records.

2.2 Use Case 2: Execution Bottleneck

A bioinformatition B, downloads some locality information for a particular species from the GBIF database and runs the bioclimatic modelling experiment upon it. A later run of the same process yields an overall execution time which is far greater than the earlier run. B determines which of the processes involved caused the increase in execution time in this experiment.

Through inspection of the execution times stored within actor provenance, B is able to determine which service/s caused the increased time for this particular process run. Thus, any major increase in the total execution time would make B conclude that the service(s) that are causing this have shortcomings and could not be trusted – as the data might have been corrupted while processing, and in effect leading to a corruption of the resultant projection image.

2.3 Use Case 3: Input Parameter Requirement

A bioinformatition runs the BDW experiment presented in fig 1. Later a reviewer analysing the workflow determines that the climate data that was used includes attributes such as temperature and rainfall. Based on this information, the reviewer could conclude that in order for her to trust the result (so as to serve a particular purpose); she requires humidity data also to be used in the experiment.

Although the reviewer places a lesser degree of trust on the input data, the process as a whole could still be meaningful to place some trust on its result. This would hence effect the overall trust on the result of the process.

2.4 Use Case 4: Data Consistency

A bioinformatition runs the BDW experiment explained in fig 1. A reviewer wants to confirm whether the data that is passed between the services are consistent in terms of their type and value – for example whether the locality data output from the GBIF query matches the input data received by the OM Algorithm within the workflow.

The examination of provenance data i.e., the I/Os of each actor for that particular process run, the reviewer is able to determine if a data an actor has generated is the same as the data received by another actor during their interaction. This assumes that each actor involved in the workflow provides provenance data that includes I/O demonstrating interaction.

2.5 Use Case 5: Data Schema Completeness

A bioinformatition runs the experiment explained in fig 1. Later a reviewer determines if all the data instances generated or consumed by the actors involved in this process run are complete in terms of their current (updated) schemas so that the result can be relied upon.

Investigating the instance of the data generated during a workflow enactment, and comparing this with its current schema, the reviewer can check the presence (or absence) of all the elements in the schema instance. This assume that there are predefined schemas for inputs and outputs for each node (actors) within the workflow. Note that the schema can change over time to reflect any changes or updates in the algorithms/actors which consume and generate data. Thus, such validation is crucial to place an overall degree of trust on the result. Some mechanism is also needed to indicate that a schema may have changed between multiple workflow enactments.

2.6 Use Case 6: Data Updates

In the BDWorld experiment shown in fig 1, a "Grid Bioinformatics Interoperation Facility" (GBIF) database is queried for a particular species to get the locality information. The GBIF database enables integration of schemas across multiple databases that contain specifies data involved in the experiment. Over time the data used for the experiment might be updated or corrected in the database. This updating will result in making new corrected and/or updated data available for use. In such a case, in order for a reviewer to trust the locality data used for their experiment, there is a requirement that the update frequency of the data source (GBIF database) exists.

By investigating the update frequency information of the GBIF database in the provenance record of this particular process run, the reviewer can place a degree of trust on the locality data used. The trust on the locality data would vary depending on when the workflow is reviewed and the frequency of updates. For example, if the workflow was reviewed following a recent update, locality data used in the workflow would not be trusted.

3 Conclusion

These use cases demonstrate various ways in which provenance data could be used to elicit trust within a workflow enactment. Various scenarios in the context of the BDWorld project have been outlined. Additional details of the BDWorld project can be found in [1].

References

[1] A. Jones, R. White, N. Pittas, W. Gray, T. Sutton, X. Xu, O. Bromley, N. Caithness, F. Bisby, N. Fiddian, M. Scoble, A. Culham and P.Williams, “BiodiversityWorld: An architecture for an extensible virtual laboratory for analysing biodiversity patterns”, UK e-Science All Hands Meeting, pp 759–765, Nottingham, UK, September 2003.

The *PROVENANCE* Project receives research funding from the European Community's Sixth Framework Programme