CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Designing and Building a Biodiversity Grid:

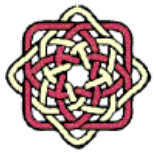
the Biodiversity World Project

A talk in the workshop "e-Research - Meeting New Research Challenges" at the Welsh e-Science Centre, 14 February 2006

Richard White, Andrew Jones, Alex Gray, Jaspreet S. Pahwa, Mikhaila Burgess

Cardiff University, UK

R.J.White@cs.cf.ac.uk

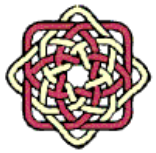


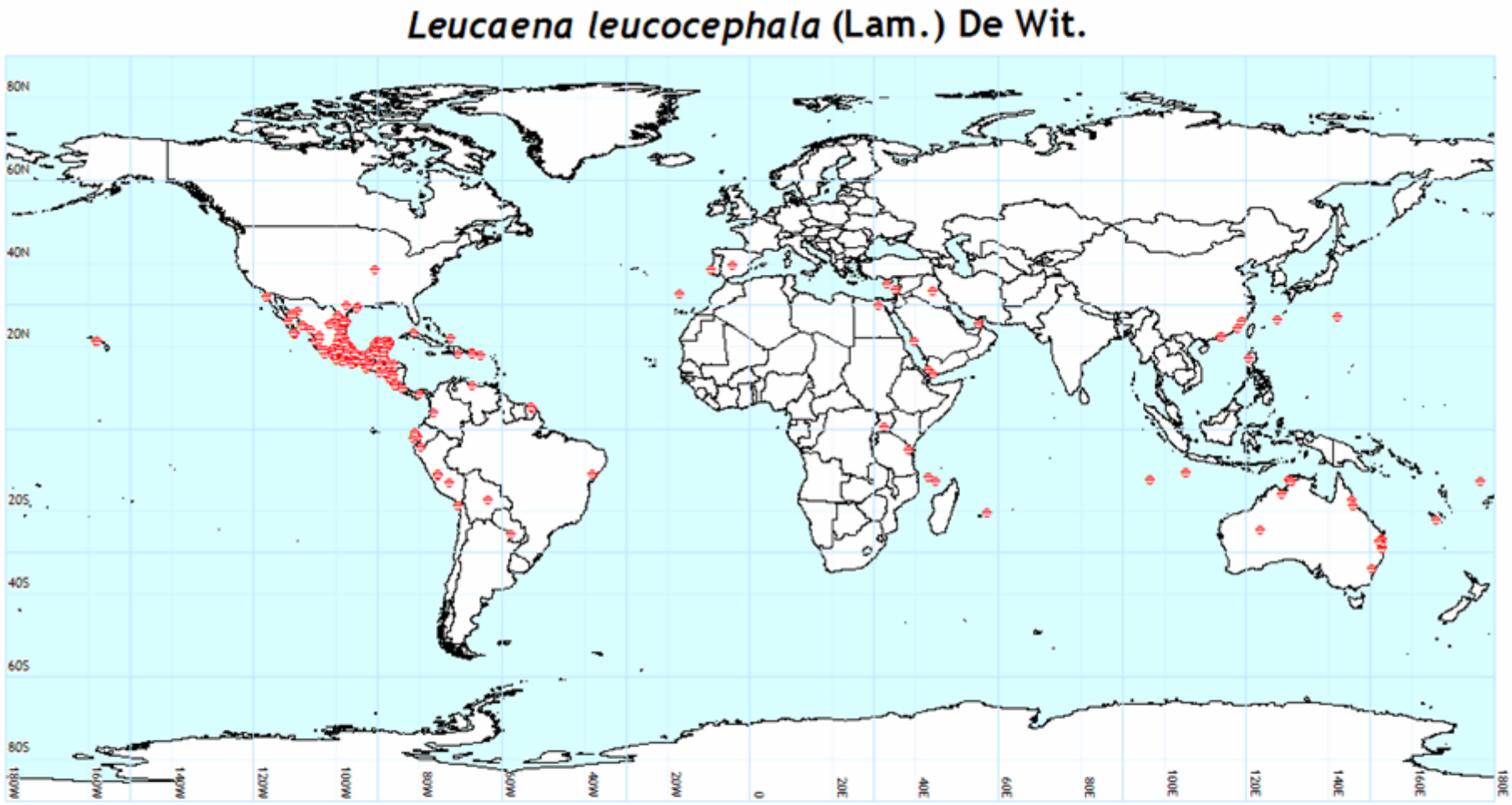
The Biodiversity World project

- 3 year e-Science project funded by the UK BBSRC research council, 2003-2006
- Universities of Cardiff, Reading and Southampton
- The Natural History Museum (London)

Some difficult biodiversity questions

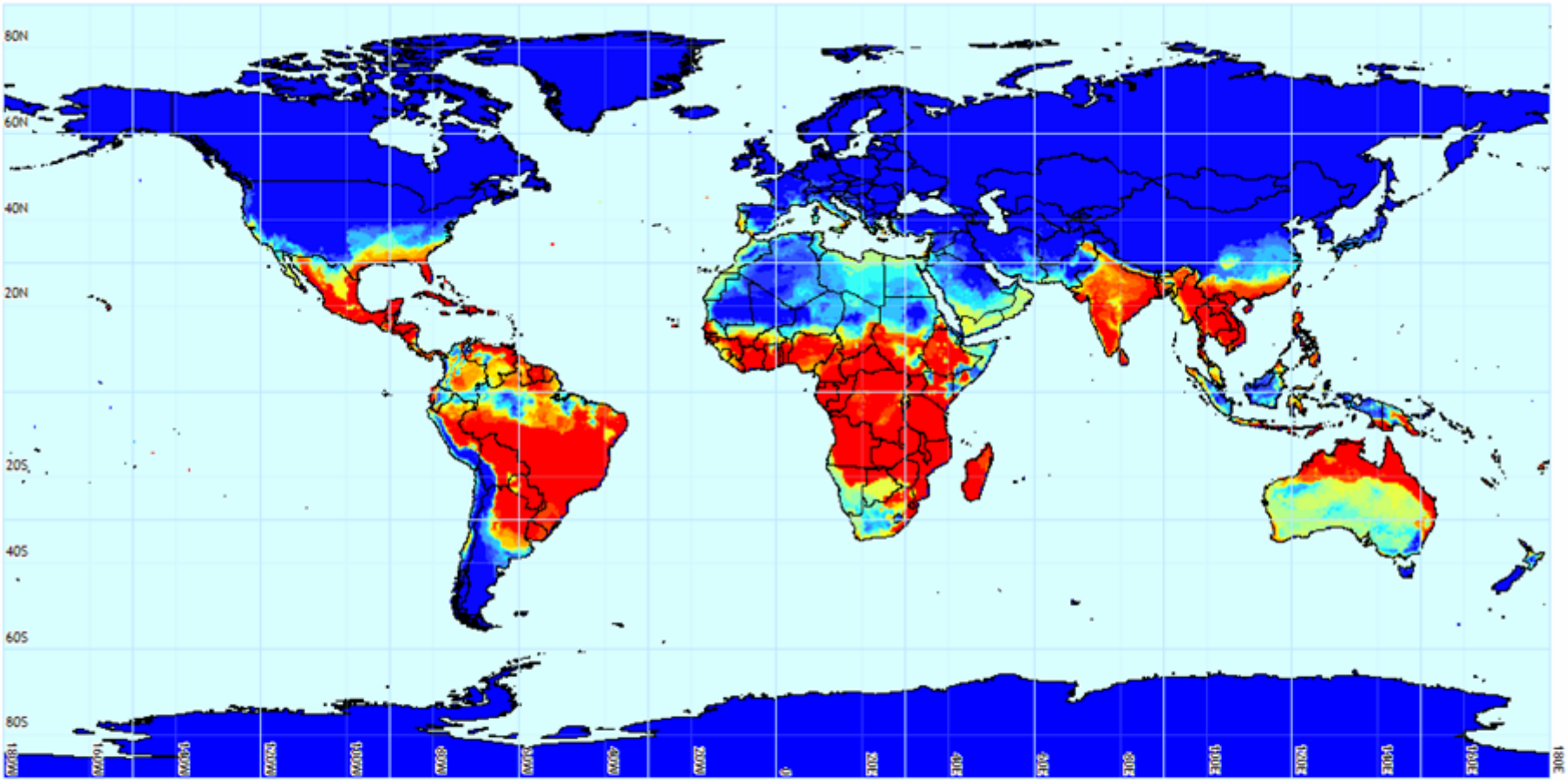
- How should conservation efforts be concentrated?
  - (example of Biodiversity Richness & Conservation Evaluation)
- **Where might a species be expected to occur, under present or predicted climatic conditions?**
  - (example of Bioclimatic & Ecological Niche Modelling)
- How can geographical information assist in inferring possible evolutionary pathways?
  - (example of Phylogenetic Analysis & Palaeoclimate Modelling)





Point data from various herbaria

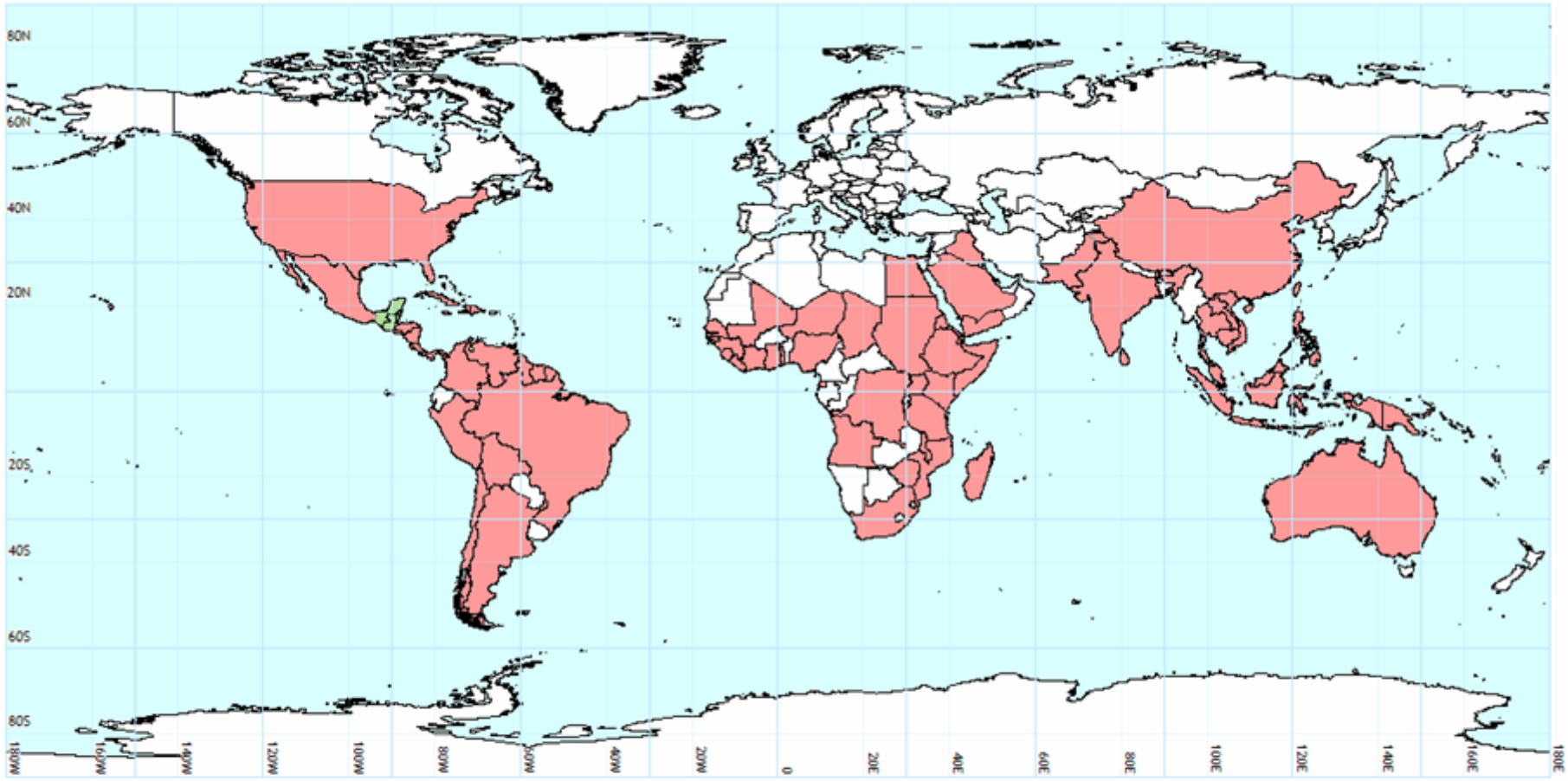
Leucaena leucocephala (Lam.) De Wit.



GARP prediction of climatic suitability

Taxonomic Database Working Group (TDWG)
WorldSat International, Inc.
Environmental Systems Research Institute, Inc. (ESRI)

Leucaena leucocephala (Lam.) De Wit.



Distribution data from ILDIS database

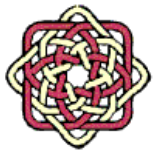
ILDIS Geography

- Introduced
- Native
- Uncertain

Taxonomic Database Working Group (TDWG)
WorldSat International, Inc.
Environmental Systems Research Institute, Inc. (ESRI)

Types of resource used in these biodiversity studies

- Data sources:
  - Catalogue of Life (names of species: Species 2000, GBIF)
  - Biodiversity data
    - Descriptive data
    - Distribution of specimens and observations
  - Geographical data
    - Boundaries of geographical & political units
    - Climate surfaces
  - Genetic sequences
- Analytic tools:
  - Biodiversity richness assessment – various metrics
  - Bioclimatic modelling – bioclimatic 'envelope' generation
  - Phylogenetic analysis (generation of phylogenetic trees)

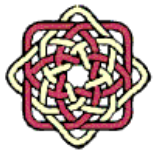


Some challenges …

- Finding the resources
- Knowing how to use these heterogeneous resources
  - Originally constructed for various reasons
  - Often little thought was given to standards or interoperability

The Biodiversity World vision (1)

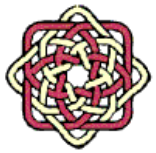
- Problem Solving Environment for Biodiversity studies –
  - Heterogeneous diverse resources
    - *Facilitating integration of both legacy and newly-developed resources*
  - Flexible workflows
  - Main challenges centre around interoperability, resource discovery, metadata, etc;
    - *High-performance computing secondary (though relevant)*



The Biodiversity World vision (2)

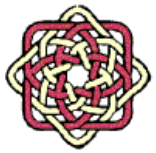
- Distinctive features:
  - a biodiversity informatics Grid
  - interoperability with heterogeneous data, complex in structure
  - resilience to infrastructure change & interoperation with other Grids
  - interactive collaboration a secondary concern

- We want to automate tasks such as the previous example analysis, as shown later

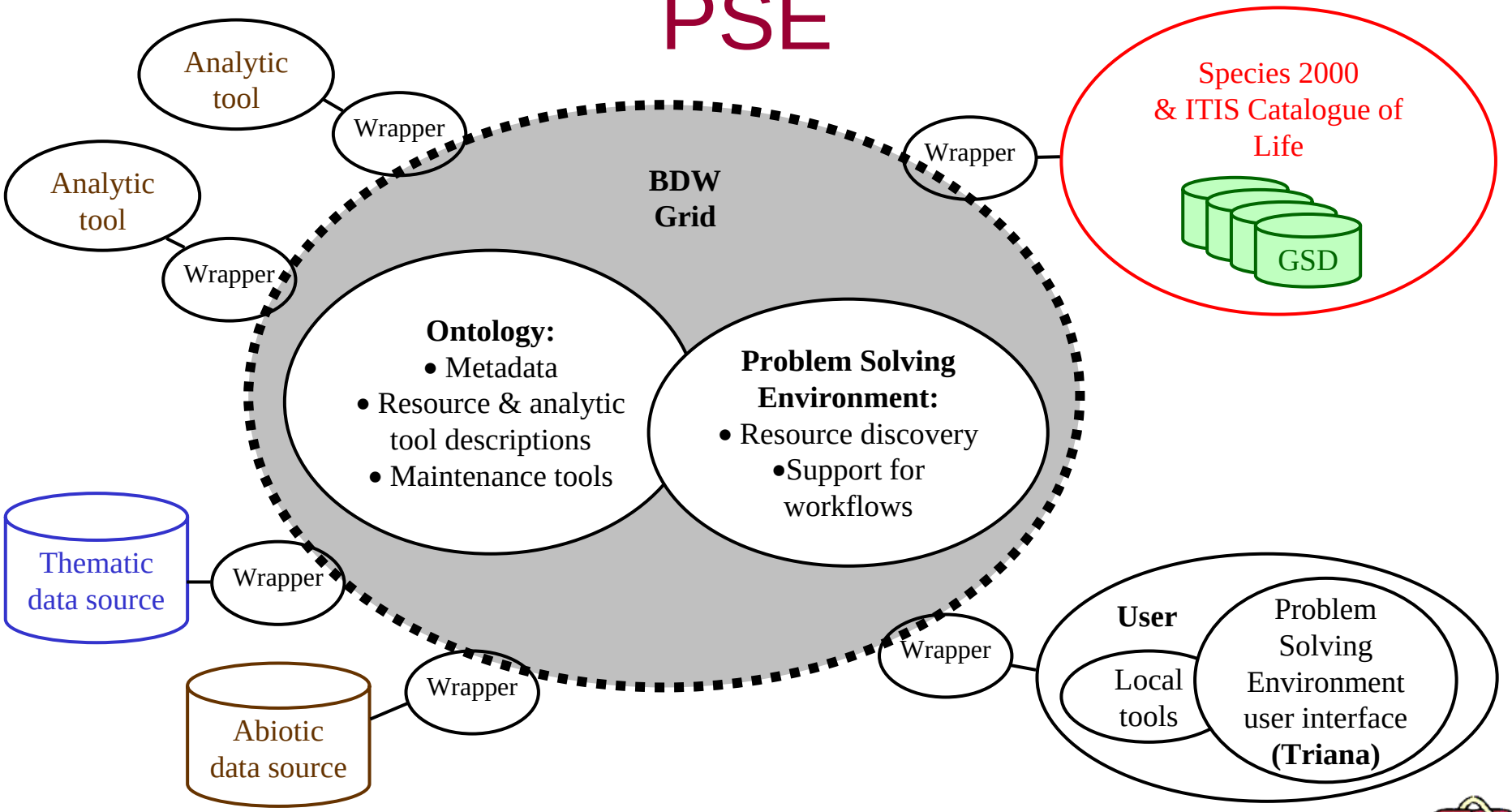


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Our architecture …



Biodiversity World as a flexible PSE



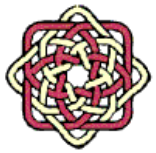
Role of metadata

Metadata is needed to enable discovery of resources and to indicate how they are to be used

- Properties to help locate appropriate resources
- Check interoperability, suggest transformations
- Provenance of data sets
- Log of work-flows executed

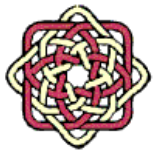
Biodiversity World Wrappers

- A mechanism to provide consistent interface to resources using a standard resource invocation mechanism
  - Operations on remote resources are invoked via the *invokeOperation(resource, operation, dataCollection)* method implemented by all the wrappers
- Wraps various kinds of resources and analytic tools
  - Insulate the core BDWorld System from heterogeneous resources
  - Retain flexibility to use various operations supported by each resource
- Solves the problem of interoperability between client and heterogeneous resources
  - Wrappers give consistent form to data retrieved from heterogeneous resources by encapsulating them into a set of standard BDWorld data types
- Can be deployed in Web Services/Grid environment



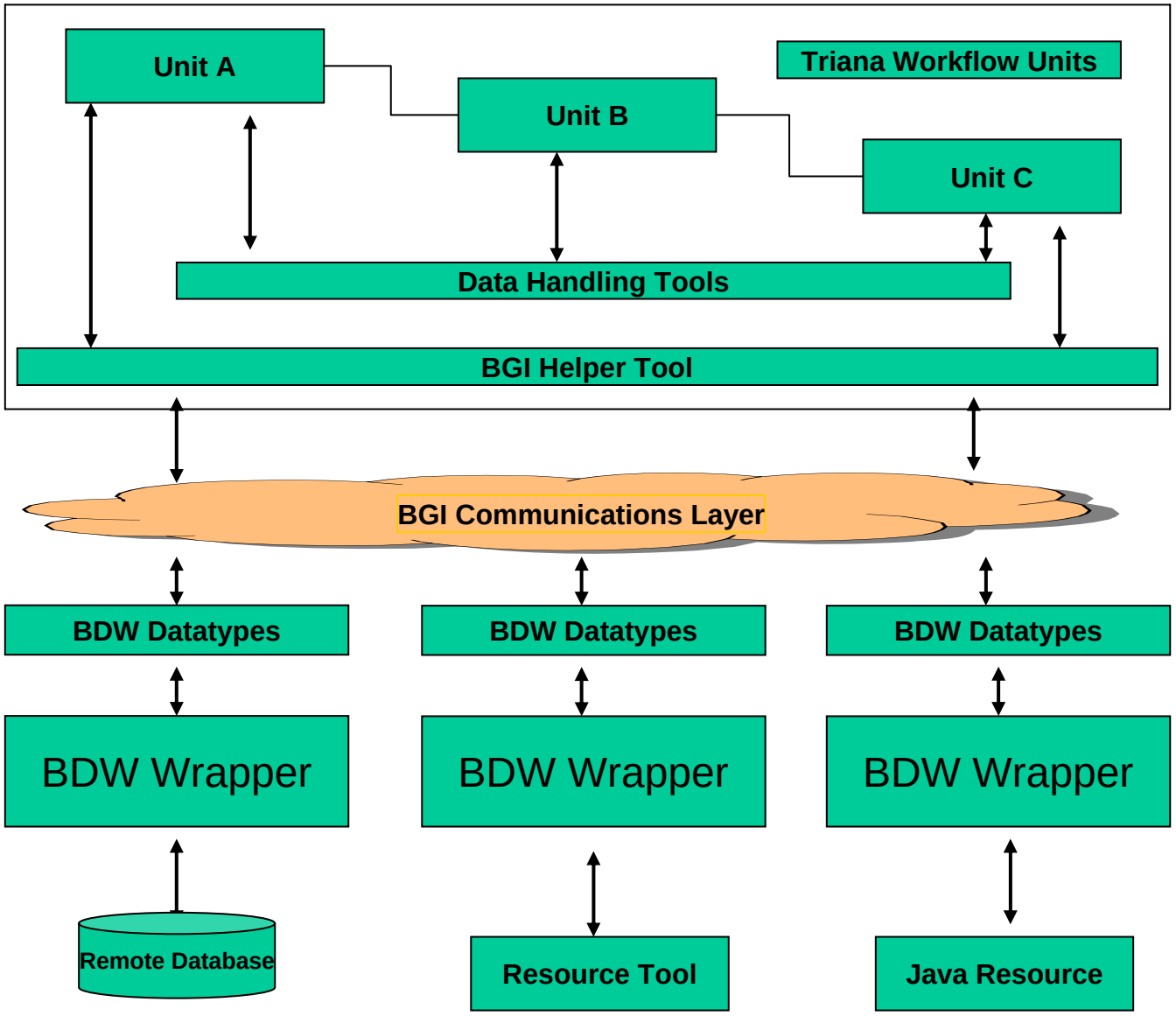
Interoperability in Biodiversity World

- Have defined Biodiversity World – Grid Interface (BGI) addressing the need to:
  - wrap resources to hide heterogeneity
  - insulate from infrastructure change
  - use metadata to cope with remaining heterogeneity

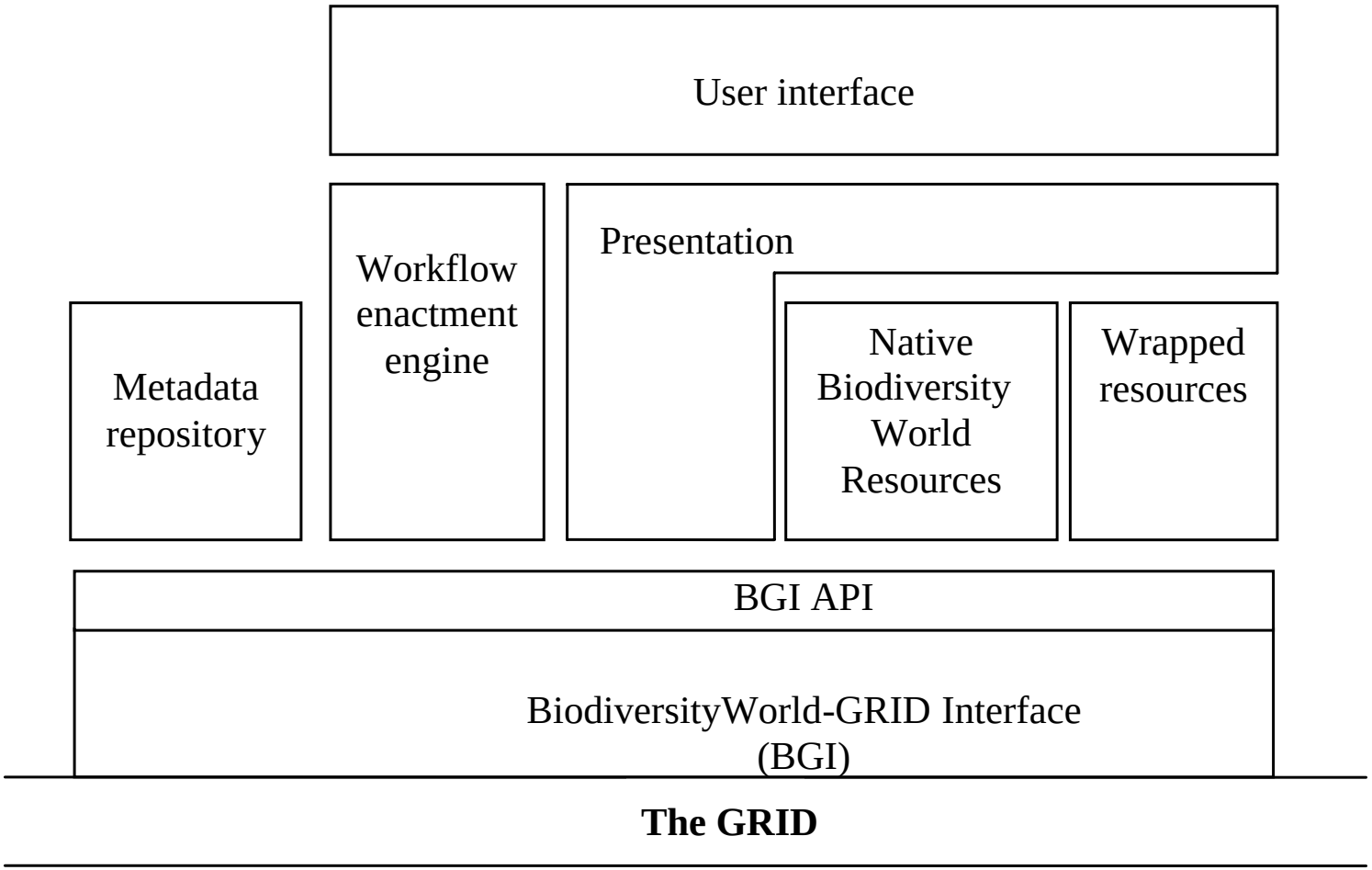


BDWorld-Grid Interface (BGI) Layer

- Provides standard mechanisms for invoking operations on heterogeneous resources
  - provides an integrated mechanism for seamless access to BDW resources via resource wrappers
  - Uses XML/SOAP messaging system for invoking operations on resource wrappers
  - Potentially interoperable with other e-Science projects
- Isolates users from Grid/Web Service complexities
  - Isolates resources/ resource wrapper implementation to enable use of web services/grid technologies as part of a separate layer
  - A Helper class is provided to the user (or software such as Triana) for using the BGI layer


Unit A
Unit B
Unit C
Triana Workflow Units
Data Handling Tools
BGI Helper Tool
BGI Communications Layer
BDW Datatypes
BDW Datatypes
BDW Datatypes
BDW Wrapper
BDW Wrapper
BDW Wrapper
Remote Database
Resource Tool
Java Resource

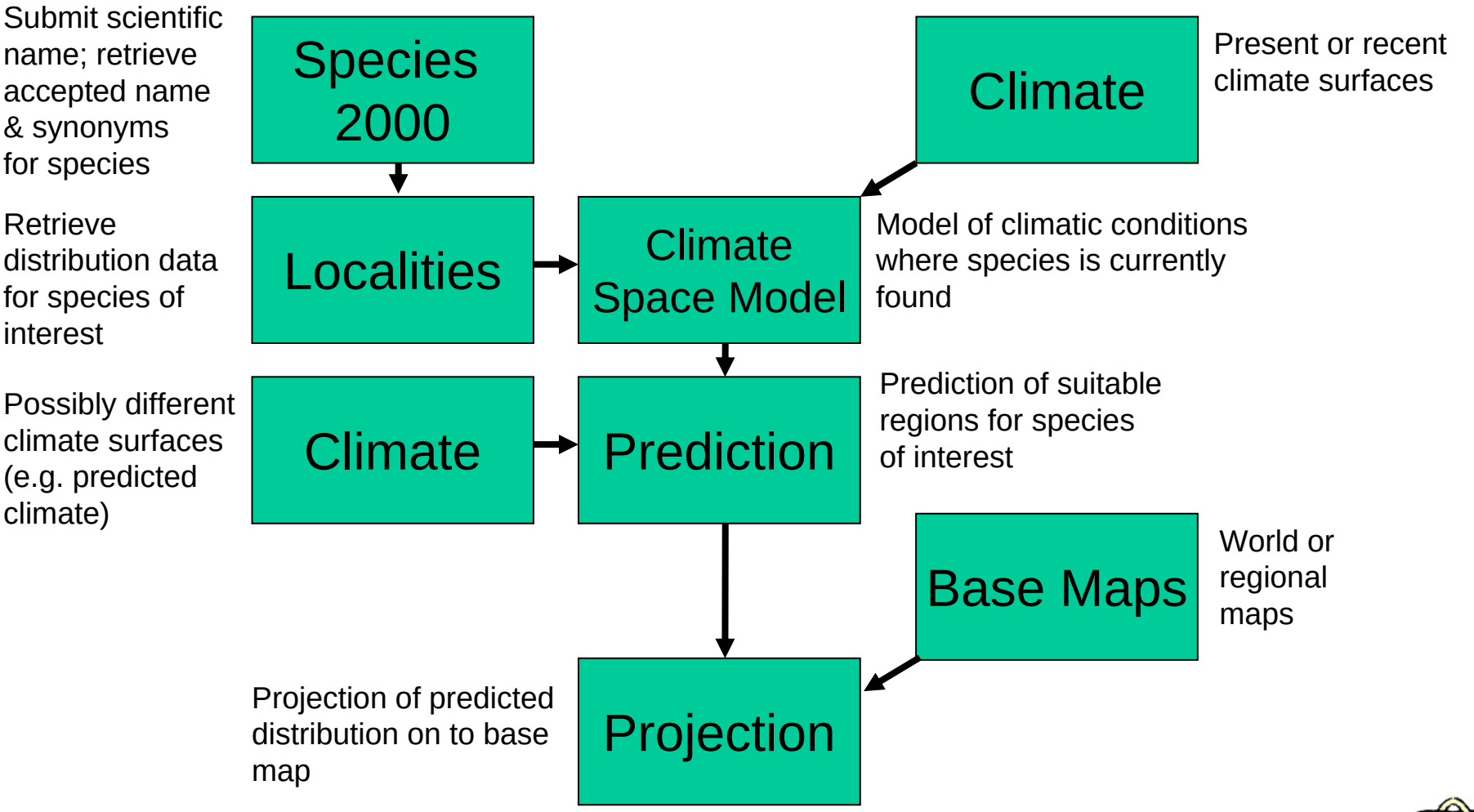
Biodiversity World architecture



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

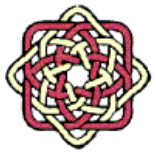
User interaction with BDWorld …

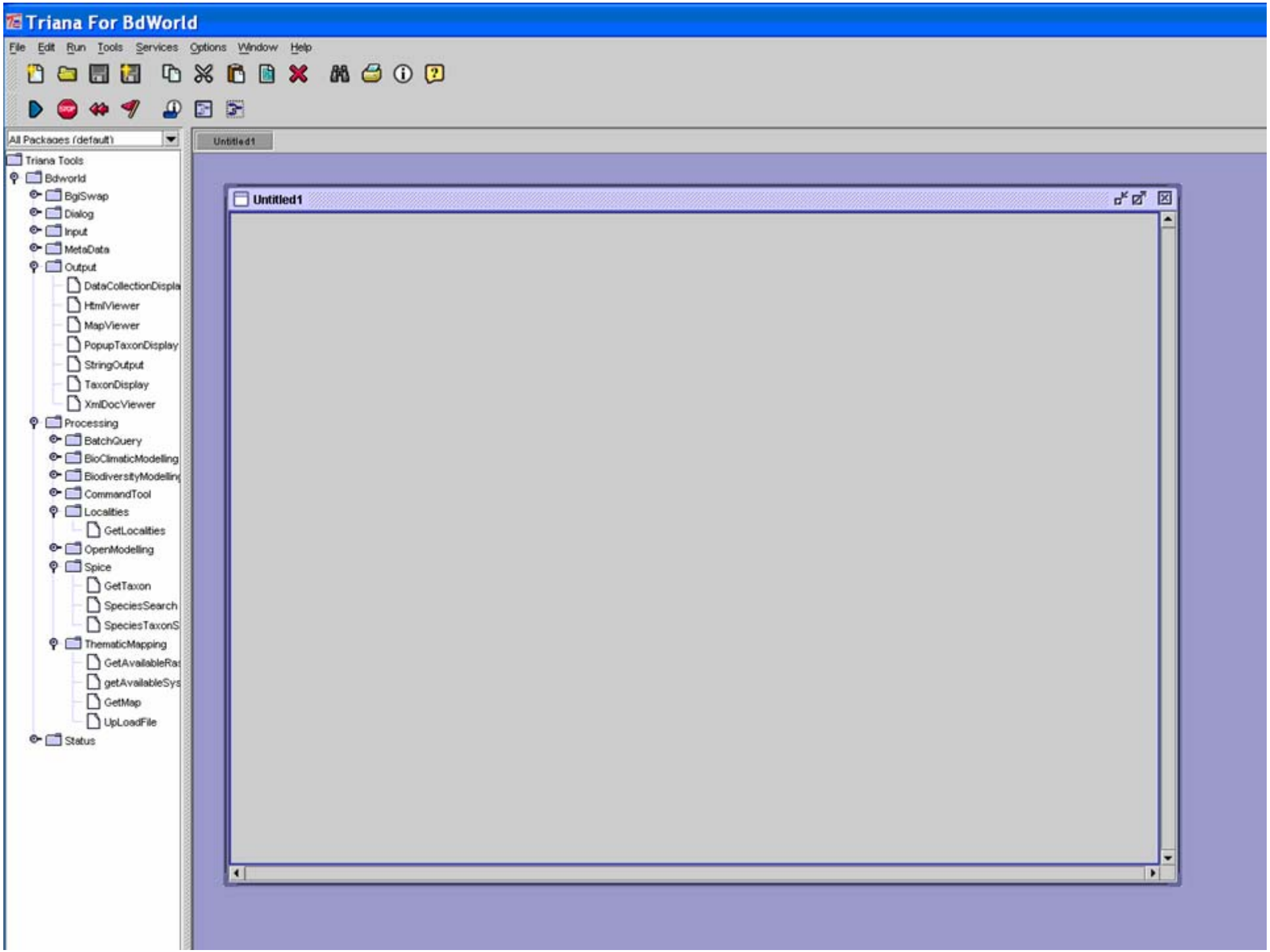
Example work-flow (Climate-space Modelling)

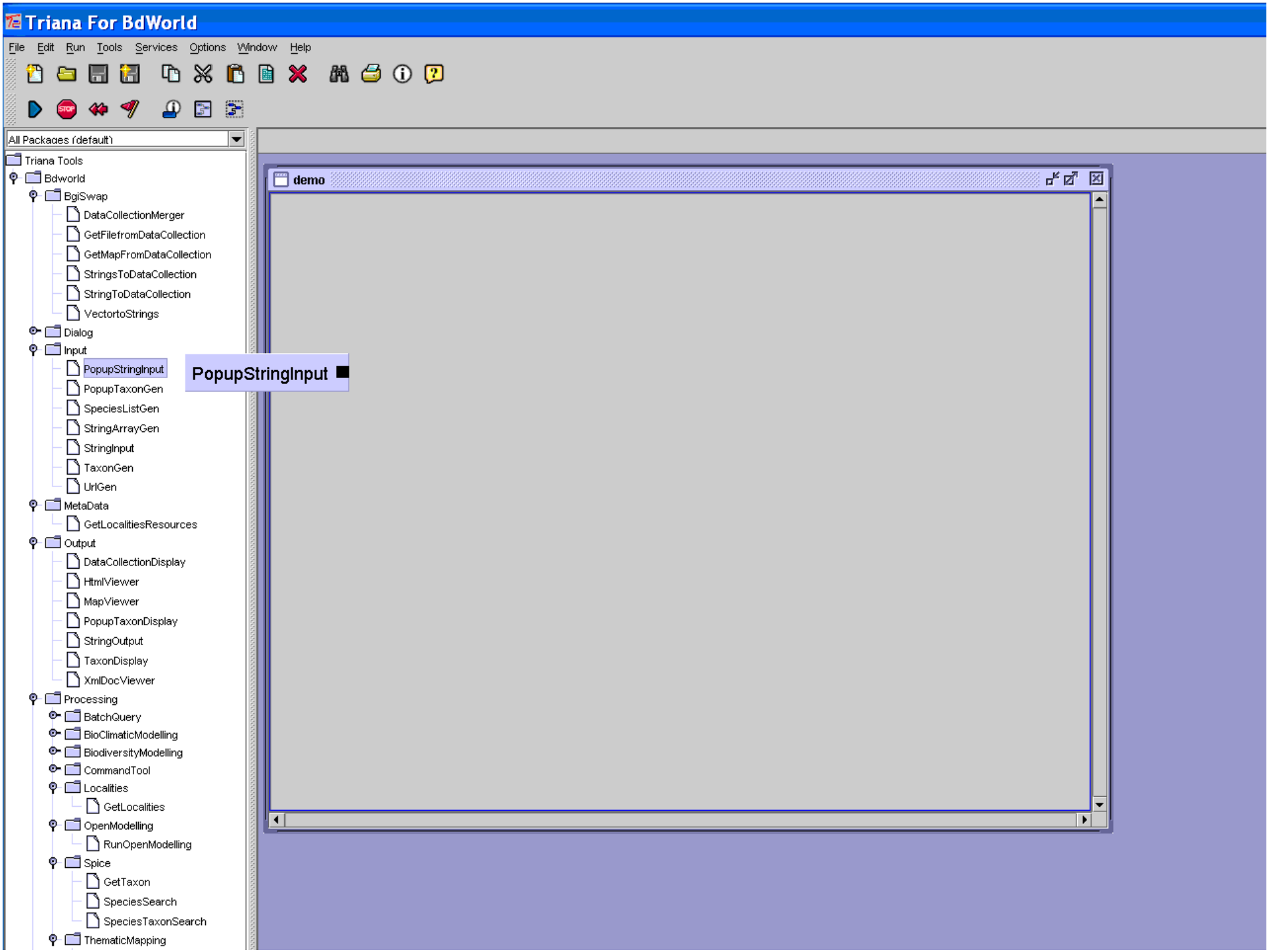


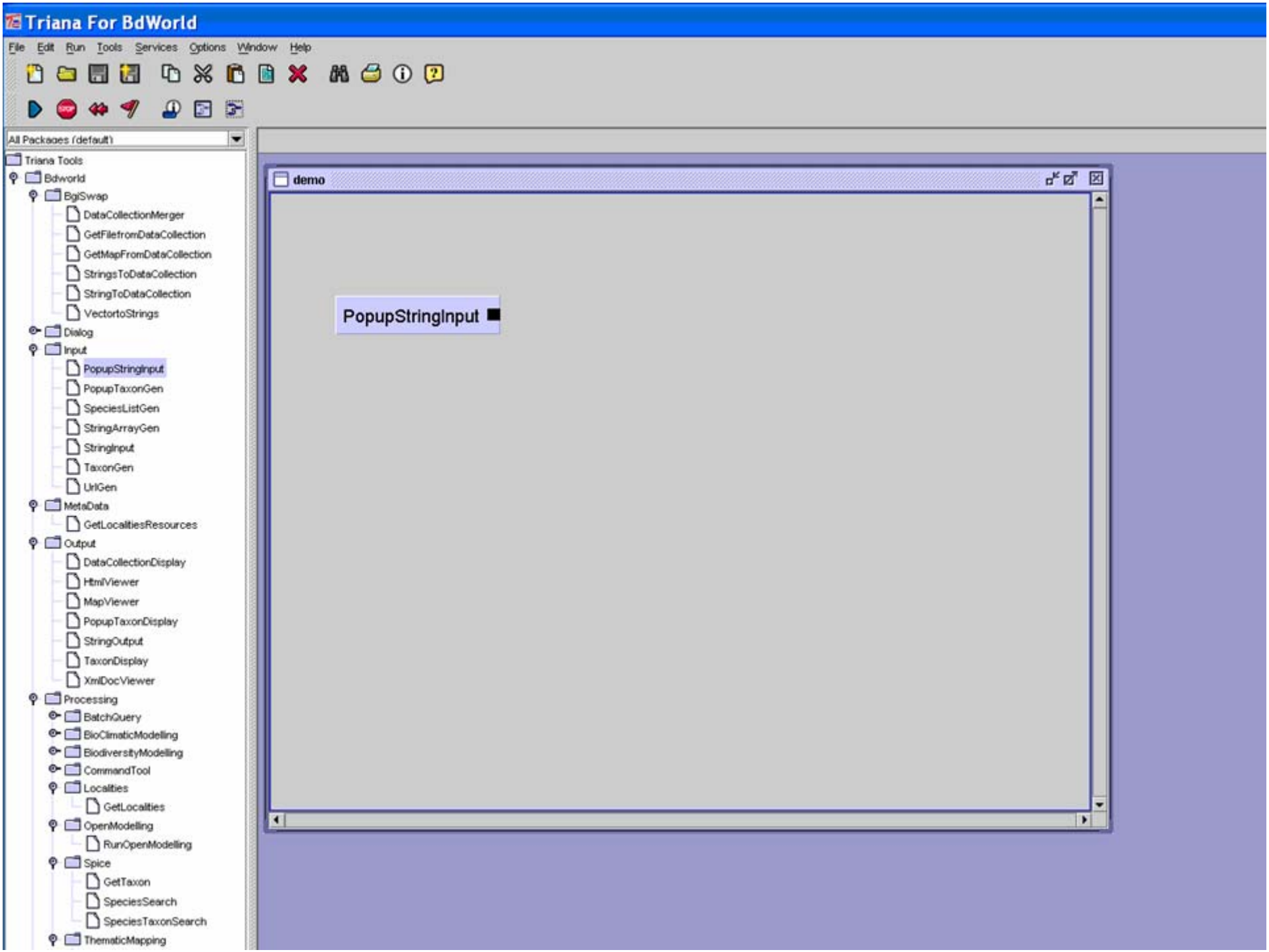
BDWorld / Triana in operation:

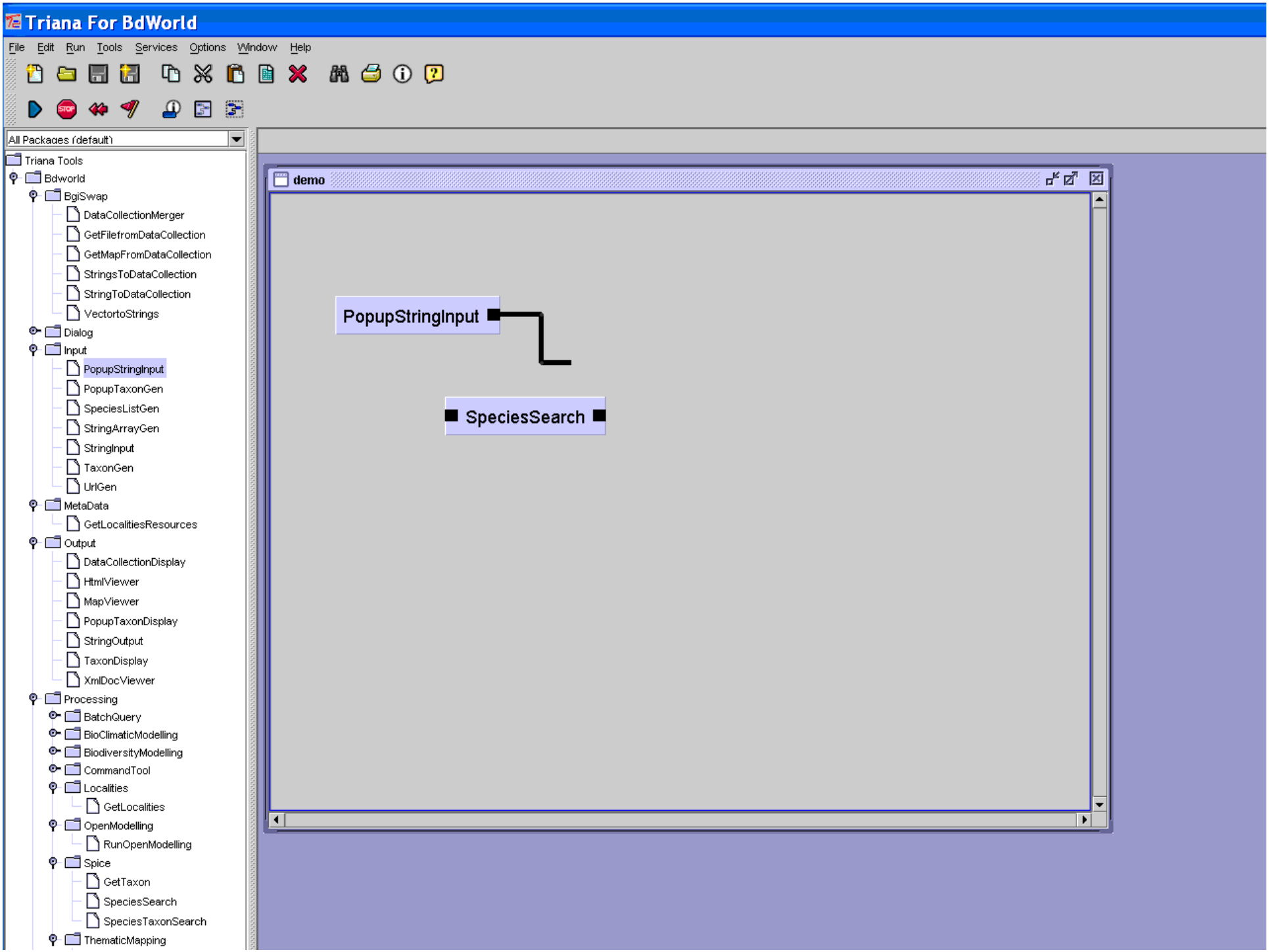
Workflow creation (design, editing)

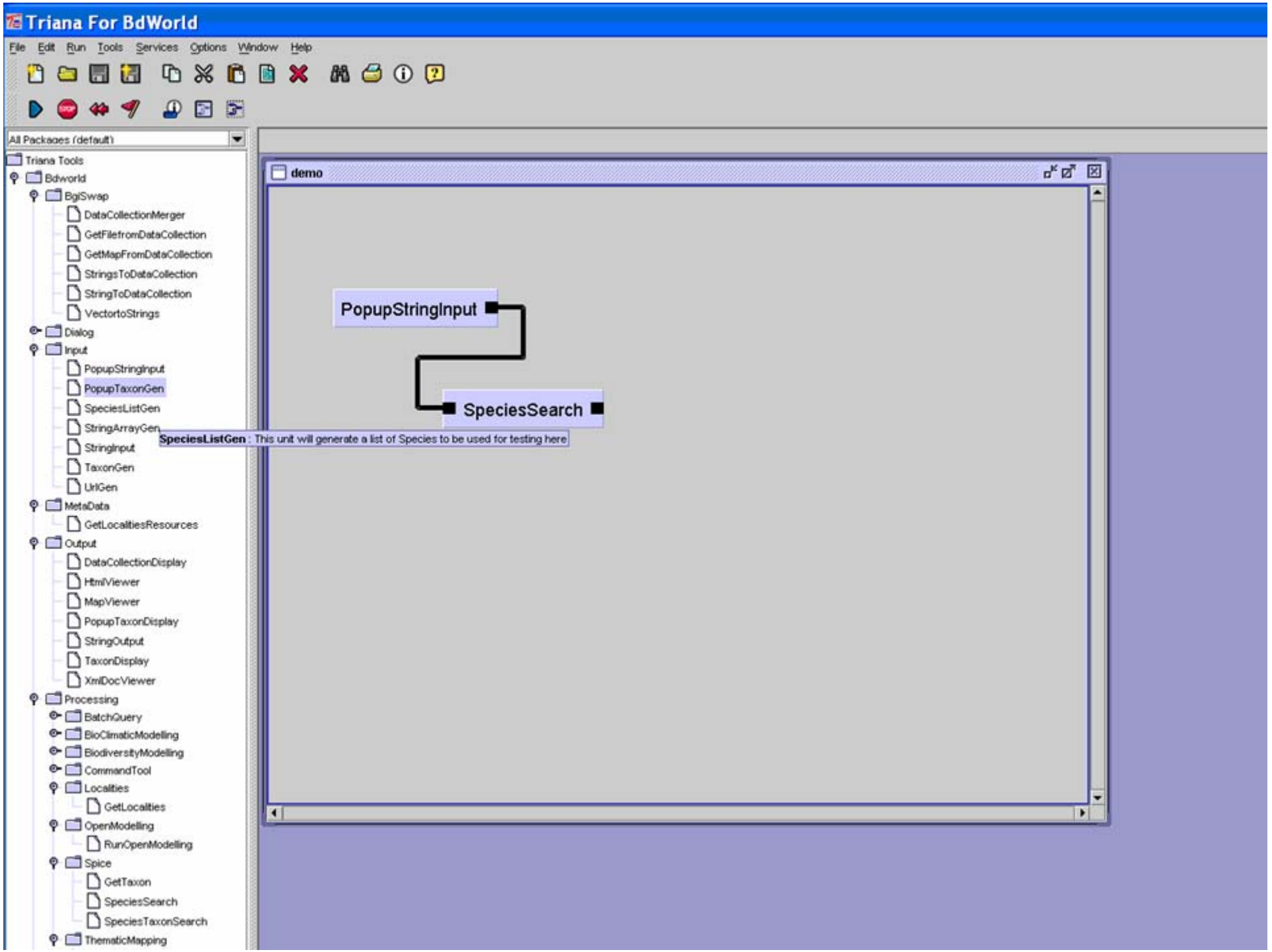


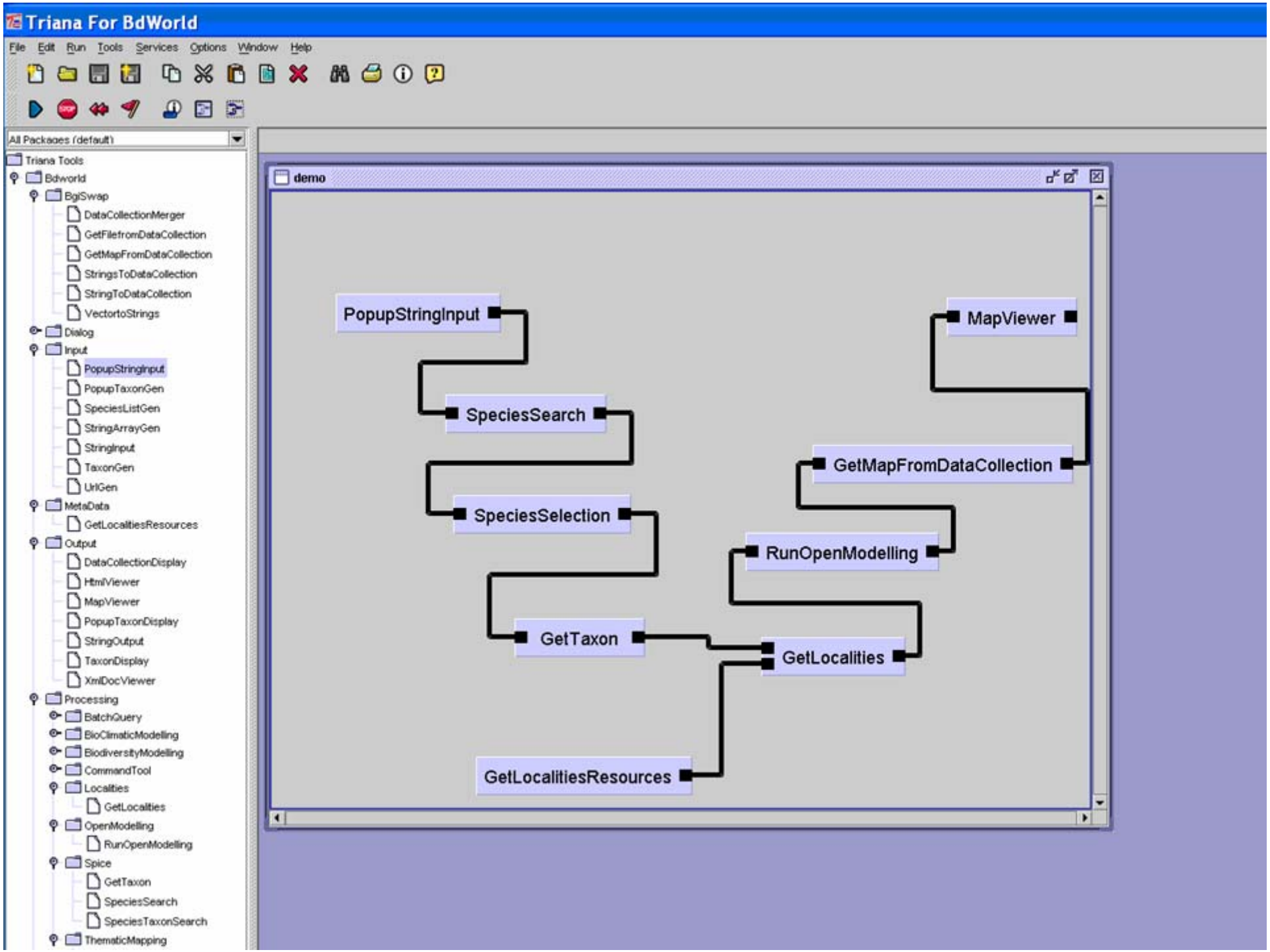

Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
Dialog
Input
MetaData
Output
DataCollectionDispla
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModellin
CommandTool
Localities
GetLocalities
OpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonS
ThematicMapping
GetAvailableRa
getAvailableSys
GetMap
UpLoadFile
Status
Untitled1
Untitled1


Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput


Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput


Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput
SpeciesSearch

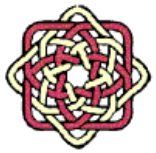

Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput
SpeciesSearch
SpeciesListGen : This unit will generate a list of Species to be used for testing here

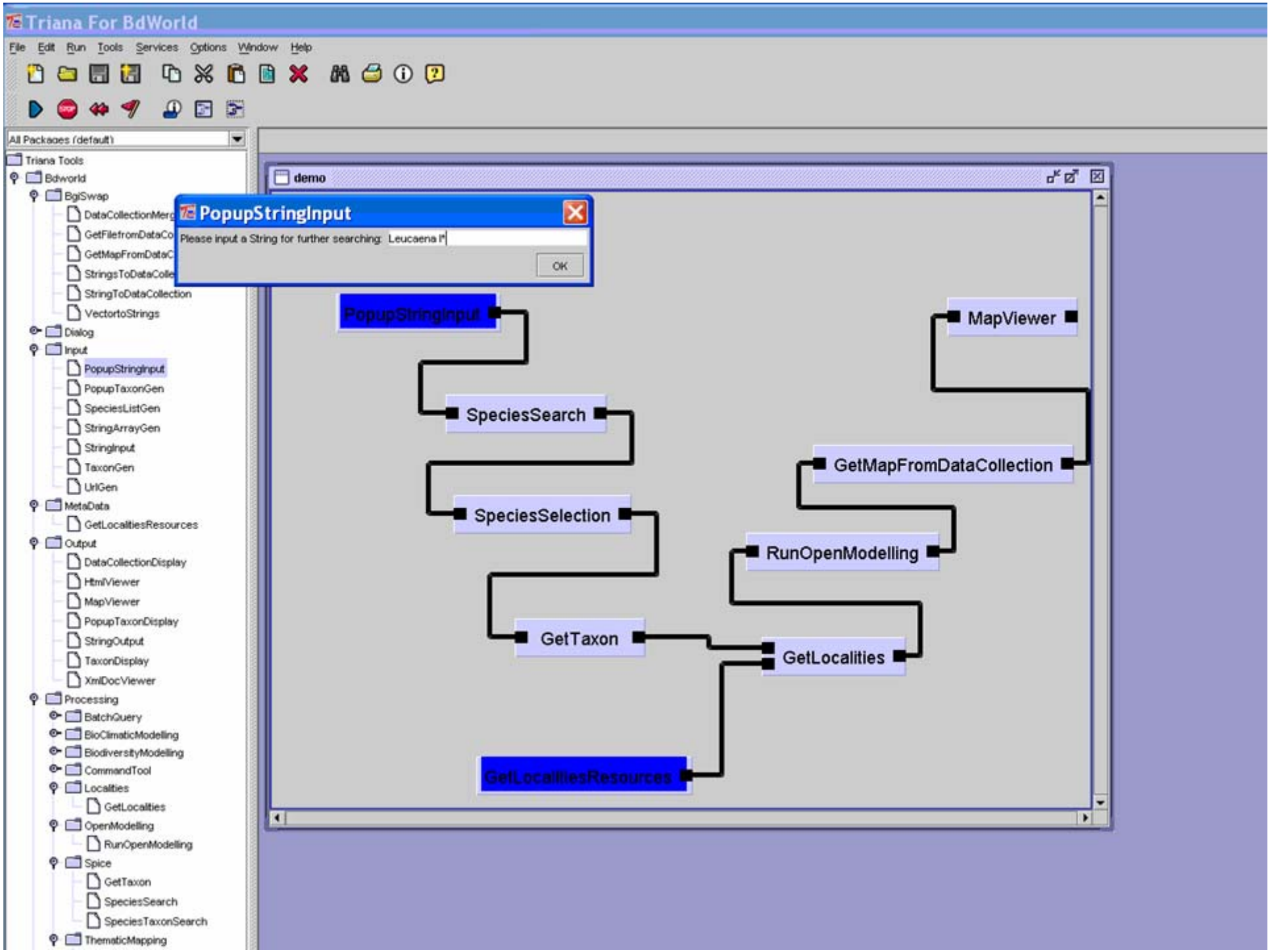

Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput
SpeciesSearch
SpeciesSelection
GetTaxon
GetLocalitiesResources
MapViewer
GetMapFromDataCollection
RunOpenModelling
GetLocalities

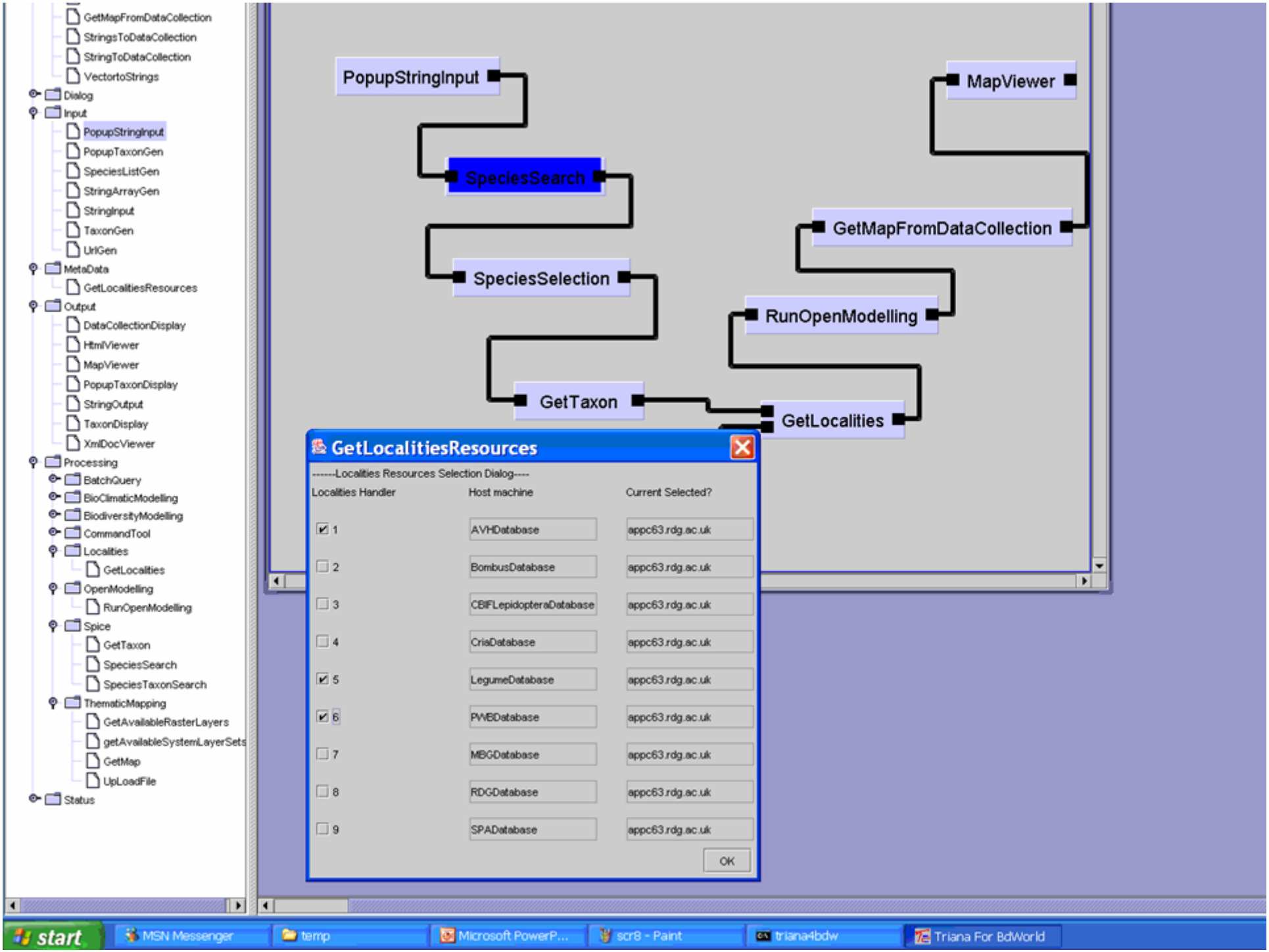
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

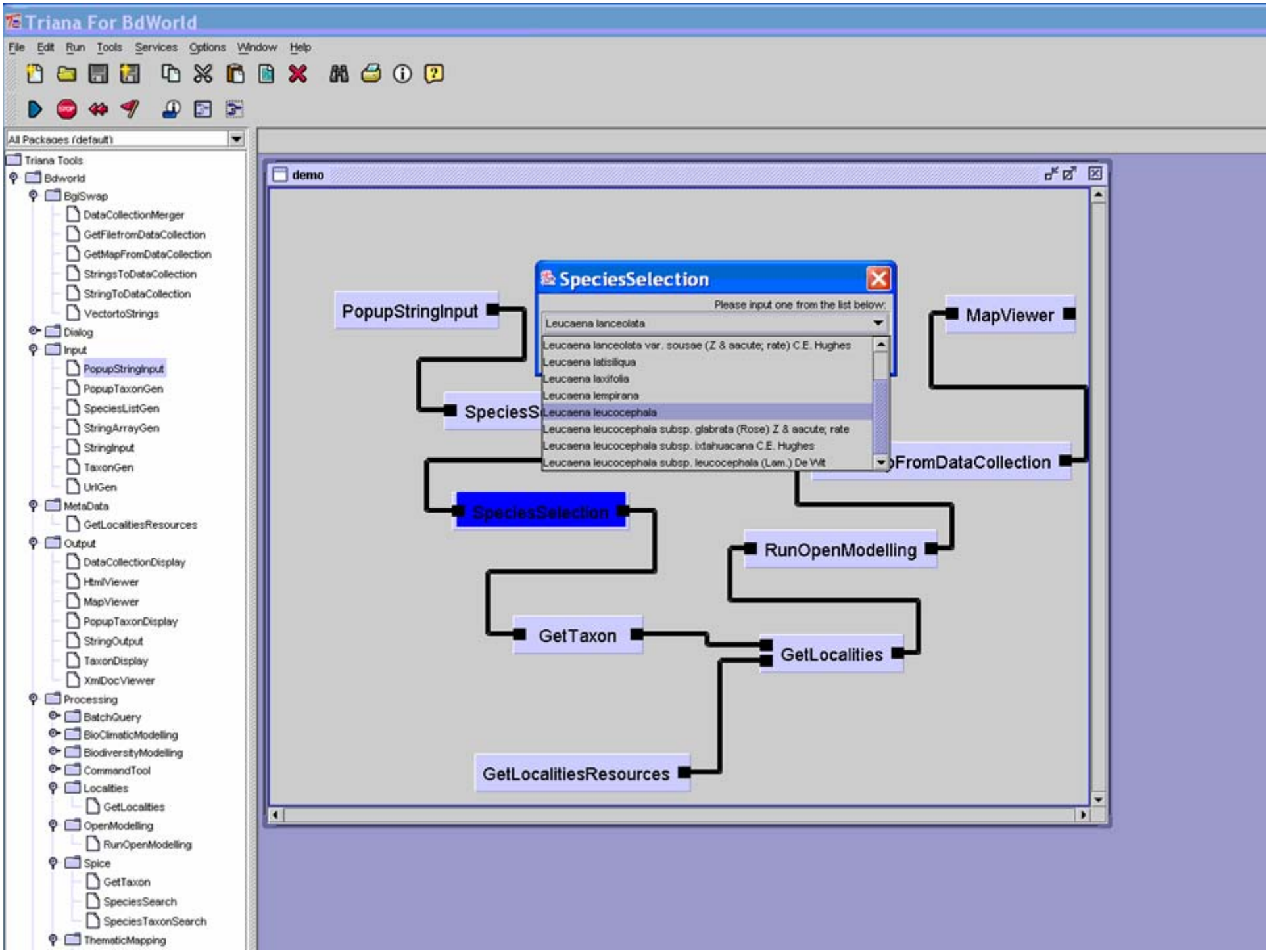
BDWorld / Triana in operation:

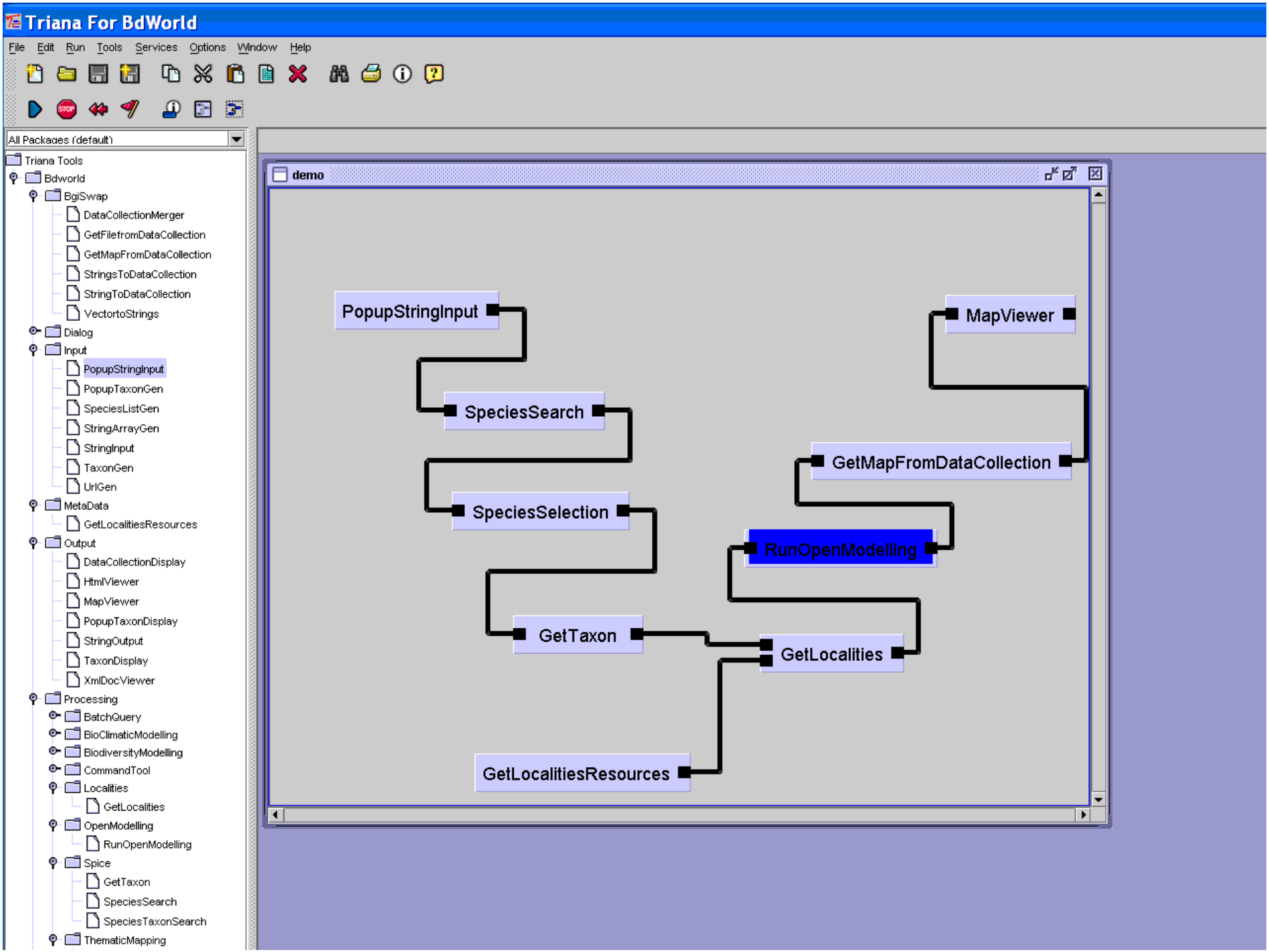
Workflow execution (enactment, run-time)

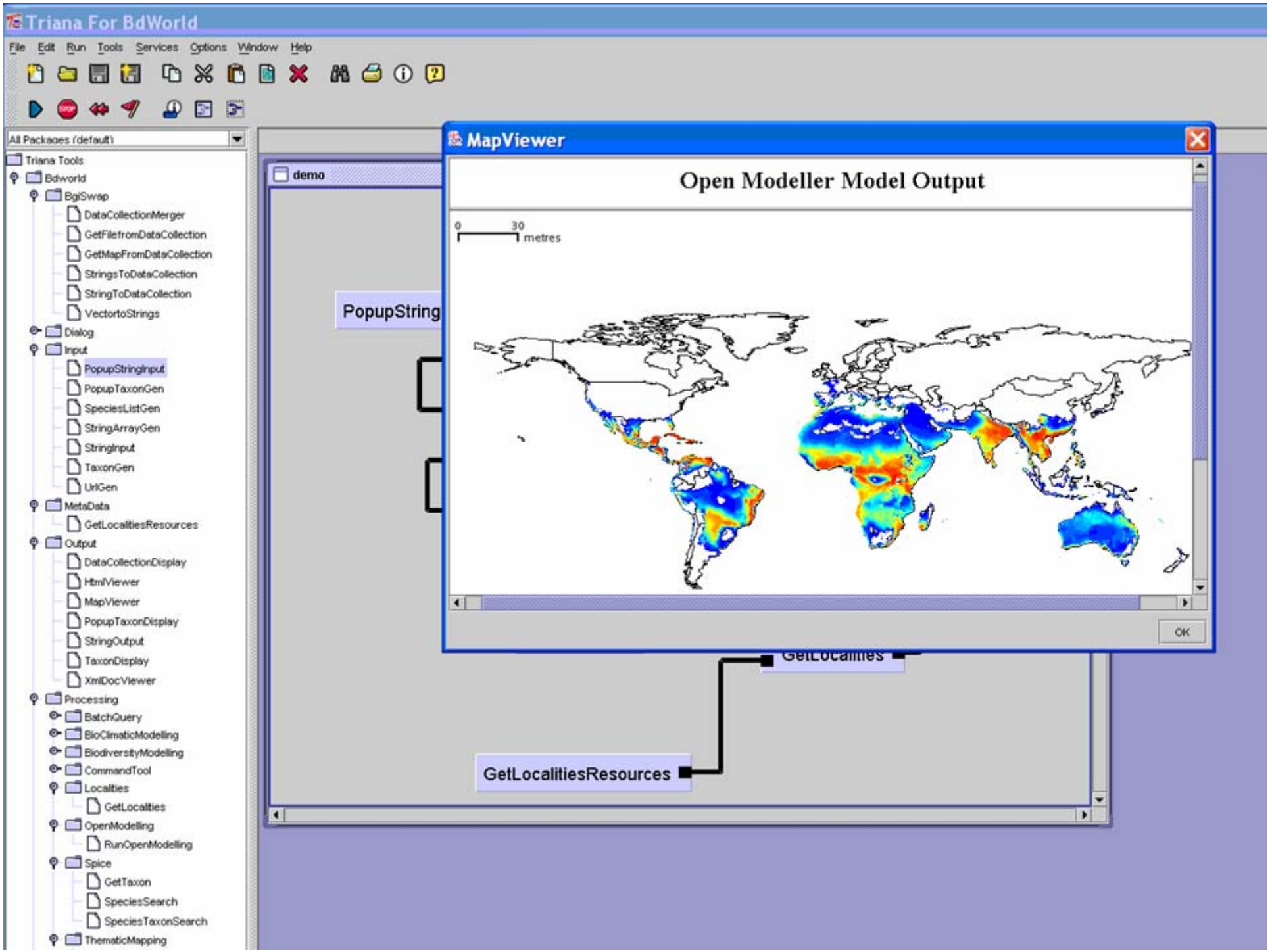



Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerg
GetFilefromDataCo
GetMapFromDataC
StringsToDataColle
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput
Please input a String for further searching: Leucaena l*
OK
PopupStringInput
SpeciesSearch
SpeciesSelection
GetTaxon
GetLocalitiesResources
MapViewer
GetMapFromDataCollection
RunOpenModelling
GetLocalities


GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
GetAvailableRasterLayers
getAvailableSystemLayerSets
GetMap
UpLoadFile
Status
PopupStringInput
MapViewer
SpeciesSearch
GetMapFromDataCollection
SpeciesSelection
RunOpenModelling
GetTaxon
GetLocalities
GetLocalitiesResources
------Localities Resources Selection Dialog----
Localities Handler
Host machine
Current Selected?
1 AVHDatabase appc63.rdg.ac.uk
2 BombusDatabase appc63.rdg.ac.uk
3 CBIFLepidopteraDatabase appc63.rdg.ac.uk
4 CriaDatabase appc63.rdg.ac.uk
5 LegumeDatabase appc63.rdg.ac.uk
6 PWEDatabase appc63.rdg.ac.uk
7 MBGDatabase appc63.rdg.ac.uk
8 RDGDatabase appc63.rdg.ac.uk
9 SPADatabase appc63.rdg.ac.uk
OK
start
MSN Messenger
temp
Microsoft PowerP...
scr8 - Paint
triana4bdw
Triana For BdWorld

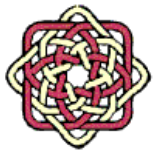

Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
SpeciesSelection
Please input one from the list below:
Leucaena lanceolata
Leucaena lanceolata var. sousae (Z & aacute; rate) C.E. Hughes
Leucaena latisiliqua
Leucaena laxifolia
Leucaena lempirana
Leucaena leucocephala
Leucaena leucocephala subsp. glabrata (Rose) Z & aacute; rate
Leucaena leucocephala subsp. ixtahuacana C.E. Hughes
Leucaena leucocephala subsp. leucocephala (Lam.) De Wit
PopupStringInput
MapViewer
SpeciesS
FromDataCollection
SpeciesSelection
RunOpenModelling
GetTaxon
GetLocalities
GetLocalitiesResources


Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupStringInput
SpeciesSearch
SpeciesSelection
GetTaxon
GetLocalitiesResources
MapViewer
GetMapFromDataCollection
RunOpenModelling
GetLocalities


Triana For BdWorld
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Bdworld
BgiSwap
DataCollectionMerger
GetFilefromDataCollection
GetMapFromDataCollection
StringsToDataCollection
StringToDataCollection
VectortoStrings
Dialog
Input
PopupStringInput
PopupTaxonGen
SpeciesListGen
StringArrayGen
StringInput
TaxonGen
UrlGen
MetaData
GetLocalitiesResources
Output
DataCollectionDisplay
HtmlViewer
MapViewer
PopupTaxonDisplay
StringOutput
TaxonDisplay
XmlDocViewer
Processing
BatchQuery
BioClimaticModelling
BiodiversityModelling
CommandTool
Localities
GetLocalities
OpenModelling
RunOpenModelling
Spice
GetTaxon
SpeciesSearch
SpeciesTaxonSearch
ThematicMapping
demo
PopupString
GetLocalities
GetLocalitiesResources
MapViewer
Open Modeller Model Output
0
30
metres
OK

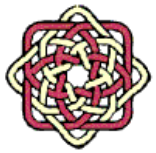
Current and future work

- What I have described up to now is more or less what was originally envisaged
- Now we have some ideas on how to improve our architecture
  - Web Services version being evaluated
  - GT4, WSRF in future



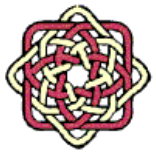
BDWorld Web Services Architecture

- Web Services is a mechanism of enabling distributed computing based on open standards
- Wrappers are now deployed in a Web Services environment which can be accessed via the BGI Layer with the assistance of a BGI Helper Tool
  - Axis SOAP engine provides the WSDL that exposes wrapper operations to outside world
- The MetadataAgent provides access to MDR via the BGI Layer



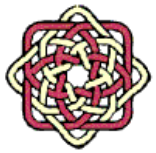
Drawbacks of Web Services

- Each web service needs to be deployed individually
- Web services are not “stateful”
  - provide mechanisms for invoking remote operations
  - but no provision for other functionality such as resource management, persistence, life cycle management, notification etc.



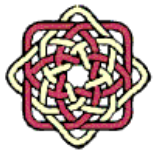
GT4 Key Concepts

- Based on Open Grid Service Architecture (OGSA)
  - OGSA defines common, standard and open architecture for Grid-based applications
- Standardises various services common to Grid applications (job management, resource monitoring and discovery, resource management, security services etc)
- Uses Web Services as underlying technology to enable distributed computing
  - But Web Services are not stateful



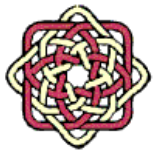
WSRF – An approach to statefulness

- WSRF provides the mechanism to keep state information by keeping the Web Service and state information completely separate
- State information is stored in an entity called a resource (not to be confused with a BDWorld resource)
  - A resource can be identified via its unique key
  - When requiring stateful interaction, a web service can be instructed to use a particular resource
  - The resources can be stored in memory or on secondary storage



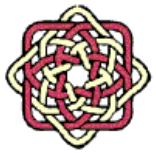
Where do we go from here?

- Present system is a proof of concept
  - Limited
    - Biodiversity exemplars only
  - Needs
    - more data resources
    - more functionality
    - additional features
      - Modelling tools
      - Virtual organisations



Workflows

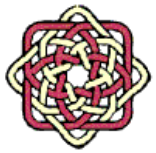
- Creating a workflow:
  - Workflows clearly good for capturing complex tasks
    - Good for ‘tweaking’ tasks
  - But is this how users think?
  - If not, we should provide an environment that supports a more exploratory approach too, e.g.
    - User tries out some small subtasks …
    - … and joins results together
    - System records interactions, so re-usable workflows can be composed



Other aspects of user interface

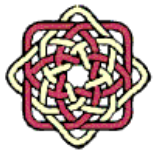
The drag-and-drop metaphor needs further research into the best ways to support

- resource discovery
- resource matching
- data management (e.g. temporary storage of intermediate results)



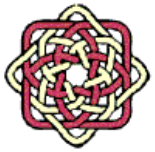
Complex interactions

- BGI not well-suited to fine-grained interaction
- Stand-alone applications difficult to wrap
  - may need, e.g., screen scraping
- We're looking at:
  - Less portable 'by-pass' mechanisms, e.g.
    - New BGI protocol
    - Existing techniques *(in extremis)* e.g. VNC
  - Plug-ins for the BDW client
  - External tools
    - (which will always be needed)



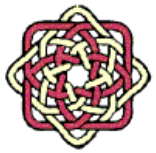
A dream

- A desktop environment in which scientists can "drag & drop" data sources, analysis and modelling tools and visualisation interfaces into a desired sequence of operations which can be run automatically
  - BDWorld just about at this stage
  - With additional features, the environment could be made richer, more productive, and support research groups.
  - Essentially a *component-based visual programming environment*
  - Not just for biodiversity!



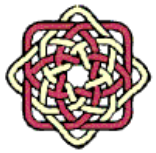
Extra functionality

- Enhanced metadata
  - Provenance and data lineage
  - Automatic electronic “lab notebook”
- Stored workflows
  - Repeatability, reproduceability
  - Re-use with different data, changed parameters
- Ontologies
  - Resource discovery
  - Usability
- Dynamic interaction of users with resources



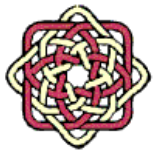
Virtual organisations

- Collaborative working environments
- Shared and private resources: data, tools
  - Controlled release of data, tools and results
- Shared experimentation
  - User authorisation / authentication
  - Access control
- Dynamic
  - Membership
  - Resources



The way forward

- New exemplars in environmental science, bioinformatics and health informatics
- Links with national and international organisations, resources, VOs
- “End users”
  - Input
  - Feedback
  - Applied use, driven by scientific priorities
- …



Acknowledgements

- Thanks to Jaspreet Singh Pahwa for the slides concerning wrappers, BGI, GT4, OGSA &WSRF
- The Triana Project for the workflow environment
- Other collaborators at
  - Cardiff University
  - The University of Reading
  - The Natural History Museum (London)
- Organisations that have co-operated with these research projects, especially
  - Species 2000
  - ILDIS (International Legume Database and Information Service)
  - Hadley Centre for Climate Prediction and Research
- BBSRC for BDWorld
- DTI, EPSRC & EU for related projects