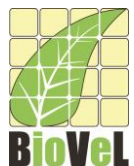




Introduction to myGrid

Aleksandra Pawlik
myGrid Team
University of Manchester

VLIZ, 2014-10-06 / 2014-10-08
<http://www.taverna.org.uk/>



*This work is licensed under a
[Creative Commons Attribution 3.0 Unported License](http://creativecommons.org/licenses/by/3.0/)*





Overview

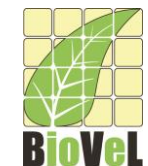
- Taverna
- BiodiversityCatalogue
- myExperiment



What is myGrid?

- “... a product of the e-Science initiative in the early 2000s to seize on the advantages of [grid] computing and translate them into the biological sciences.”
- Mixed team of developers, domain scientists and researchers
- Building technical infrastructure and tools supporting the ***in silico* experiments**
- Various domains: bioinformatics, biodiversity, systems biology, social science, astronomy, chemistry, ...
- Open source

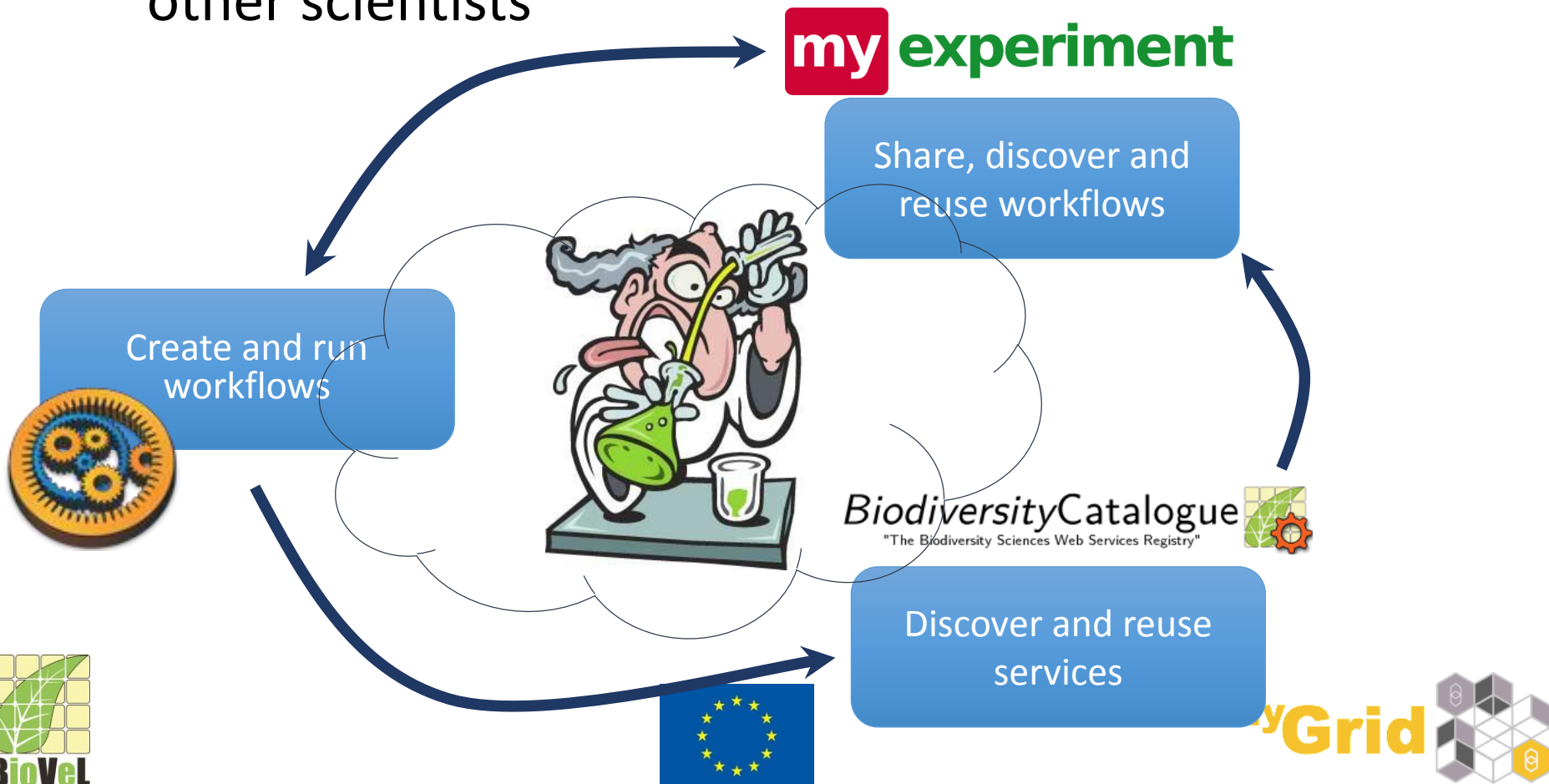
Software | Services | Content | Skills | Community





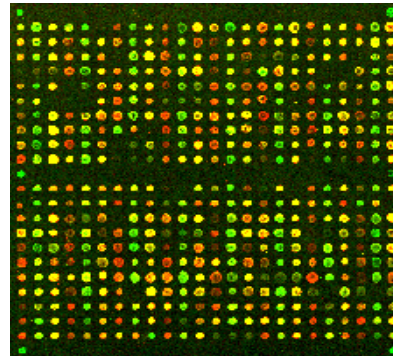
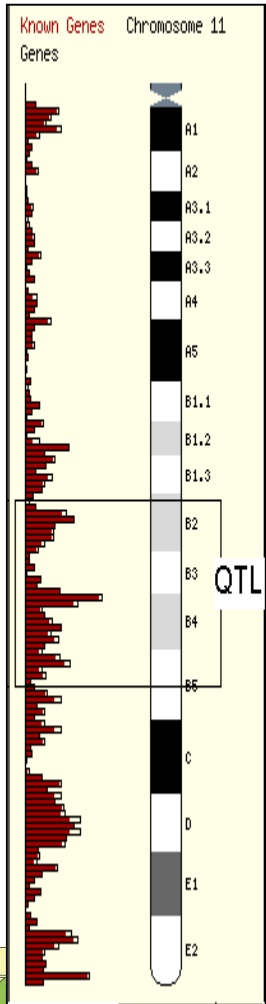
(Selected) myGrid Tools

- “ Help scientists get on with science and get on with other scientists”





A Problem: Huge Amounts of Data



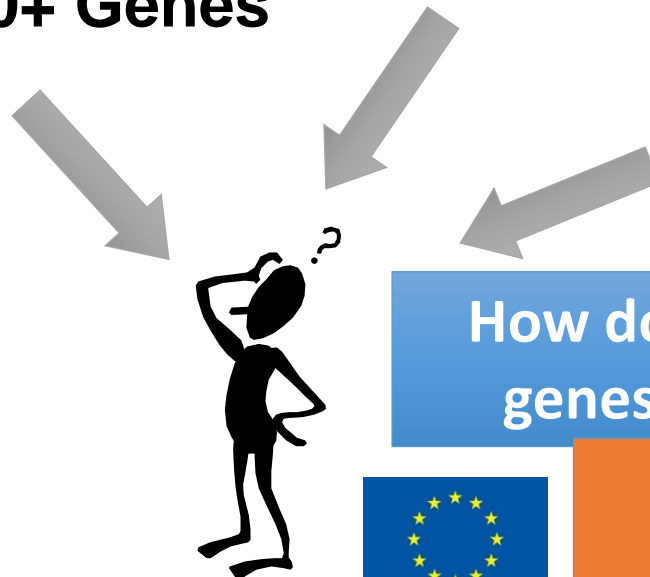
Microarray
1000+ Genes



Next Generation
Sequencing

**10,000+
Genes**

QTL regions
100+ Genes



How do I look at ALL the
genes systematically?

How do I identify
candidate genes?



(Some) Issues with Current Approaches

- Scale of analysis task overwhelms researchers – **large amounts of data**
 - Cannot use full data sets
 - Cannot rerun experiments
- Constant change in data - need for reruns of experiments (re-analysis) to check for new and updated information
- Scientists often have to manually click through various Web sites and pass their data
- Incompatible data formats
- Error proliferation from any of the listed issues – notably human error

Solution: Automate



One Solution - Workflows

- General technique for explicit describing and executing a process
- Describes **what** you want to do, including the **services** to use
- Automation of scientific process enables
 - Easier rerun
 - Parameter tweaking
 - Easier comparison of results
 - Science reproducibility



Creation and execution of workflows



Taverna



Taverna

- Workflow management system for creating and executing workflows
- Allowing to connect various data sources and services processing them
- A bit of history:
 - First released in 2004
 - Currently Version 2.5
 - 3rd generation in alpha
- Freely available, open source (LGPL license)
 - Windows, Mac OS and Linux
- Extensible – allowing for collaboration and new functionalities
- Domain specific versions, including one for biodiversity
- <http://www.taverna.org.uk>



Taverna Workflows

- Sophisticated automated analysis pipelines for **data-driven** research or processes
 - Define **how** you want your data to flow
- Facilitate **data flow** links between the services:
 - Output data of service A is the input of service B
- Service **types**: data resources, analysis tools, knowledge resources
- Service **location**:
 - Remote services: Web services or scripts on remote machines
 - Local services: scripts and libraries on users' machines

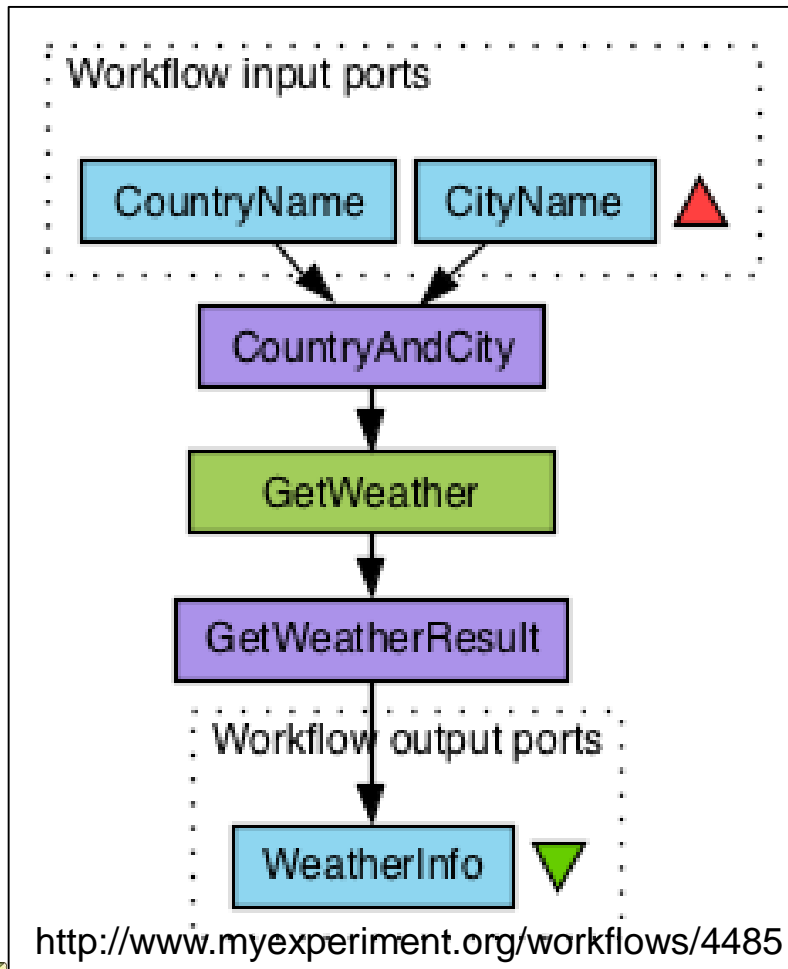


Taverna Workflows – continued

- Advanced flow control:
 - **List handling/iteration** strategy for data inputs
 - **Loops** (until a condition is met)
 - **Control links** to determine execution order
- Help overcome problems with service interoperability and integration
 - Shim services



Example Taverna Workflow



- Get the weather forecast of the day given the city and the country
- Green box is a Web service
- Purple boxes are local XML services to assemble/extract XML
- Blue boxes are workflow input and output ports
- Arrows define the direction of data flow

Supported Services

- SOAP/WSDL Web services
- REST Web services
- Interaction
- R statistical services
- Local or remote (via ssh) scripts:
 - Python, Perl, Java, ...
- Spreadsheet data import
- XPath and text manipulation services
- WebDav data management services
- BioMart
- and more!

- Third party extensions





Who Provides the Services?

- Open domain services and resources
- Third party
- We do not own them
- We did not build them
- We enforce no common data model
- You can include your own services and resources too!



Data and Provenance Collection

- Collection of workflow run provenance – **who, when, where**
- Workflows can generate vast amount of data – Taverna can help manage and track:
 - Data, metadata, and workflow provenance
- Allows checking back over past results, comparing workflow runs and sharing workflow runs with colleagues
 - Facilitates collaboration and reproducibility in science
- Inspecting intermediate results when designing and debugging workflows
 - Test with small dataset before running over a million of datasets

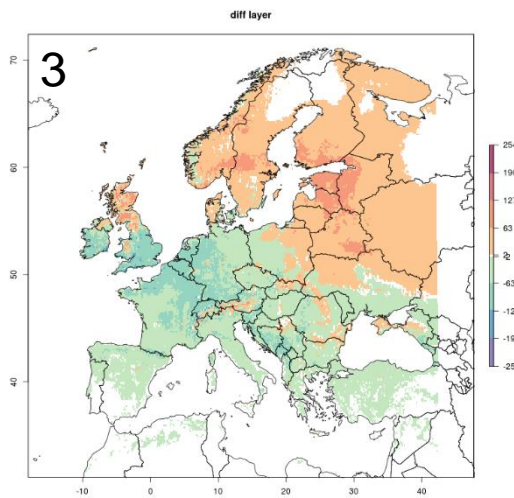
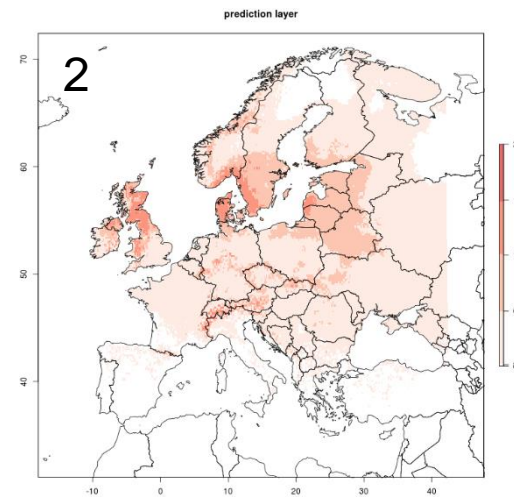
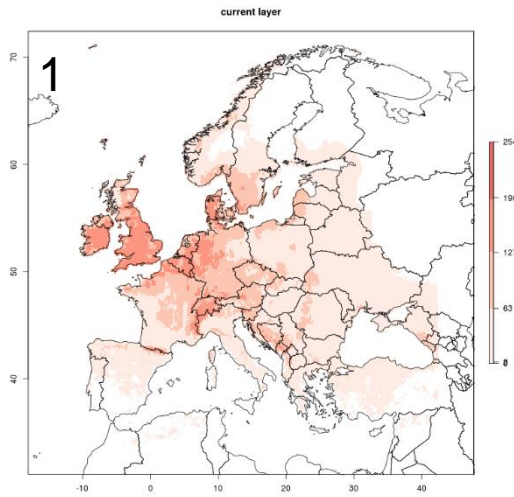


Success Story: Essential Biodiversity Variables for butterflies

- Wide ranging experiment for **all Butterflies in Europe** – 624 species including synonyms
- Carried out by Yuliya Fetyukova and Hannu Saarenmaa from University of Eastern Finland
- Work carried out using workflows created by BioVeL project and run on the BioVeL portal
- Users interact with the workflow runs on the portal



Niche projection



Erynnis tages

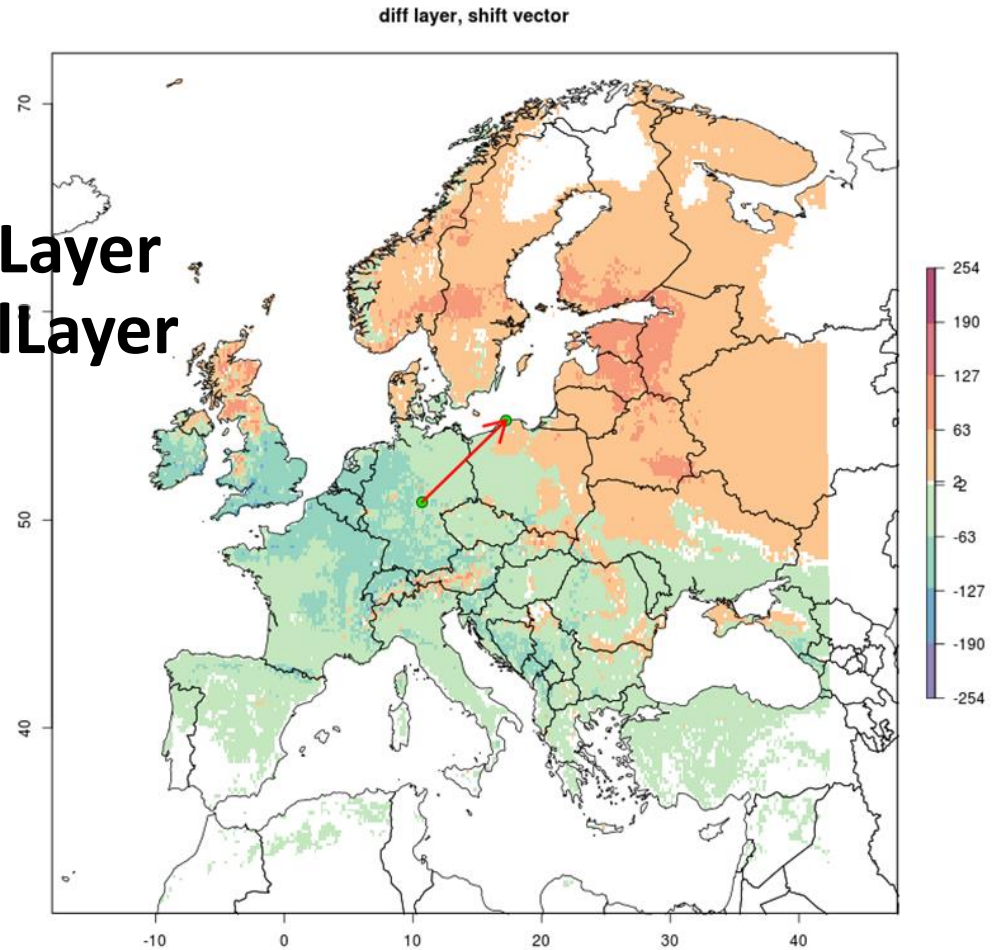
1. 1960 – 1990 projection
2. 2050 projection
3. Difference Layer

Red – increase
Green - decrease



Niche shift

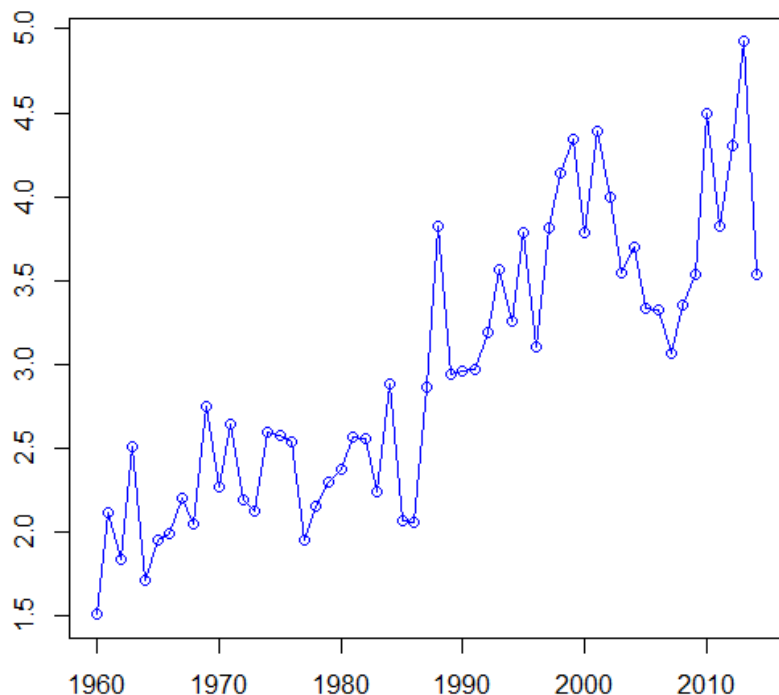
- *Erynnis tages*:
- $\text{diffLayer} = \text{predictionLayer} - \text{historicalLayer}$
- Shift: 618.61 km



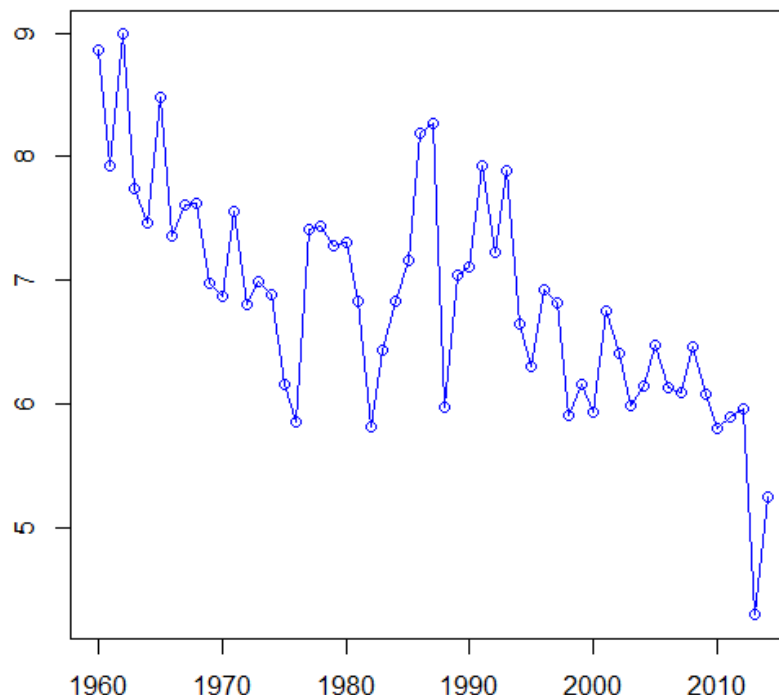


Relative abundance

- *Pieris napi*



- Hesperiiidae





Taverna Summary

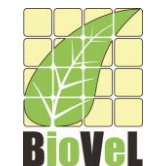
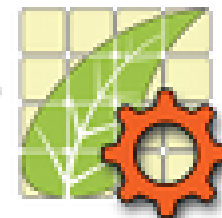
- Data Analysis Pipelines
- Machinery for coordinating the execution of services and linking together resources
- Repetitive and mundane boring stuff made easier



Discover and reuse services

*Biodiversity*Catalogue

"The Biodiversity Sciences Web Services Registry"





BiodiversityCatalogue

- A public centralised and curated registry of Biodiversity Web services that can be used in workflows
- Allow anyone to register, discover and curate Web services
- Community oriented with expert guidance
- Open content
- Open source (BSD license)
- <http://www.biodiversitycatalogue.org>



Motivation

- Guesstimate: many publicly available on-line services for biodiversity
- **Where**
 - can I find them? advertise them?
- **What**
 - do they do? can I use them for?
 - is the cost? the licenses?
- **How**
 - do they work? up to date are they?
- **Who**
 - provides them? uses them?
 - recommends them?





Google is good, but ...

- Unified way of describing services
 - How to invoke a service
 - Input and output data examples
 - Documentation
 - Who else is using this service
 - How reliable a service is
 - Monitoring history
 - Service variants, deployments



What Can BiodiversityCatalogue Users do?

	Non-registered user	Registered user	Curator
Search	✓	✓	✓
Register a service	✗	✓	✓
Receive notifications	✗	✓	✓
Add annotations	✗	✓	✓
Manage all services	✗	✗	✓



Benefits for Service Providers

- Free!
- Easy to REGISTER services
- Great exposure
 - The right demographic
 - Instantly searchable/discoverable
- Notifications
 - When your service goes down
 - When someone 'likes' your service
 - When someone annotates your service
- Community-aided ANNOTATION



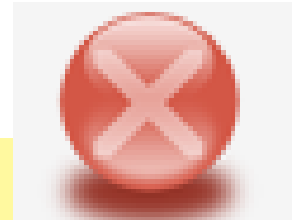
Benefits for Community Members

- **EXPLORE** Web services
 - Full text search
 - Browse
 - Filtering
- Comprehensive service descriptions
 - Annotated and verified by the community



Benefits for Community Members – cont'd

- Web service **MONITORING**
 - Services change and get outdated
 - Long term reliability
 - Testing on a daily basis



Monitoring status: **FAILED**

The last check failed

Note : Test is failing since Saturday December 08 , 2012

Last checked: about 14 hours ago



Monitoring status: **PASSED**

The last check for this test was successful

Last checked: about 14 hours ago



Various Information Feeds

- Latest services
- Latest activity/annotations
- Latest monitoring changes
- Top contributors

Subscribe to updates with an RSS reader



Latest

Activity

Monitoring

Services

Contributors

11 days ago



BioSTIF upload service has a test **change status** from UNCHECKED to **PASSED**



BioMaS has a test **change status** from UNCHECKED to **PASSED**



BiodiversityCatalogue Summary

Discover *“Web services are hard to find”*

Register *“My Web services are not visible”*

Annotate *“Web services are poorly described”*

Monitor *“Web services can be volatile”*



Share, discover and reuse workflows

my experiment



myExperiment

- Social networking site for scientists
 - “Facebook for Scientists”
 - Virtual Research Environment
- A pool of scientific workflows
- Enables scientists to **share, reuse** and **repurpose** workflows
- Helps to reduce time-to-experiment, share expertise and avoid reinvention
- Helps to build communities and form relationships
- <http://www.myexperiment.org>



myExperiment Resources

- Groups
 - myExperiment provides a central location for workflows from one community/group
- Workflows
 - Taverna
 - KNIME
 - Galaxy
 - Kepler
 - ...



myExperiment Resources – cont'd

- General files
 - Presentations
 - Data files
 - Auxiliary files
 - ...
- Packs
 - Sets of resources in myExperiment
 - Like a Research Object or a full experiment
 - Workflow, input data, results, logs, provenance, metadata, slides, paper



Workflow Sharing, Ownership and Attribution or Resources

- Fine control over privacy
- myExperiment allows you to say
 - Who can look at your workflow
 - Who can download your workflow
 - Who can modify your workflow
 - Who can run your workflow
- Workflow ownership and attribution
 - Users do not need to start from scratch – reuse or modify existing workflows
 - Attribute/credit original author



myExperiment Summary

Discover *“Find workflows suitable for your science, see what other are doing.”*

Share *“Share your and other people’s expertise”*

Reuse *“Reduce reinvention, form relationships with people working on same or similar things”*

Repurpose *“Reduce time-to-experiment, attribute original author”*



Registering on myExperiment

- Go to <http://www.myexperiment.org> and click the Register button

First time visitor? Try these videos:

- [Project Introduction](#)
- [Bioinformatics Case Study](#)

Use myExperiment to...

- [Find Workflows](#)
- [Share Your Workflows and Files](#)
- [Create and Find Packs of Items](#)
- [Find People and Make Friends](#)
- [Create and Join Groups](#)
- [Build your Profile and Reputation](#)
- [Tag and Rate things](#)
- [Write Reviews and Comments](#)

Explore

Register

or Login:

Username or Email:

Password:

Remember me:

Or use OpenID:

(eg: name.myopenid.com)

Login

[Forgot Password?](#)

About myExperiment

- [Join the Mailing List](#)
- [myExperiment Publications](#)
- [For Developers](#)
- [Give us Feedback](#)
- [The BioCatalogue Project](#)



Registering on myExperiment - 2

- Fill in the ReCAPTCHA
- It is best to register with a Username and Password
- Decide if you want to receive e-mail
- Click the grey Register button



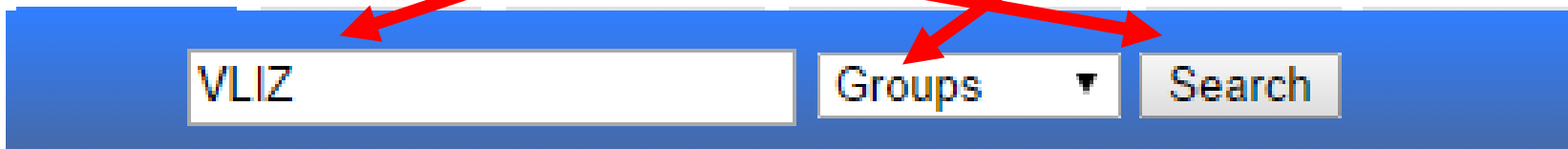
Confirming registration

- An e-mail will be sent to the account you registered with
- Open the link in the e-mail to confirm your account
- You can now log into myExperiment



Joining the workshop group

- Log onto myExperiment
- In the search box, enter **VLIZ**, select **Groups** and click **Search**

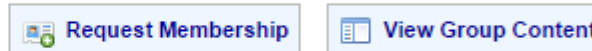




Joining the workshop group - 2

- Click the **Request Membership** button

Home > Groups > VLIZ Workshop Oct 2014



Group: VLIZ Workshop Oct 2014

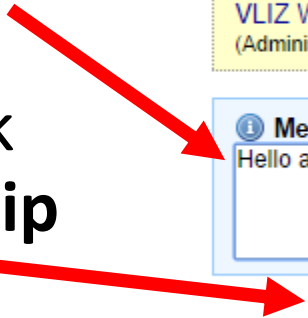
New Membership Request

- You should say hello to the group admin, then click **Send membership request**

Send a request to join the group:
VLIZ Workshop Oct 2014
(Administrator: Alan Williams)

Message to group administrator:
Hello admin person

Send membership request Return to Group





Take a Breath - There is More

- Taverna Command Line Tool, Server, Player, On-line
- Taverna integrations with:
 - IPython
 - Galaxy
- SysMO SEEK
- BioVeL
 - Portal for sharing and executing Taverna workflows



More information

- myGrid
 - <http://www.mygrid.org.uk>
- Taverna
 - <http://www.taverna.org.uk>
- BiodiversityCatalogue
 - <http://www.biodiversitycatalogue.org>
- myExperiment
 - <http://www.myexperiment.org>

