myGrid Tools for Scientists

Aleksandra Nenadic, myGrid team
University of Manchester

Overview

- Taverna
- BioCatalogue
- myExperiment

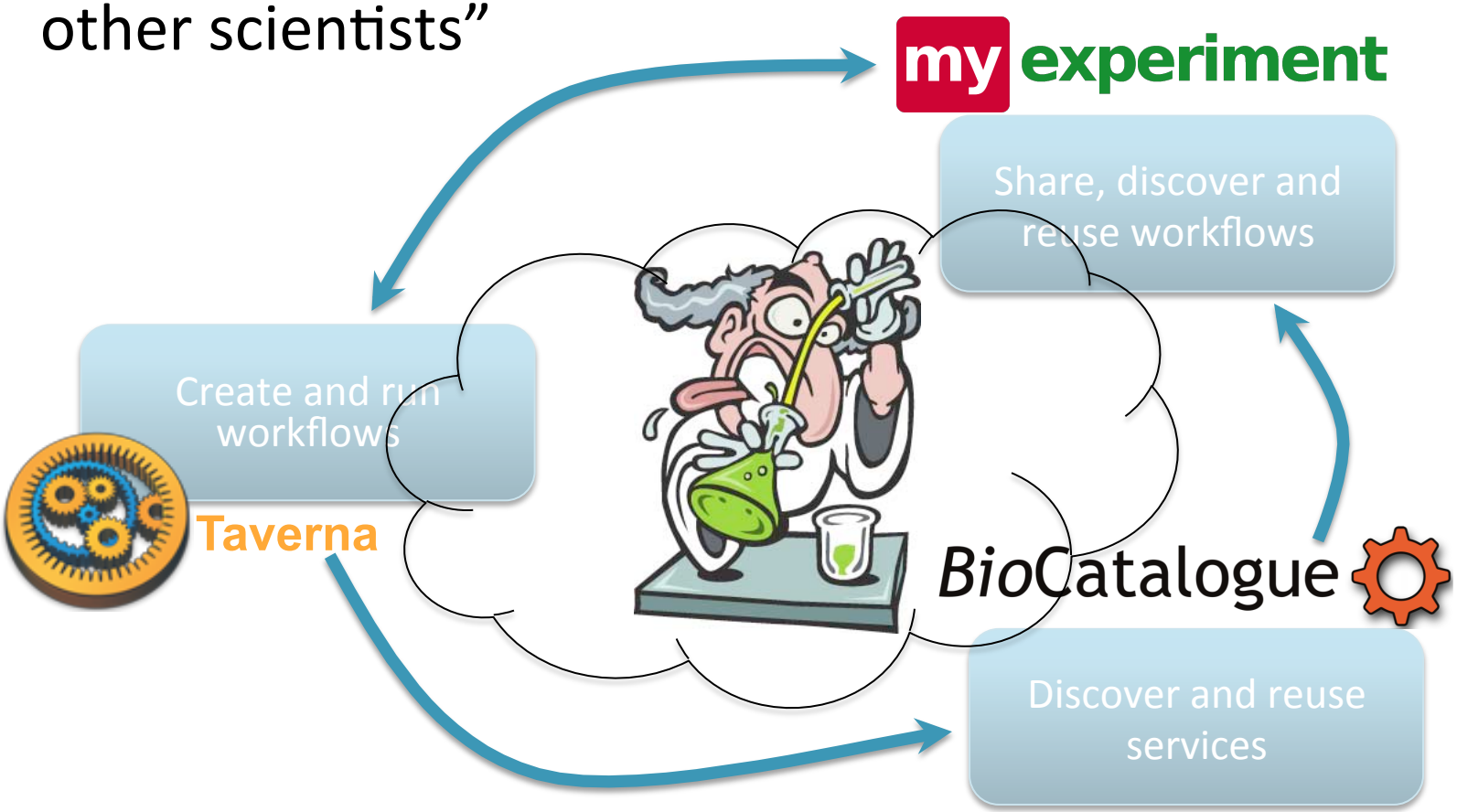
What is myGrid?

- "… 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 tools

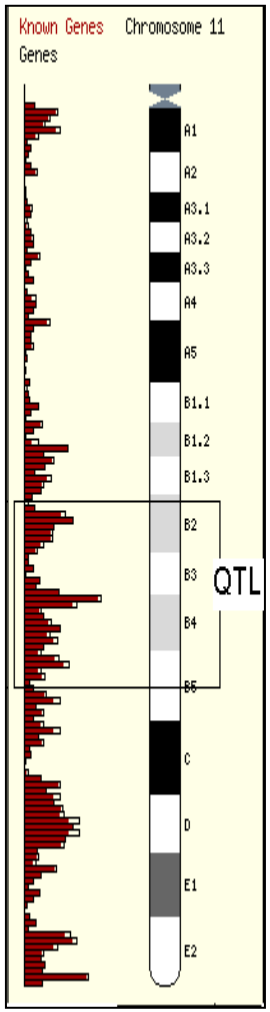
Software | Services | Content | Skills | Community

(Some of) myGrid Tools

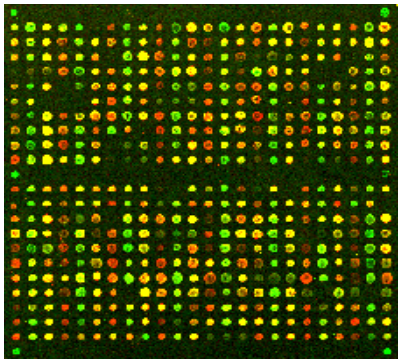
- " Help scientists get on with science and get on with other scientists"



A Problem: Huge Amounts of Data



QTL regions
100+ Genes



Microarray

1000+ Genes

Next Generation Sequencing

10,000+ 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 due to data size
- Constant change in data - need for reruns of experiments (re-analysis) to check for new and updated information
- Scientists often have to manually click though 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/scientific experiment
- Describes **what** you want to do on a high level
- Automation of scientific process enables:
  - Avoiding manual steps
  - Easier rerun
  - Parameter tweaking
  - Easier comparison of results
  - Science reproducibility

MANCHESTER 1824
The University of Manchester

The University of Manchester

SCAPE

BioVeL

myGrid

Creation and execution of workflows

Taverna

Taverna

- http://www.taverna.org.uk
- Workflow management system for creating and executing workflows
- A bit of history:
  - First released in 2004
  - Currently 2nd generation of software
    - Version 2.5 released in May 2014
  - 3rd generation in alpha
- Freely available, open source (LGPL license)
  - Windows, Mac OS and Linux
- Extensible
  - Via various plugins

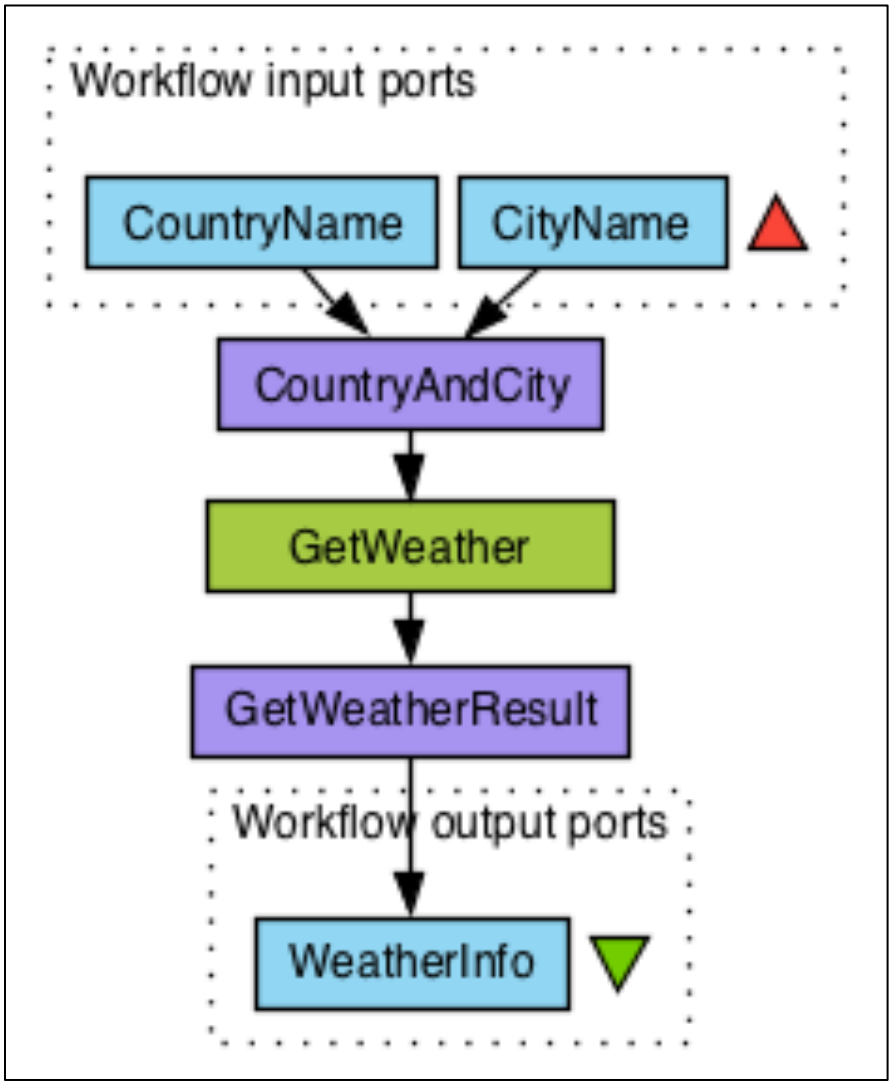
Taverna Workflows

- Sophisticated automated analysis pipelines for **data-driven** research or processes:
  - Define **how** you want your data to flow
  - Describe **what** you want to do with data, including the **services** to use
- 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 – cont'd

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

Example Taverna Workflow

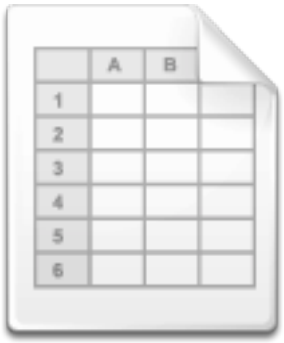


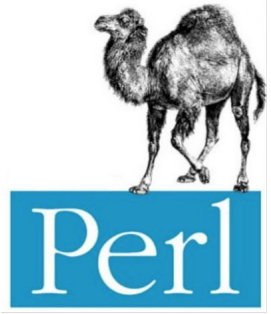
- Get the weather forecast of the day given the city and the country
- Green box is a Web service
- Purple boxes are local XML services to assemble/ extract XML (**shims**)
- Blue boxes are workflow input and output ports
- Arrows define the direction of data flow

Supported Services

- SOAP/WSDL Web services
- REST Web services
- 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
- myGrid does not own them
- myGrid did not build them
- myGrid does not enforce any 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**
- **Data lineage** - data life cycle - origin and how it moves over time
- 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: Sleeping Sickness Case Study

- Successful use of Taverna to assist with gene analysis relating to sleeping sickness (trypanosomiasis) in African cattle
  - Joint study: Universities of Manchester, Liverpool, and International Livestock Research Institute (Nairobi)
  - Some cattle breeds more disease-resistant than others
  - Genetical diff. between resistant and susceptible cattle?
  - Can we breed cattle resistant to infection?
- Two key genes have been identified
- Breeding trials have started with one of the genes to see if new lines of resistant cattle can be raised

Sleeping Sickness Case Study – cont'd

- BBC news:
  - http://www.bbc.co.uk/news/10403254
- EPSRC Pioneer Magazine 2014:
  - http://www.epsrc.ac.uk/newsevents/pubs/pioneer-edition-12-issued-june-2014/
- Watch a video on YouTube:
  - http://www.youtube.com/watch?v=hmIErdZwFS0

Taverna Summary

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

Discover and reuse services

BioCatalogue

- http://www.biocatalogue.org
- A public centralised and curated registry of Life Science 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)

Motivation

- Guesstimate: thousands publicly available on-line services in Life Sciences
- **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 BioCatalogue 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+xFiltering
- 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**

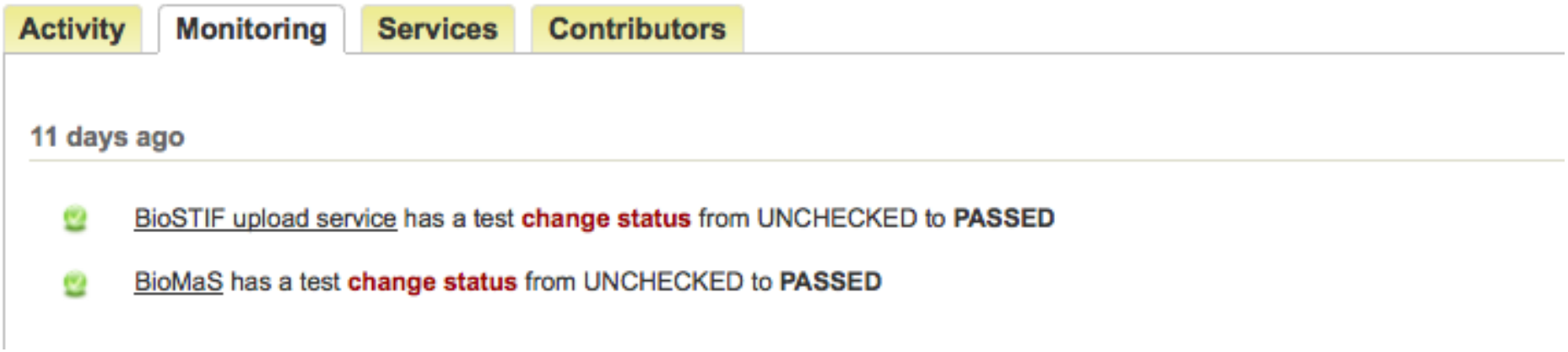
The last check for this test was successful

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



BioCatalogue 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"*


BioCatalogue

Demo


Share, discover and reuse workflows

myExperiment

- http://www.myexperiment.org
- 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

myExperiment Resources

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

myExperiment Resources – cont'd

- General files
  - Presentations
  - Data files
  - Auxiliary files
  - …
- Packs
  - Sets of resources in myExperiment
  - Like a **Research Object** or a full scientific experiment
    - Workflows, input data, results, logs, provenance, metadata, publications, presentations, etc.

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"*


myExperiment

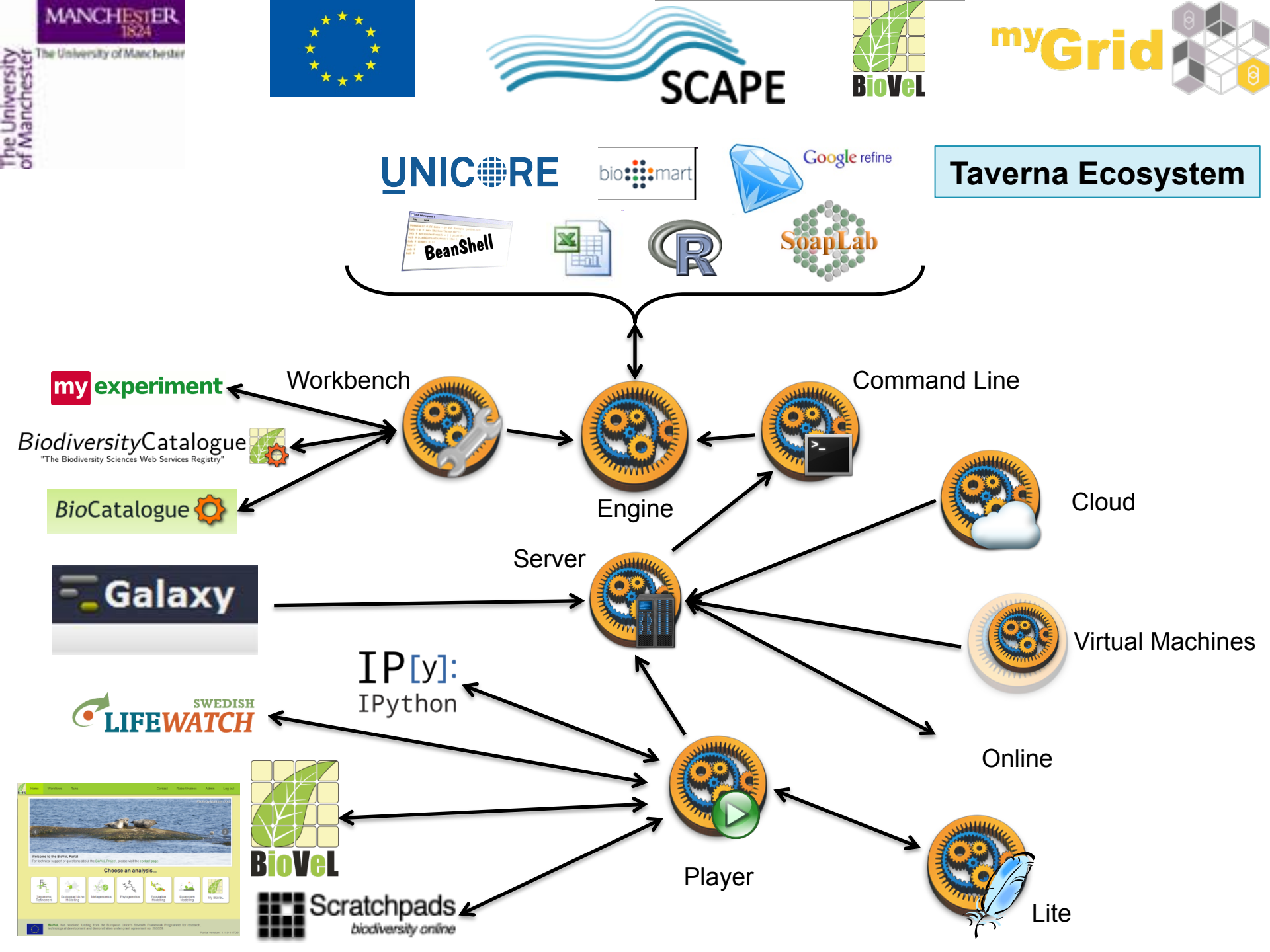
Demo

The Rest of the Day …

- Taverna introduction demo
- Taverna introduction – tutorial
- Taverna advanced features – tutorial
- Take a deep breath – there is more to myGrid!

Take a Breath - There is More

- Taverna Command Line Tool, Server, Player, On-line
- Taverna integrations with:
  - IPython
  - Galaxy
- SysMO SEEK
  - Data, models and SOPs (Standard Operating Procedures) repository
- BioVeL
  - Portal for sharing and executing Taverna workflows


MANCHESTER 1824
The University of Manchester
SCAPE
BioVeL
myGrid
UNICORE
bio mart
Google refine
Taverna Ecosystem
BeanShell
SoapLab
my experiment
Workbench
Command Line
BiodiversityCatalogue
"The Biodiversity Sciences Web Services Registry"
BioCatalogue
Engine
Cloud
Server
Galaxy
Virtual Machines
IP[y]:
IPython
SWEDISH
LIFEWATCH
Online
BioVeL
Player
Scratchpads
biodiversity online
Lite

More information

- myGrid
  - http://www.mygrid.org.uk
- Taverna
  - http://www.taverna.org.uk
- BioCatalogue
  - http://www.biocatalogue.org
- myExperiment
  - http://www.myexperiment.org

Acknowledgement:
Based on presentations by C.Goble, K. Wolstencroft, P. Fisher, J. Bhagat