Scientific Programming 14 (2006) 195–208
IOS Press

Supporting the construction of workflows for biodiversity problem-solving accessing secure, distributed resources

J.S. Pahwa[a,*], A.C. Jones[a], R.J. White[a], M. Burgess[a], W.A. Gray[a], N.J. Fiddian[a], R.O. Smith[a], A.R. Hardisty[a], T. Sutton[b], P. Brewer[b], C. Yesson[b], N. Caithness[b], A. Culham[b], F.A. Bisby[b], M. Scoble[c], P. Williams[c] and S. Bhagwat[c]
[a]*Cardiff School of Computer Science, Cardiff University, Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA, UK*
[b]*School of Plant Sciences, The University of Reading, Berks RG6 6AS, UK*
[c]*The Natural History Museum, Cromwell Road, London SW7 5BD, UK*

Abstract. In the Biodiversity World (BDW) project we have created a flexible and extensible Web Services-based Grid environment for biodiversity researchers to solve problems in biodiversity and analyse biodiversity patterns. In this environment, heterogeneous and globally distributed biodiversity-related resources such as data sets and analytical tools are made available to be accessed and assembled by users into workflows to perform complex scientific experiments. One such experiment is bioclimatic modelling of the geographical distribution of individual species using climate variables in order to explain past and future climate-related changes in species distribution. Data sources and analytical tools required for such analysis of species distribution are widely dispersed, available on heterogeneous platforms, present data in different formats and lack inherent interoperability. The present BDW system brings all these disparate units together so that the user can combine tools with little thought as to their original availability, data formats and interoperability. The new prototype BDW system architecture not only brings together heterogeneous resources but also enables utilisation of computational resources and provides a secure access to BDW resources via a federated security model. We describe features of the new BDW system and its security model which enable user authentication from a workflow application as part of workflow execution.

Keywords: Biodiversity, phylogeny, biogeography, climate change, workflow

1. Introduction

Many individual scientists and institutions working with biodiversity data create their own data and analysis resources often for a narrow range of uses. Interoperability among such resources can sometimes be challenging, as they were not originally designed to be used together as part of a larger system and often do not conform to any recognised standard. Similarly, analytical tools to perform specific tasks have often been developed as stand-alone executables operating on data in a defined format so that if facilities from more than one tool are required, a user will frequently have to resort to transferring and converting data between tools manually. There is a lack of automated means of correlating biodiversity and ecosystem data from various sources for its analysis and use in models and statistical tools to derive useful biological results [33].

For example, suppose a scientist wishes to investigate where a particular species of plant or animal might be expected to occur, given estimated past or predicted future climatic conditions. To investigate this bioclimatic modelling problem requires access to species distribution data, to tools that can model the climate char-

*Corresponding author. Tel.: +44 (0)29 2087 4812; Fax: +44 (0)29 2087 4598; E-mail: J.S.Pahwa@cs.cardiff.ac.uk.

ISSN 1058-9244/06/$17.00

acterising the locations where the species is to be found, to data pertaining to the climate at the time of interest, and to map images on to which the predicted distribution can be projected for viewing. Many tools routinely used by scientists to perform such experiments are legacy tools which are not interoperable.

In addition to the biodiversity resources and tools the scientist also requires access to computational resources on which to conduct biodiversity experiments. The duration of an experiment depends upon its nature, number of variables involved and the amount of biodiversity resources required. For example, performing species distribution experiments for a large number of species under different climatic scenarios and using different modelling algorithms is complex and requires access to high-performance or clustered computational resources in order to achieve an acceptably short experiment duration. The challenge with providing access to computational resources is that these computational environments are not readily available and therefore they need to be designed and implemented based on available computational nodes, biodiversity resources, researchers' needs and using middleware such as Condor [44]. We require an environment which not only brings together heterogeneous biodiversity resources and analytical tools but also meets the computational requirements of biodiversity experiments in order to meet increasing user and workflow requirements. It is also important to address secure access to analytical tools and data sources available in the BDW environment for performing complex biodiversity experiments. Resources such as data sources are owned and managed by particular organisations or communities of researchers, and some resources contain a large amount of intellectual property to which their owners wish to regulate access. Results derived from biodiversity experiments can also contain sensitive information. For example, a biodiversity experiment may use species data pertaining to a species that is endangered or under a specialist conservation program and project its distribution across the world on to a map. It is important to provide restricted access to raw locality data in order to avoid undesirable or illegal exploitation of endangered species. Therefore providing a secure access to biodiversity data for its use by authorised users is of significant importance.

The BDW system provides a framework for biodiversity problem solving by providing access to widely dispersed and disparate data sources and analytical tools as part of a Problem Solving Environment (PSE). BDW is a biology-driven project and is being actively used for biodiversity informatics research in three main exemplar study areas: (a) biodiversity richness analysis and conservation evaluation, (b) bioclimatic modelling and global climate change, and (c) phylogenetic analysis and biogeography. In all three of these research areas existing data sources and analytical tools are widely dispersed, available on different platforms and present data in different formats. The system brings together these data sources and analytical tools; provides scientists with tools which allow ready access to resources originally designed for use in isolation and gives them the ability to compose these resources into complex workflows. This is an improvement on the manual process of performing individual activities of a biodiversity experiment using individual systems which are not linked. The system enables chaining of data processing operations and provides flexibility both in the choice of the kinds of resources to be used and in the sequence of operations to be performed. The system is extensible so that new resources and tools can be added to it.

The new BDW architecture which is presently being developed and tested extends the developed earlier in the BDW project, architecture, in order to allow us to experiment with ways to allow the user who wants the ability to control the execution of a workflow with a greater degree of flexibility. In this paper we describe both the current architecture and the new architecture and highlight the areas where significant improvements have been made. For example, as part of the new architecture we demonstrate the application of (i) a new resource invocation model which enables utilisation of computational resources for distributing workload across available nodes, (ii) the Ganglia [10] cluster monitoring system and middleware from the Condor project to meet the computational needs of workflow tasks, and (iii) a new security model implemented in the BDWShib module which uses the capabilities of the Shibboleth [19] security framework - a federated Single Sign-On (SSO) and attribute exchange framework which provides secure access to resources accessible via the web.

The paper is organised as follows. Section 2 provides information on three exemplar study areas. Section 3 describes the architecture of the current BDW system. Section 4 presents a bioclimatic modelling workflow enacted using the current BDW PSE for modelling the distribution of species of bean plant family, Fabaceae (Leguminosae). In section 5 we describe the new BDW system architecture. Section 6 describes the architecture of the BDWShib module for providing secure access to resources available via the BDW PSE. Sec-

tion 7 briefly describes some of the existing Grid and computational environments and projects in the area of biodiversity, bioinformatics and biology which provide access to diverse resources and tools in their environments. Section 8 presents conclusions and further work.

2. The three example scenarios

The BDW system is being actively used for biodiversity research in three main exemplar study areas. The three areas not only fall within the team's area of expertise but are also representative of types of experiments which are performed in the area of biodiversity informatics.

2.1. Biodiversity richness analysis and conservation evaluation

A key issue in biodiversity science and one that also contributes directly to international conservation policy is the analysis of biodiversity richness patterns for a particular taxon (typically a group of related species) around the world. When performing biodiversity richness analysis using the BDW system, the first step for an investigator using the system is to have a name for the target taxon. Because of instabilities in the nomenclature of species, databases containing relevant biological data may index data associated with that taxon using a different name from the one supplied by the investigator. Taxonomic verification, often the first step in any analysis using the BDW system, involves the retrieval of an authoritative list of taxa and their associated synonyms. In our case this is obtained from the Species 2000 (www.sp2000.org) project's [29] 'Catalogue of Life' so that records indexed under either the accepted name or its synonyms can be retrieved and amalgamated. It is recognised that this is only a partial solution to the taxonomic verification problem: others (e.g. Kennedy et al. [30]) have drawn attention to the ambiguity of scientific names and the need for disambiguation with context and definition information, but this information is often not available for existing data organised by species. If it is, and appropriate tools to handle such more fully-specified concepts are available, then these could be incorporated into the BDW system.

In the second stage of the workflow, a distribution data set of specimens or observations belonging to the target taxon is composed from a variety of sources around the world. The final stage is for the distribution data set to be input to the WorldMap system [45], a specialist biodiversity analysis package designed to assist in selecting priority areas for biodiversity conservation. The WorldMap system uses species distribution data to compute a wide range of diversity measures, which it displays on a species richness map, that can then be used for further analyses. The simplest form of analysis is to identify areas of high species richness. As part of research in this area we are currently using species of butterflies from a database of Canadian butterflies provided by the Canadian Biodiversity Information Facility (CBIF) [2]. Access to the database is provided in the BDW system via a resource wrapper.

2.2. Bioclimatic modelling and global climate change

The subject of global climate change and its effect on the biological world is of great importance. Although significant progress has been made in recent years towards understanding how man is affecting the world's climate system, much less is known about how these changes are likely to affect the distribution and diversity of plant and animal species. A rapidly developing area of biodiversity analysis is to model the envelope of climatic and ecological conditions under which a single species lives, deducing this from known features of the places where it is recorded. Such a bioclimatic model can be used to calculate a potentially wider set of areas where the species might occur, or predict its future distribution under changing climatic conditions and to project these on to a map of the world. This can be used to predict the responses of the species which may become endangered or conversely become a pest presenting a new or increased threat.

2.3. Phylogenetic analysis and biogeography

Phylogenetic analysis comprises a variety of methods for attempting to discover the evolutionary relationships between groups of organisms, and typically produces an evolutionary tree, or phylogeny, to describe these relationships. The BDW system is used for phylogenetic analyses of various taxonomic groups. A standard phylogenetic workflow includes: searching the EMBL [9] DNA sequence database for sequences; using this to produce a phylogeny using parsimony or maximum likelihood techniques [43]; and estimating the age of species and lineages within the tree using techniques of temporal calibration [40]. This phylogenetic workflow can be integrated with the bioclimatic

modelling workflow of Section 2.2. This has led us to the study of bioclimatic models from an evolutionary perspective, in particular to study the impact of historical climate change on the evolution of Mediterranean plant groups. We have developed ancestral bioclimatic models for the lineages of the sundews (a group of carnivorous plants in the family Droseraceae), and the popular garden flower genus *Cyclamen* (Myrsinaceae). These models have been projected into estimates of historical climate scenarios for 8–10 million years ago, to produce plausible estimates of ancestral distributions in geographical space. We have found that they demonstrate clear phylogenetic patterns coincident with historical shifts in climate, and they are helping us to understand the impact of climate change on the evolution of plants.

3. The current BDW system architecture

We have outlined the requirements of the BDW system and the way they have influenced the current architecture as part of an earlier publication [28]. An important requirement is to bring about interoperability between a diverse set of local and remote legacy databases and applications for a researcher to use these resources in designing complex workflows in a Grid environment. The architecture of the current BDW system is presented in greater detail in [36]. However in the present paper we provide an overview of the present system architecture in order to familiarise the reader with the important system components.

From the perspective of a biodiversity researcher, the system consists of a graphical, desktop-based workflow tool. The system allows linking of entities to create a workflow. By linking entities the user brings together several local or remote resources, legacy applications and analytical tools for conducting biodiversity experiments. A typical experiment involves retrieval of data from local or remote data sources and its processing by one or several applications or tools in a sequence or in parallel to derive useful results. From a system perspective, the BDW system has a multilayered architecture where interoperable components of each layer provide a set of operations and abstract the functionality of components of lower level layers from the layers above. Adopting a multilayered architecture has enabled the usage of remote resources in a web service-enabled Grid environment for executing workflow tasks. Figure 1 illustrates the current BDW system architecture. We describe important components of the system below.

3.1. The triana workflow management system

The BDW system uses the Triana workflow management system to provide workflow capabilities. The reasons for choosing Triana include its portability with different systems using Java, easy workflow assembly and its window-based visualisation tools for both designing and executing workflows. In Triana, a unit represents a wrapped resource or a tool. Triana invokes these wrapped resources as part of a workflow execution. It is inherently a flow based workflow application where data or control arriving at a particular unit triggers its execution. It has embedded features for authoring tasks and workflow units and provides tools for working with web service-based systems. A key point is that Triana has been developed locally at Cardiff and we have direct access to the Triana team for support. The version of Triana used in the BDW system has been extended to support resources and analytical tools for biological analyses in our web services-based Grid environment. In our extended version of Triana, the toolbox contains tools provided for use in the BDW system in addition to standard Triana tools. This includes tools pertaining to resource wrappers (described in Section 3.2), helper tools for accessing remote resources, configuration tools, data transformation tools (similar in purpose to Taverna shims [35] and KEPLER adaptors [23]) and data parsing tools. The tools can be simply dragged on to a workflow design panel as workflow units and linked with other workflow units to quickly assemble a workflow, as in the example shown later in Fig. 2. These tools are provided as part of BDW client-side libraries and integrated with the Triana workflow system for performing analyses in the BDW exemplar study areas.

3.2. Resource wrappers

In the BDW system, biodiversity resources are wrapped using BDW resource wrappers which provide an invocation mechanism that allows invocation of operations on heterogeneous resources in a standard manner. We have implemented wrappers for remote resources, local data stores or data stores accessible via JDBC drivers, cache databases, command-line tools, legacy applications such as WorldMap [45] and MATLAB, environmental modelling tools such as openModeller [16], wrappers for performing batch queries and wrappers for cataloguing, retrieval and visualisation of spatial data (including model outputs) within the BDW environment. The BDW system provides researchers

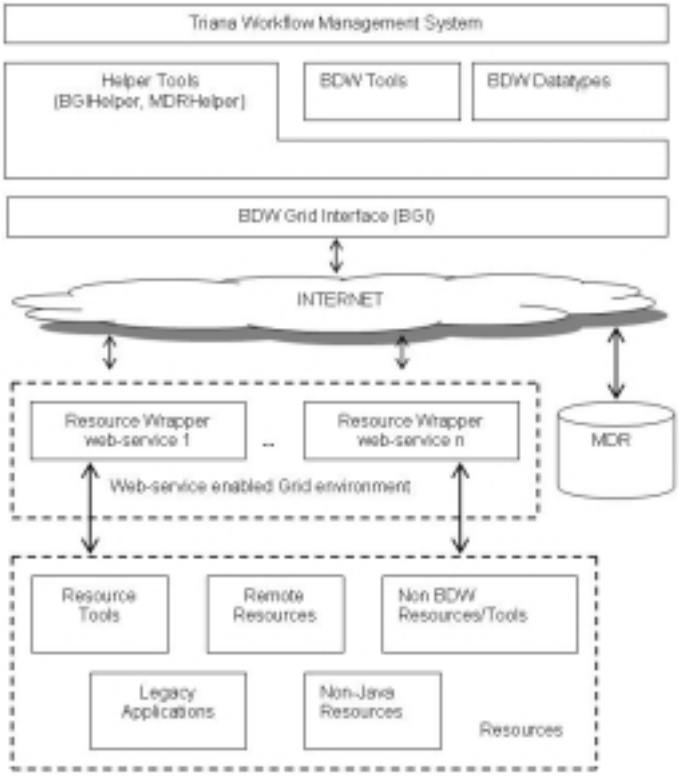

Fig. 1. The Current BDW system architecture.

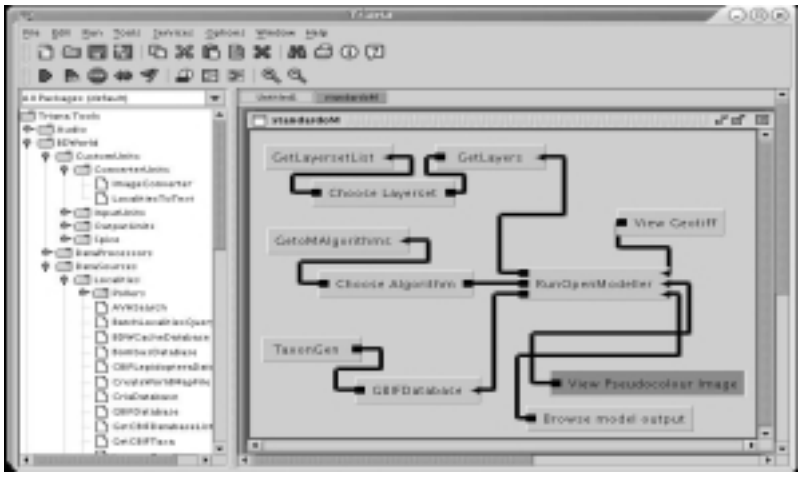

Fig. 2. A bioclimatic modelling workflow.

with wrappers supporting remote data resources from various parts of the world, local data sources, analytical tools and legacy applications. Some of the important remote resources wrapped in the BDW environment include the data portal of the Global Biodiversity Information Facility (GBIF) [11] and Australia's Virtual Herbarium (AVH) [7] – an online botanical information resource providing access to six million specimen records. Resource wrappers implemented so far have been written in the Java programming language and are deployed as web services using the Apache Axis [6] web services container. In the BDW PSE

the web service interfaces provide access to operations on resources and tools used in our exemplar study areas. Web service interfaces enable their standard invocation using the HTTP/XML-based SOAP messaging protocol from within the Triana application. It also allows usage of these resources and tools in distributed and computational environments when invoked from remote clients.

3.3. The BDW Grid Interface (*BGI*)

BDW workflow units access resource wrappers deployed in the web services environment via the BGI layer (see Fig. 1) which insulates Triana from the Grid resources and provides a standard mechanism for invoking operations on them. Within the multi-layered architecture of the BDW system the BGI acts as a bridge between the workflow units and the resource wrappers and insulates users from the complexities of resource wrappers deployed as web services. We provide a helper class called the BGIHelper for workflow units to use the BGI for invoking operations on resources. This provides workflow units with high-level access to the resource wrapper operations without needing to know about the underlying web services enabled Grid middleware which hosts resource wrappers.

3.4. The metadata repository (*MDR*)

In order to locate resources and build workflows, metadata is needed to enable the selection of resources meeting appropriate criteria. Metadata is currently being used to provide information about resources available to the BDW system. The present interface to the MDR facilitates querying the repository to find essential information regarding how a resource can be invoked, such as its location, the operations it supports, the parameters required by these operations and the returned data.

4. A bioclimatic modelling workflow

Currently the existing BDW system is being used to run bioclimatic modelling experiments on the whole of the plant family Fabaceae (bean family). This is a diverse family containing around 20,000 species representing around one twelfth of all flowering plants. They are distributed on all continents (excluding Antarctica) and are found within every major ecosystem. Global-scale models are currently being run under a number of climate scenarios and modelling algorithms in an attempt to quantify the potential risk of extinction of species of the family over the next 50 years. A degree of complexity is involved in supporting data and tools from different sources and in different formats which the BDW system manages as part of its PSE. In the research reported here around 1800 species are being modelled using 3 bioclimatic algorithms, under 4 climate scenarios consisting of 22 global climate surfaces: a total of approximately 22,000 bioclimatic modelling experiments.

Figure 2 illustrates a complete bioclimatic modelling workflow built and run using Triana. The central tool *RunOpenModeller* is a tool based upon openModeller [16] – an open-source spatial distribution modelling library providing a uniform method for modelling distribution patterns using different algorithms. The openModeller library is wrapped in a resource wrapper for use in the BDW environment.

RunOpenModeller requires three inputs:

1. The Localities data, which is provided by running the GBIFDatabase search tool, a resource wrapper for the GBIF portal [11]. This tool expects a string object pertaining to a taxon as an input and returns a localities collection (i.e. genus, species, latitude, longitude for each returned specimen) as output.
2. A parameter indicating which algorithm implemented in openModeller for modelling the potential distribution of species is required, such as GARP [42], BIOCLIM [34] or CSM [39].
3. A collection of layers (typically climate layers such as temperature and precipitation values on a geographical grid) which are supplied by other workflow units.

Using the specified algorithm, openModeller constructs a bioclimatic model by interpolating the climatic data at the point localities of the specimens specified by the localities collection. The bioclimatic model for an area of interest is projected under present or predicted future climate parameters specified by an appropriate layers collection. It finally displays the results for interpretation by overlaying the projection on to an appropriate base map. Figure 3 illustrates an example model output produced by executing the workflow illustrated in Fig. 2.

5. The new BDW system architecture

The present BDW system brings together heterogeneous resources for a researcher to conduct biodiver-

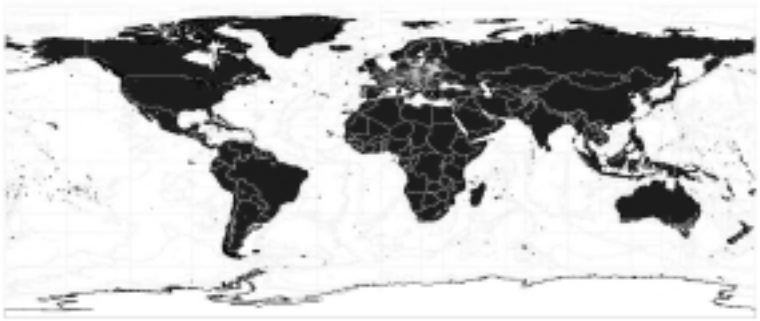

Fig. 3. Example model output for the clover species *Trifolium patens* Schreber (a member of the bean family). The map shows areas (shaded regions across Europe, South America, Asia and Australia) predicted to be suitable for the species in the 2050's using the bioclimatic modelling algorithm GARP and the Hadley Centre climate model using the SRES A1F climate scenario [26].

sity experiments. It serves the purpose of providing a standardised set of operations and interfaces for accessing heterogeneous resources and executing workflow components on remote nodes very well. However the existing architecture also presents a number of challenges when it comes to accessing computational resources for executing workflow tasks. It also does not provide the functionality of distributing user jobs across several nodes. Additionally the present architecture requires certain client-side libraries to be installed which restricts availability of the BDW resources to other applications or workflow tools. We believe that these restrictions have to be addressed so that workflows could be composed and executed with a greater degree of control and flexibility by a knowledgeable user where the user is free to choose the workflow manager of his choice for accessing resources in the BDW PSE. The limitations which the new architecture addresses are also described in greater detail in our earlier publication [37].

5.1. Approaches for accessing BDW resources in the new architecture

The new BDW architecture not only brings together heterogeneous resources and tools as before but also provides the ability to utilise available computational resources. This enhances the scalability of the system and provides the user with more control and choice in choosing appropriate computational nodes for executing workflow tasks. Figure 4 illustrates the new BDW architecture which uses middleware from the Condor [44] project for distributing the workload across the available execute nodes. In the new architecture BDW resources can be accessed in the two different ways described below.

The first method of accessing resources in the computational pool is for advanced users who would like first to identify the resources which are available and then match the resources with the requirements of the workflow tasks. For example, a user might be interested in nodes having any or all of the following capabilities such as higher processing speed, large disk storage, less network usage, lower average CPU load for a machine in a given duration of time, etc. In the BDW system, information about host metrics is provided by the Ganglia [10] cluster monitoring system. Giving the advanced user the ability to view host metrics provides users with the flexibility of using the node or nodes of their choice. Once the user identifies the node, the workflow manager can directly access the resource wrapper web service running on the node as shown in Fig. 4. A workflow manager can be instructed to allocate a particular workflow task to the preferred node by creating a workflow unit which represents the task. Bindings between the workflow unit and the resource wrapper web service available on a preferred compute node can be established by importing the WSDL of the web service into the workflow environment. This design enables invocation of different resources available in different compute nodes when creating a workflow. However, this design requires replication of BDW resources and wrappers across several nodes so that if a user's preferred node is not available or is busy serving another user, the user can choose from other available nodes which provide the same set of services.

In the second method of accessing BDW resources available in the BDW computational environment a user submits a job to a BDW Condor web service component which creates a job description file and submits the job to the Condor central manager. The Condor central manager then decides which node to run the job on based on available nodes. This design allows allocation

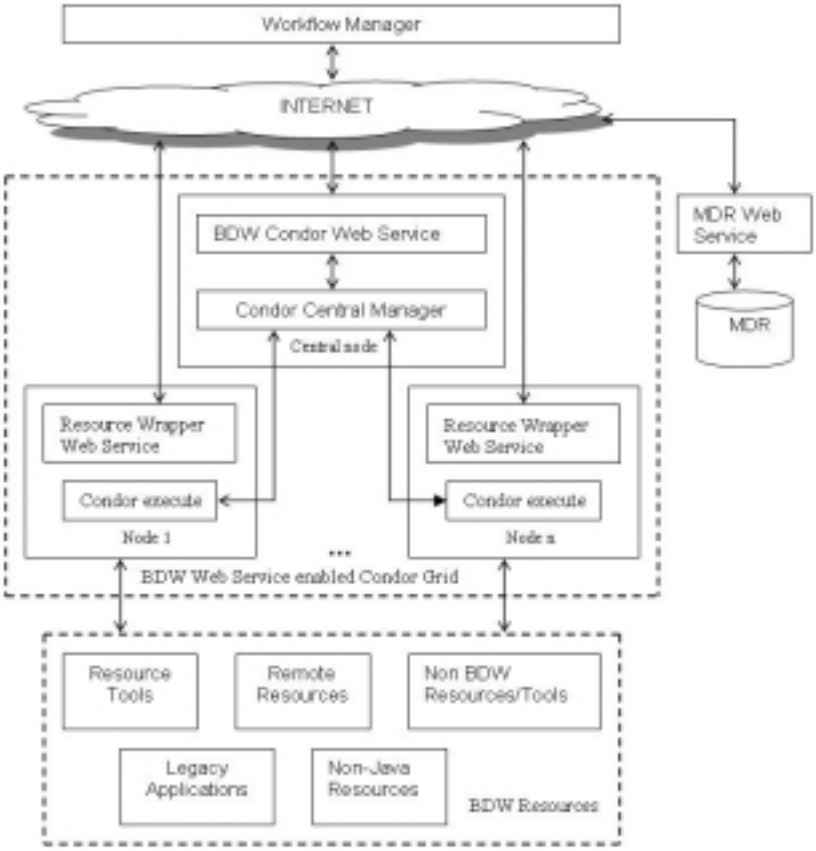

Fig. 4. The new BDW system architecture.

of computational resources dynamically and running of a large number of iterative tasks perhaps with different variables every time and spreading of work load across available execute nodes. This design can be adapted for Grid environments where Grid resources such as computational nodes change dynamically. The Condor middleware provides a mechanism of first transferring the required libraries to the computational node before running a job. This will allow BDW tasks to be executed in dynamically changing computational environments. Condor also allows utilisation of computational resources outside the BDW domain via the mechanism of flocking so that resources available from other Condor pools can also be used. For example as part of a future possibility BDW jobs can be run in Condor pools available from wider university networks. The use of the Condor middleware also allows submission of a large number of jobs as part of a single batch which are then queued to be executed on available nodes. This allows us to run several modelling experiments in parallel in available nodes. Although the execution time may be increased for a particular job waiting in the queue, the units in the Triana application are not blocked because of long waiting time. Triana units can be configured to pick up the results at a later stage or poll the *Job Progress* web service at regular intervals to check whether the execution is completed.

In both approaches, access to BDW resources available in the computational environment is provided directly via the web service interfaces. By combining the two approaches as part of the new architecture we have provided flexibility for workflow mangers to access computational resources of their choice. The two approaches can also be used together as part of a single workflow. While some tasks directly access a resource wrapper web service in a workflow, the others can be instructed to access resources via the BDW Condor web service. A node in the BDW PSE can be configured to run jobs submitted by the Condor central manager; and also process resource wrapper web service requests which are submitted directly by workflow managers at the same time. This allows better utilisation of com-

putational nodes and particularly those nodes with fast processors and better multitasking capabilities.

The new architecture allows external applications or workflow tools to access resources available in the BDW environment without requiring client-side libraries. However, at the present stage these resources can only be accessed individually from external applications. They cannot be composed into workflows using external workflow tools because bringing together heterogeneous resources in a workflow also requires addressing interoperability issues. Presently we provide data transformation tools at the client side which allow heterogeneous resources to interoperate when they are linked together in a workflow in our environment for the Triana workflow application only. As the work is still in progress we are presently making further resources and tools available in our environment. We are also investigating the issues pertaining to interoperability of heterogeneous resources as part of the European Network for Biodiversity Information (ENBI) [5] consortium for inter-linking and integration of the varied biodiversity information types (e.g. on taxonomy, collections, observation, etc.) that are distributed among different databases.

6. Secure access to the BDW PSE

In order to provide secure access to resources in the BDW PSE we have developed a security model called BDWShib which uses the capabilities of the Shibboleth security framework briefly introduced in Section 1. When accessing resources protected by the Shibboleth framework, applications such as web browsers are normally used. The novelty of this part of our work is that the BDWShib system provides programmatic access to Shibboleth protected resources, whereas in most other applications the Shibboleth framework is used to provide a user based access to protected resources which we describe in Section 6.1. As part of our approach we describe how Shibboleth authentication can be performed in non-browser based environments such as desktop-based applications which interact with web-accessible Shibboleth protected resources on behalf of its users. This approach then allows us to perform user authentication from within desktop-based Triana workflow application during workflow execution.

6.1. The Shibboleth security framework

The web based Shibboleth framework provides a mechanism for exchanging attributes across different organisations for the purpose of authorisation [20]. It enables a user to access a protected resource or service at a remote domain (Service Provider or SP) by using the user's own home security domain (Identity Provider or IdP) to perform user authentication. The framework uses X.509 certificates for the underlying secure attribute exchange which involves sending a user's identity information to the user's home institution. An important advantage the framework provides is that the user is not required to possess an X.509 certificate. This is because Shibboleth allows inter-institutional sharing of resources within a trusted federation where it is the responsibility of the home institutions to authenticate their users. Therefore Shibboleth directs the users to their home institution for authentication. The information which is exchanged as attributes helps to determine whether to grant the user access to the resource at the SP. Shibboleth uses Security Assertion Mark-up Language (SAML) [15], an OASIS standard for exchanging authentication and authorisation statements, between the IdP and the SP. When the user is authenticated, the Shibboleth component at the SP establishes and maintains a session with the user's web browser on behalf of the resource the user is accessing. This session consists of cookies which are passed between the web browser and web server. The cookies are associated with a security context which holds the user's authentication information and a set of attributes describing the user's identity. The Shibboleth framework does not support logout functionality. The user can access the resource more than once without repeating the Shibboleth authentication process until the cookies expire or are deleted from the user's machine.

6.2. Using the Shibboleth framework for the BDW system

The Shibboleth framework enables the creation of a federation to build trust relationships between participating organisations for inter-institutional access of resources. These organisations exchange attributes using the Shibboleth protocols and abide by a common set of policies and practices [20]. The Shibboleth framework is appropriate for the BDW application domain because the users of the BDW system and resources available in the BDW PSE span more than one organisation. The Shibboleth framework also uses the user's

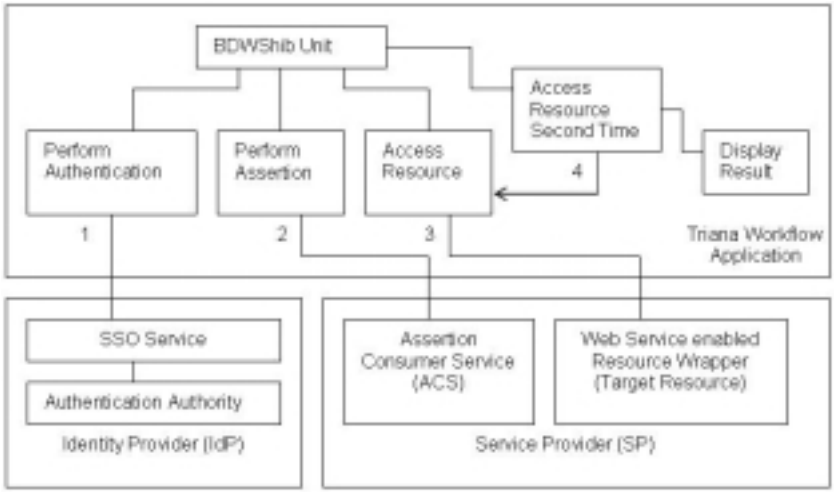

Fig. 5. The BDWShib system architecture (inside the Triana workflow application).

credentials available from their home organisation for authentication. Hence by using Shibboleth the responsibility of user authentication can be devolved to the user's home institution. This avoids the need to create a separate authentication system that is exclusive to the BDW system. The Shibboleth framework also provides a scalable and extensible solution for managing access to resources. By using the Shibboleth framework it is possible to accommodate a growing number of users from different institutions as part of a federated access management arrangement. The Shibboleth framework is also being adopted by a number of higher educational institutes in the UK to develop the next generation access-management system for their users.

6.3. The BDWShib system architecture

The BDWShib system enables user authentication for a resource protected by the Shibboleth framework. The system in its present version is being designed and implemented to be used from within the Triana application. However, the system can also be modified to be used as a standalone application. On the client side, the BDWShib system provides a non-browser based solution for performing user authentication. This scenario is applicable to desktop-based systems using Java or environments where direct interaction of a user with the Shibboleth protected resource is not possible or is not required. For example, in our case the workflow units in the desktop-based Triana application have to interact with Shibboleth protected resources in the BDW PSE on behalf of its users. This scenario is also applicable to environments where a set of tasks are required to be batch processed at certain times during the day or at regular intervals without requiring user presence at all times. Hence in the BDW PSE with the BDWShib system we use Shibboleth for gaining programmatic access to resources as opposed to the interactive user-based access for which Shibboleth is normally used.

Using the prototype BDWShib system, we presently provide secure access to data from the database of the International Legume Database & Information Service (ILDIS) [13]. The ILDIS project aims to document and catalogue the world's legume species diversity in a readily accessible form. The ILDIS project website provides access to the database of legumes where a user can search the database by providing the scientific name of a legume species and retrieve all the data pertaining to that species. The BDWShib module inside the Triana workflow application provides programmatic access to the ILDIS database for use when conducting a biodiversity experiment.

The BDWShib system is implemented using the Java programming language and uses the Apache Axis toolkit. The system also uses the Jakarta Commons HttpClient [14] library to request web pages from the Shibboleth protected resource wrapper web service in the BDW PSE using the secure HTTPS protocol from the Triana application. After the user is authenticated, invocation of operations and exchange of information using SOAP messages also takes place via the HTTPS protocol. The BDWShib module is incorporated into the Triana application as a tool within its toolbox. The tool can be dragged on to the workflow design panel just like any other task as a workflow unit and linked with other units to quickly assemble a workflow. Fig-

ure 5 describes the architecture of the BDWShib system and the sequence of operations which are performed in order to authenticate a user and allow authorised access to a protected resource. The architecture of the system can be summarised as a series of steps which the system performs.

Step 1. When the BDWShib workflow unit is executed by the workflow manager it pops up a window (Fig. 6) which performs the Single Sign-On (SSO) on user's behalf. In the SSO window the user provides username and password for authentication. The 'Target Resource' field identifies the protected resource(s) the user is trying to access (which are resource wrapper web services in our case). The 'Perform Authentication' component first accesses the target resource for identifying details such as URL of the IdP, Assertion Consumer Service (ACS), etc. as shown in Fig. 6. The component then performs user authentication by sending the username and password via the HTTPS protocol to the SSO Service at the IdP. The SSO service processes authentication requests and obtains authentication statements for the user. If authentication is successful the SSO service responds to the BDWShib system with a digitally signed SAML response containing an authentication assertion for delivering to the SP. The other components of the Shibboleth framework at the IdP side such as the Attribute Authority and the Artefact Resolution Service are not shown in Fig. 5.

Step 2. As part of this step, the BDWShib workflow unit, via the 'Post Assertion' component in the BDWShib system, issues a HTTP POST request to the ACS at the SP. The ACS is presented with the authentication assertion returned by the SSO service in the previous step. The ACS processes the authentication assertion and establishes a security context at the SP using a cookie-based session [18]. The ACS service finally redirects the user to the target resource. The other Shibboleth framework component at the SP called the Attribute Requester is not shown in Fig. 5.

Step 3. Once the security context is established at the SP in step 2, the 'Access Resource' component is allowed to access the resource. At this stage, we can invoke operations on the ILDIS database resource wrapper web service by exchanging SOAP messages with the web service. The session which is established with the SP at step 2 using cookies is shared with the Axis toolkit for invoking operations on the target resource via the HTTPS protocol.

Step 4. By using the same security context established in Step 2 more than one operation can be invoked at the target resource without the user having to go through the authentication process again. This is possible using the 'Access Resource' component which keeps Java objects pertaining to resource access (and cookie information) alive in the memory until the workflow is terminated. The operations can belong to the same resource wrapper or other resource wrapper web services pertaining to different resources available in a given compute node. After data is retrieved from the web service it is used for further analysis in the BDW PSE, or is displayed to the user.

7. Related work

The field of bioinformatics is diverse and a number of research projects in its different areas such as genetic studies, structural studies of cells and tissues, cellular processes, etc. have developed tools for the application of informatics techniques to biological data to address the challenges identified. The web based myGrid project [25] provides middleware for conducting *in silico* experiments in biology. One example application of the project is in the area of genetic studies for gaining new insights into diseases having a strong genetic component (such as Graves' disease) with the aim of aiding the process of designing novel therapies. The GeneGrid [27] provides access to resources and tools to biologists interested in the development of antibodies and drugs. The BASIS project [31] serves the biology of ageing at the cell, tissue and organism level by providing access to diverse biological resources to conduct experiments such as constructing a virtual ageing cell.

The Science Environment for Ecological Knowledge (SEEK) [17] project provides a range of tools for accessing and analysing a wide range of ecological and biodiversity data from heterogeneous sources. The project provides the KEPLER [32] workflow tool based on an actor-oriented modelling approach for modelling and analysing scientific workflows. An important research area of the SEEK project is Ecological Niche Modelling (ENM) for understanding biodiversity patterns. Although similarities can be drawn between the BDW and the SEEK projects which have overlapping approaches they differ in application areas. One of the important research areas of the SEEK project is analysing the effect of climate change on over 2000 mammal species in North America [38] by using data resources of the distributed Mammal Networked Information System (MaNIS) [1] consortium.

Current methods for phylogenetic analysis such as those used in the work described in Section 2.3 are

Fig. 6. The BDWShib SSO window.

limited to small data sets. CIPRES (Cyberinfrastructure for Phylogenetic Research) [8] is a project which is developing models, algorithms and tools applicable to large data sets containing hundreds of thousands of biomolecular sequences. One of several ways for users to access these tools will be through the Kepler workflow interface [4].

The Pegasus portal [41] is a web based computational portal which allows access to Grid resources via a standard web browser using HTTP(S) protocol. This portal provides an approach for creating abstract workflows using Chimera [24] by specifying the metadata description of the desired data to be generated or analysis to be done. The portal supports two applications, LIGO [21] in gravitational-wave astronomy and Montage [22] for generating astronomical image mosaics. The P-GRADE [3] Grid Portal is a workflow-oriented Grid portal which enables execution of job workflows using Grid middleware based on Globus technology. The GENIUS [12] portal provides web-based access to the EGEE Grid infrastructure.

8. Conclusions and further work

The BDW system brings together disparate resources and analytical tools for biodiversity researchers to solve problems in biodiversity and analyse biodiversity patterns. The system allows linking of these tools and resources into a workflow so that different activities performed as part of an experiment are automated and the experiment as a whole is conducted more efficiently. This is an improvement on the manual process of performing each activity individually using separate systems which are not linked. The new BDW architecture provides users with the additional flexibility to utilise computational resources either by submitting workflow tasks to the BDW Condor web service or invoking web service enabled resource wrappers directly whilst conducting an experiment in the BDW PSE. The BDW system, based on the present architecture, is being actively applied in three exemplar study areas of biodiversity informatics. However we are also progressively making BDW resources available as part of the new architecture. We aim to utilise the National Grid Service (NGS) facilities such as Condor pool and data storage available locally at the Welsh e-Science Centre (http://www.wesc.ac.uk/) to meet the increasing computational requirements of biodiversity experiments.

Acknowledgments

The project was funded by a research grant from the UK Biotechnology and Biological Sciences Research

Council (BBSRC). The BDWShib part of the project was funded by the Joint Information Systems Committee (JISC) as part of the Core Middleware Infrastructure Programme. We are grateful to a good number of collaborators for making data and tools available to the project. In particular we thank Species 2000 and the Hadley Centre for Climate Prediction and Research for providing us with valuable resources. We also thank an anonymous reviewer for helpful suggestions, Cardiff University's Information Services Directorate and the Triana team for the support they provided to the project.

References

[1] The Mammal Networked Information System (MaNIS) website. [Online]. Available: http://manisnet.org/, 2001.
[2] The Canadian Biodiversity Information Facility Portal. [Online]. Available: http: //www.cbif.gc.ca/portal/digir-toc.php, 2003.
[3] The P-GRADE Grid Portal. [Online]. Available: http://www.lpds.sztaki.hu/pgportal/, 2005.
[4] CIPRES Software. [Online]. Available: http://www.phylo.org/sub_sections/software.htm, 2006.
[5] European Network for Biodiversity Information. [Online]. Available: http://www.enbi.info/forums/homedir/clusterIII.php, 2006.
[6] The Apache Web Services Project. [Online]. Available: http://ws.apache.org/axis/, 2006.
[7] The Australia's Virtual Herbarium Website. [Online]. Available: http://www.chah.gov.au/avh/avh.html, 2006.
[8] The Cyberinfrastructure for Phylogenetic Research CIPRES project website.[Online]. Available: http://www.phylo.org/, 2006.
[9] The European Bioinformatics Institute Website. [Online]. Available: http://www.ebi.ac.uk/, 2006.
[10] The Ganglia Monitoring System. [Online]. Available: http://ganglia.sourceforge.net/, 2006.
[11] The GBIF portal. [Online]. Available: http://www.gbif.org/, 2006.
[12] The GENIUS Portal Overview. [Online]. Available: http://egee.cesnet.cz/en/user/genius-guide.pdf, 2006.
[13] The International Legume Database & Information Service Website. [Online]. Available: http://www.ildis.org/, 2006.
[14] The Jakarta Commons HTTP Client. The Apache Jakarta Project. [Online]. Available: http://jakarta.apache.org/commons/httpclient, 2006.
[15] The OASIS Security Services (SAML) TC. [Online]. Available: http://www.oasis-open.org/committees/tc_ home.php?wg_abbrev=security, 2006.
[16] The openModeller. [Online]. Available: http://openmodeller.sourceforge.net/, 2006.
[17] The SEEK Project Proposal. [Online]. Available: http://seek.ecoinformatics.org/Wiki.jsp?page=SEEKProjectProposal, 2006.
[18] The Shibboleth Architecture Technical Overview. [Online]. Available: http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf, 2006.
[19] The Shibboleth Project – Internet2 Middleware. [Online]. Available: http://shibboleth.internet2.edu/, 2006.
[20] The Shibboleth Target Deployment Guide. [Online]. Available: http://www.switch.ch/aai/docs/shibboleth/internet2/1.2/deploy-guide-target1.2.1.html, 2006.
[21] A. Abramovici, W.E. Althouse, R.W.P. Drever, Y. Gursel, S. Kawamura, F.J. Raab, D. Shoemaker, L. Sievers, R.E. Spero and K.S. Thorne, LIGO: The Laser Interferometer Gravitational-Wave Observatory, *Science* **256**(5055) (1992), 325–333.
[22] G.B. Berriman, J.C. Good, A.C. Laity, A. Bergou, J. Jacob, D.S. Katz, E. Deelman, C. Kesselman, G. Singh, M. Su and R. Williams, Montage: A Grid Enabled Image Mosaic Service for the National Virtual Observatory. *Proc. Astronomical Data Analysis Software and Systems* (*ADASS*) *XIII*, ASP Conference Series, 314, 2003.
[23] S. Bowers and B. Ludascher, *Actor-Oriented Design of Scientific Workflows*, In 24th Intl. Conf. on Conceptual Modeling (ER 2005). LNCS, 2005.
[24] I. Foster, J. Voeckler, M. Wilde and Y. Zhao, *Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation*, In Proc. 14th Conference on Scientific and Statistical Database Management, 2002.
[25] C. Goble, C. Wroe and R. Stevens, *The MyGrid Project: Services, Architecture and Demonstrator*, In Proc. UK e-Science All Hands Meeting 2003 (AHM'03), Nottingham, UK, Sept. 2003.
[26] C. Gordon, C. Cooper, C.A. Senior, H. Banks, J.M. Gregory, T.C. Johns, J.F. Mitchell and R.A. Wood, The Simulation of SST, Sea Ice Extents and Ocean Heat Transports in aversion of the Hadley Centre Coupled Model without Flux Adjustments, *Climate Dynamics* **16** (2000), 147–168.
[27] P.V. Jithesh, N. Kelly, S. Wasnik, P. Donachy, T. Harmer, R. Perrott, M. McCurley, M. Townsley, J. Johnston and S. McKee, *Bioinformatics Application Integration in GeneGrid*, In *Proc. UK e-Science All Hands Meeting 2005* (*AHM'05*), Nottingham, UK, Sept. 2005.
[28] A.C. Jones, R.J. White, W.A. Gray, F.A. Bisby, N. Caithness, N. Pittas, X. Xu, T. Sutton, N.J. Fiddian, A. Culham, M. Scoble, P. Williams, O. Bromley, P. Brewer, C. Yesson and S. Bhagwat, *Building a Biodiversity GRID*, In Grid Computing in Life Science: First International Workshop on Life Science Grid, Revised selected and invited papers (LNCS/LNBI 3370), Kanazawa, Japan, 2004. Springer-Verlag.
[29] A.C. Jones, X. Xu, N. Pittas, W.A. Gray, N.J. Fiddian, R.J. White, J.S. Robinson, F.A. Bisby and S.M. Brandt, *SPICE: A Flexible Architecture for Integrating Autonomous Databases to Comprise a Distributed Catalogue of Life*, In Proc. 11th International Conference on Database and Expert Systems Applications, London, UK, 2000. Springer-Verlag.
[30] J.B. Kennedy, R. Kukla and T. Paterson, *Scientific Names are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition are Required for Accurate Data Integration*, In Proc. 2nd Intl. Workshop on Data Integration in the Life Sciences (DILS), San Diego, USA, 2005. Springer LNBI 3615, 80–95.
[31] T.B.L. Kirkwood, R.J. Boys, C.S. Gillespie, C.J. Proctor, D.P. Shanley and D.J. Wilkinson, Towards an E-Biology of Ageing: Integrating Theory and Data, *Nature Reviews Molecular Cell Biology* **4** (2003), 243–249.
[32] B. Ludascher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, J. Tao and Y. Zhao, *Scientific Workflow Management and the Kepler System*, In Special Issue on Scientific Workflows, Concurrency and Computation: Practice and Experience, 2005.

[33] D. Maier, E. Landis, J. Cushing, A. Frondorf and A. Silberschatz, Research Directions in Biodiversity and Ecosystem Informatics, in: *Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics*, J.L. Schnase, ed., NASA Goddard Space Fight Center, Greenbelt, Maryland, June 22–23, 2001.

[34] H.A. Nix, A biogeographic analysis of Australian elapid snakes, in: *Australian Flora and Fauna Series Number 7*, R. Longmore, ed., Canberra, Australian Government Publishing Service, 1986, pp. 4–15.

[35] T. Oinn, M. Greenwood, M. Addis, M.N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M.R. Pocock, M. Senger, R. Stevens, A. Wipat and C. Wroe, Taverna: lessons in creating a workflow environment for the life sciences, *Concurr Comput Pract Exper* **18**(10) (2005), 1067–1100.

[36] J.S. Pahwa, P. Brewer, T. Sutton, C. Yesson, M. Burgess, X. Xu, A.C. Jones, R.J. White, W.A. Gray, N.J. Fiddian, F.A. Bisby, A. Culham, N. Caithness, M. Scoble, P. Williams and S. Bhagwat, *Biodiversity World: A Problem-Solving Environment for Analysing Biodiversity Patterns*, In 6th IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2006), Singapore, May 16–19, 2006.

[37] J.S. Pahwa, R.J. White, A.C. Jones, M. Burgess, W.A. Gray, N.J. Fiddian, T. Sutton, P. Brewer, C. Yesson, N. Caithness, A. Culham, F.A. Bisby, M. Scoble, P. Williams and S. Bhagwat, *Accessing Biodiversity Resources in Computational Environments from Workflow Applications*, In Workshop on Workflows in Support of Large-Scale Science (in conjunction with the 15th IEEE International Symposium on High Performance Distributed Computing), Paris, France, June 20, 2006.

[38] D.D. Pennington, D. Higgins, A.T. Peterson, M.B. Jones, B. Ludaescher and S. Bowers, Ecological Niche Modeling Using the Kepler Workflow System, in: *Workflows for eScience: Scientific Workflows for Grids*, (Chapter 8), I.J. Taylor, E. Deelman, D. Gannon and M.S. Shields, eds, Springer, 2006.

[39] M.P. Robertson, N. Caithness and M.H. Villet, A PCA-based modelling technique for predicting environmental suitability for organisms from presence records, *Diversity and Distributions* **7** (2001), 15–27.

[40] M.J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, *Bioinformatics* **19** (2003), 301–302.

[41] G. Singh, E. Deelman, G. Mehta, K. Vahi, M.-H. Su, G.B. Berriman, J. Good, J.C. Jacob, D.S. Katz, A. Lazzarini, K. Blackburn and S. Koranda, *The Pegasus Portal: Web Based Grid Computing*, In Proc. ACM symposium on Applied computing (2005), Santa Fe, New Mexico, 2005, 680–686.

[42] D. Stockwell and D. Peters, The GARP modelling system: Problems and solutions to automated spatial prediction, *International Journal of Information Science* **13** (1999), 143–158.

[43] D.L. Swofford, PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods), Version 4.0 b10. Sinauer Associates, Sunderland, Massachusetts, 2002.

[44] D. Thain, T. Tannenbaum and M. Livny, *Distributed Computing in Practice: The Condor Experience*, (Vol. 17), In Concurrency and Computation: Practice and Experience, 2005, 323–356.

[45] P. Williams, Biodiversity and WorldMap. [Online]. Available: http://www.nhm.ac.uk/research-curation/projects/worldmap/index.html, 2006.