

Architektura dużych projektów bioinformatycznych

Bartek Wilczyński

bartek@mimuw.edu.pl

<http://www.mimuw.edu.pl/~bartek>

**Wykład 6. - O przepływach pracy –
Kepler, Taverna i Galaxy**
16. XI. 2015

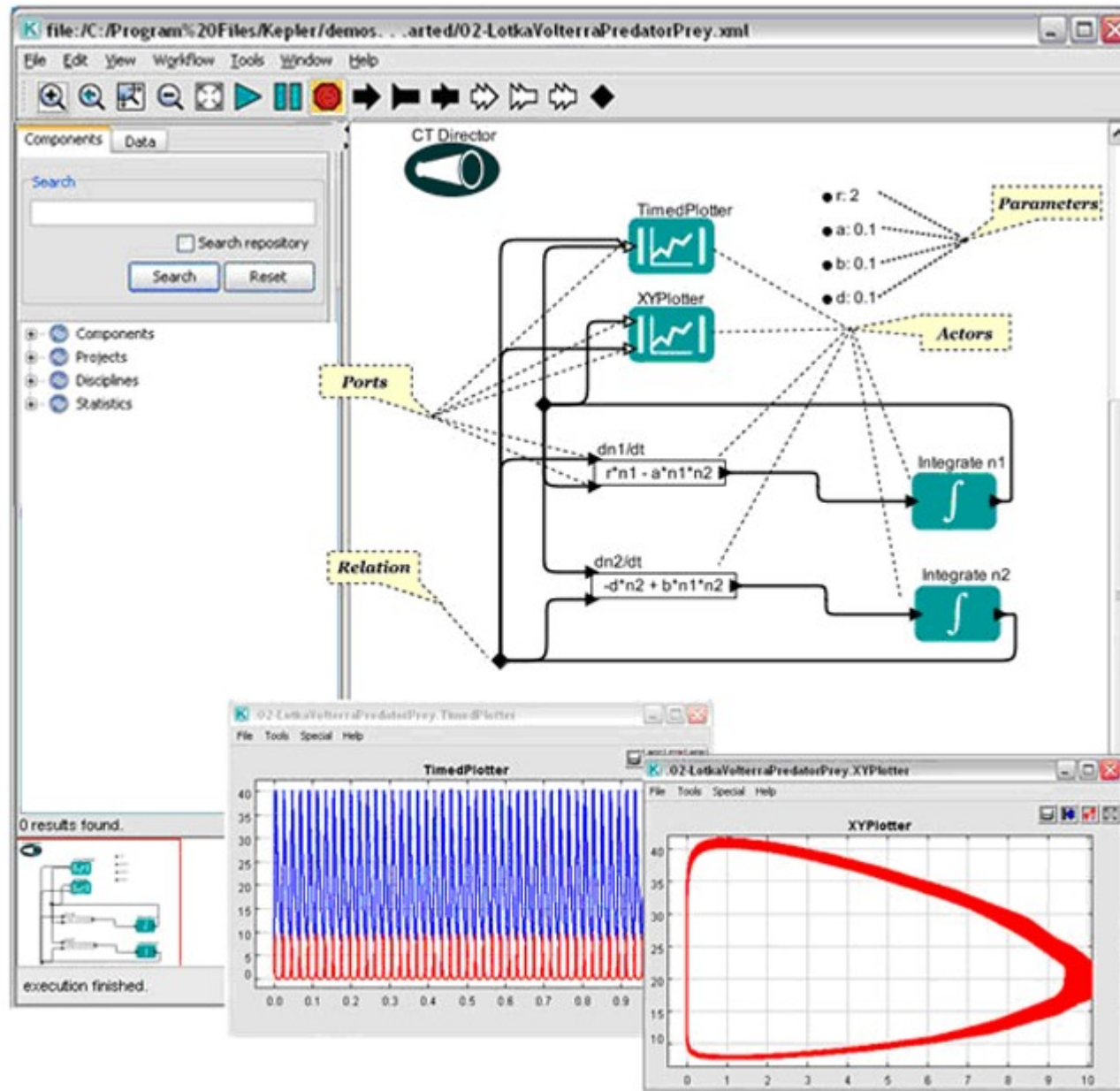
Plan na dziś

- Usługi sieciowe i projekty typu grid
- O pomysłach na przepływy pracy – Kepler (2002), Taverna workbench (2004), i wiele innych
- Pomysły na wymianę workflow'ów - myExperiment, i usług – biocatalogue
- Porównanie podejść – rozproszonego z użyciem aplikacji klienta (Taverna itp) i klient-serwer przez przeglądarkę (Galaxy)

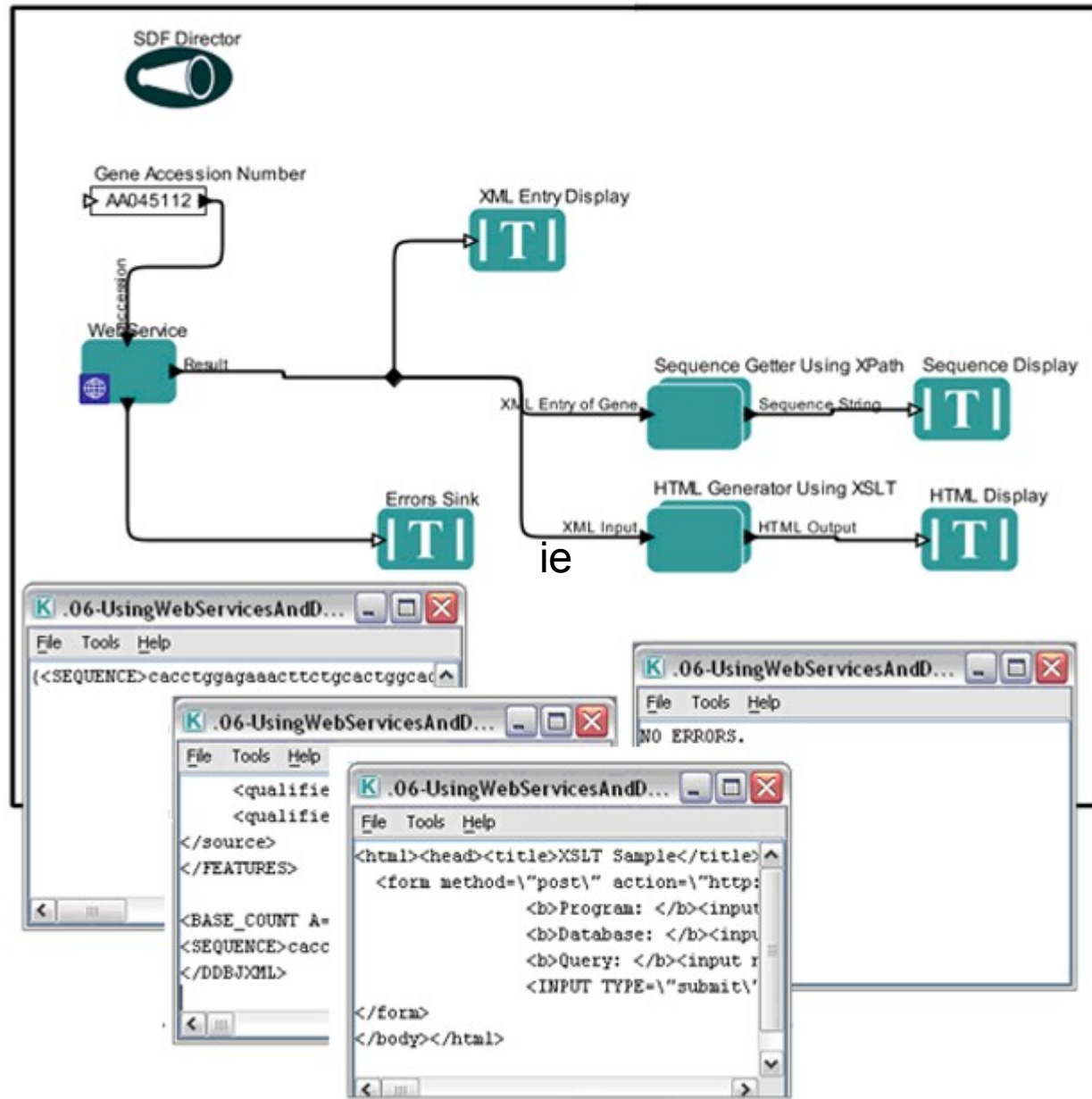
Projekt Kepler

- Pomyślany jako interfejs do pakietu Ptolemy (1999), służącego do złożonego modelowania matematycznego procesów fizycznych
- Napisany w Javie
- Pierwsze wersje w roku 2002
- Obecnie w wersji stabilnej (ostatnie wydanie w połowie 2013)
- Napisany w Javie
- Open Source(BSD), wspierany przez granty NSF

Model Lotki-Volterra w Keplerze



Pobieranie sekwencji DNA z usługi sieciowej



Cechy Kepler'a

- Środowisko uruchomieniowe Ptolemy (równoległe działające systemy)
- Interfejsy do czujników eksperymentalnych
- Wersje na GRID
- BioKepler dla bioinformatyków
- Raczej dla eksperymentatorów
- Silne wsparcie w środowisku fizyków doświadczalnych

Taverna Workbench

- Rozpoczęta jako oddzielny projekt w roku 2002
- Wsparcie grantami badawczymi, obecnie konsorcjum
- Licencja LGPL
- Zasadniczo oparta na pomycie reproducible research
- Oryginalnie pomyslanana jako interfejs do webserwisow (bez obliczen lokalnych)
- Wersje dla astronomow, bioinformatykow, itp.
- Obecnie rowniez do obliczen lokalnych, interfejs do R, Galaxy itp

Pobieranie sekwencji z Ensembl

Taverna Workbench v1.5.1.6

Design Results Discover

Search Watch loads

Local Services

- Notification Processor
- Local Java widgets
 - String Constant
 - BSF scripting host
 - AbstractProcessor - Processor for abstract taskdescriptions
 - RShell - Run R/S scripts through RServe
 - Beanshell scripting host
- WSDL @ http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jws?wsdl
- WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl
- WSDL @ http://soap.bind.ca/wsd1/bind.wsdl
- WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl
- WSDL @ http://soap.genome.jp/KEGG.wsdl
- WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
- Biomart service @ http://www.biomart.org/biomart
- Biomoby @ http://mobycentral.icapture.ubc.ca/cgi-bin/MOBY05/mobycentral.pl
- SeqHound @ seqhound.blueprint.org
- Soaplab @ http://www.ebi.ac.uk/soaplab/emboss4/services/

Advanced model explorer

Workflow Object properties

Add Nested Workflow Offline

Workflow object	Retrie	Delay	Backof	Thread	Critica
BiomartAndEMBOSSAnalysis					
Workflow inputs					
Workflow outputs					
outputPlot					
HSapIDs					
MMusIDs					
RNorIDs					
Processors					
FlattenImageList	0	0	1	1	
getMMsequence	0	0	1	1	
getRNsequence	0	0	1	1	
getHSsequence	0	0	1	1	
hsapiens_gene_ensembl	0	0	1	1	
GetUniqueHomolog	0	0	1	1	
CreateFasta	0	0	1	1	
seqret	0	0	1	5	
emma	0	0	1	5	
plot	0	0	1	5	
emma	0	0	1	5	
Data links					
CreateFasta:fasta-seqret:sequen					
GetUniqueHomolog:HSOut-getHS					
GetUniqueHomolog:MouseOut-ge					

```
graph TD; A[hsapiens_gene_ensembl] --> B[GetUniqueHomolog]; B --> C[getMMsequence]; B --> D[getRNsequence]; B --> E[getHSsequence]; C --> F[CreateFasta]; D --> F; E --> F; F --> G[seqret]; G --> H[emma]; H --> I[plot]; I --> J[FlattenImageList]; J --> K[outputPlot]; J --> L[HSapIDs]; J --> M[MMusIDs]; J --> N[RNorIDs];
```

Rendering done.

BioCatalogue

The screenshot shows the BioCatalogue website homepage. At the top, there is a navigation bar with links for 'Getting Started', 'About Us', 'Contact Us', and 'API Docs'. Below this is a search bar and a navigation menu with 'Home', 'Services', 'Register a Service', 'Service Providers', 'Search by Data', and 'Latest'. The main content area features a central banner with the text 'The BioCatalogue: providing a curated catalogue of life science Web services' and a statistics bar stating 'BioCatalogue currently has 1185 services, 247 service providers and 796 members'. Below the banner are four service categories: 'DISCOVER' (finding the right Web Service), 'REGISTER' (easily register Web Services), 'ANNOTATE' (describe and annotate services), and 'MONITOR' (services change and get outdated). On the left side, there are sections for 'Helpful Links' and 'Citing BioCatalogue'. On the right side, there is a 'Site Announcements' section with several news items.

myExperiment x BioCatalogue - Home x

https://www.biocatalogue.org

Google

BioCatalogue
The Life Science Web Services Registry

Getting Started | About Us | Contact Us | API Docs

Sign up | Sign in

Search: Go! | Home | Services | Register a Service | Service Providers | Search by Data | Latest

Home »

SHARE EMAIL

The BioCatalogue: providing a curated catalogue of life science Web services

BioCatalogue currently has **1185 services**, **247 service providers** and **796 members**

Helpful Links

- [Getting started with the Catalogue](#)
- [General info on Web services](#)
- [Best practice guide for developing Web services](#)
- [SSI's top tips on creating Web services](#)
- [Executing Web services](#)
- [Turn your command line application\(s\) into Web services](#)

Citing BioCatalogue

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orłowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.

BioCatalogue: a universal catalogue of web services for the life sciences

Nucl. Acids Res. (2010) 38: 689-694
doi:10.1093/nar/gkq394

Site Announcements

- Scheduled Maintenance**
By [Niall Beard](#) (7 months ago)
- BioCatalogue Rails 3 Upgrade**
By [Aleksandra Nenadic](#) (11 months ago)
- Soaplab EMOSS services end of life: 31st January 2013**
By [Robert Haines](#) (about 2 years ago)
- KEGG WSDL/SOAP services end of life: 31st December 2012**
By [Robert Haines](#) (about 2 years ago)
- BioCatalogue Scheduled Downtime 14th August 2012**
By [Robert Haines](#) (over 2 years ago)

[More](#)

"Web Services are hard to find" DISCOVER

- Find the right Web Service
- Powerful search and filtering
- Information from providers and community

[More info](#)

"My Web Services are not visible" REGISTER

- Easily register Web Services
- Instantly available to everyone
- Providers can advertise, describe and monitor their Services

[More info](#)

"Web Services are poorly described" ANNOTATE

- Anyone can describe and annotate
- Ongoing expert curation
- Social curation by the community

[More info](#)

"Web Services are volatile" MONITOR

- Services change and get outdated
- BioCatalogue monitors Services
- Monitors availability and reliability

[More info](#)

MyExperiment



myExperiment makes it easy to **find, use and share scientific workflows** and other **Research Objects**, and to build **communities**.

All Search

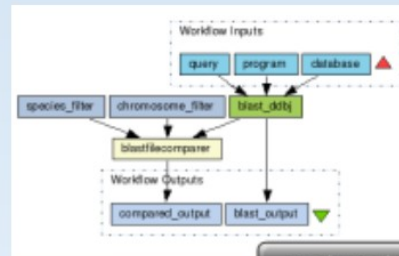
First time visitor? Try these videos:

- [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



[Find Workflows](#)

[About myExperiment](#)

[Join the Mailing List](#)

[myExperiment Publications](#)

[For Developers](#)

[Give us Feedback](#)

[The BioCatalogue Project](#)

Register

or Login:

Username or Email:

Password:

Remember me:

Or use OpenID:

(eg: name.myopenid.com)

[Login](#)

[Forgot Password?](#)

Galaxy

- Środowisko do uruchamiania narzędzi bioinformatycznych
- Architektura klient-serwer, ze wsparciem dla historii danych i analiz
- Pozwala na zapisywanie historii i dzielenie się nimi w ramach serwera
- Edytor “przepływów”, pozwalający na uruchamianie tych samych analiz na wielu zbiorach danych
- Zachowuje elementarne zabezpieczenia danych
- Pozwala na publikację wyników analiz

Publiczny serwer galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0%

New high performance job execution options are available! See [the wiki](#) for more information.

Tools


search tools

Get Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity

NGS TOOLBOX BETA
Phenotype Association
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: RNA-seq
NGS: Picard (beta)
NGS: Variant Analysis
NGS: VCF Manipulation
snpEff
BEDTools
EMBOSS

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Want help?
Get answers.

 GALAXY EXPLAINED

Tweets

Galaxy Project @galaxyproject 14 Nov
A case study for cloud based high throughput analysis of NGS data using Globus Genomics system, Bhuvaneshwar et al. [bit.ly/112U5OV](#)
Expand

CRS4 Orione @CRS4Galaxy 13 Nov
Seven Mycoplasma hyosynoviae strains assembled using Orione @CRS4Galaxy [goo.gl/XMMci1](#)
Retweeted by Galaxy Project
Expand

Galaxy Project @galaxyproject 13 Nov
RNA-Seq Analysis of Differential Gene Expression in Electroporated Chick Embryonic Spinal Cord, Viecelli & Yan, JoVE [bit.ly/112kZ9U](#)
Expand

Tweet to @galaxyproject

History

search datasets

Unnamed history
0 bytes

This history has been deleted

This history is empty. You can [load your own data](#) or [get data from an external source](#)

PENN STATE

JOHNS HOPKINS UNIVERSITY

TACC

iPlant Collaborative™

The Galaxy Team is a part of [the Center for Comparative Genomics and Bioinformatics at Penn State](#), and the [Department of Biology](#) and at [Johns Hopkins University](#).

This instance of Galaxy is utilizing infrastructure generously provided by the [iPlant Collaborative](#) at the [Texas Advanced Computing Center](#), with support from the [National Science Foundation](#).

The Galaxy Project is supported in part by [NSF](#), [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Johns Hopkins University](#).

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own [local Galaxy instance](#) or run [Galaxy on the cloud](#).

Workflow editor

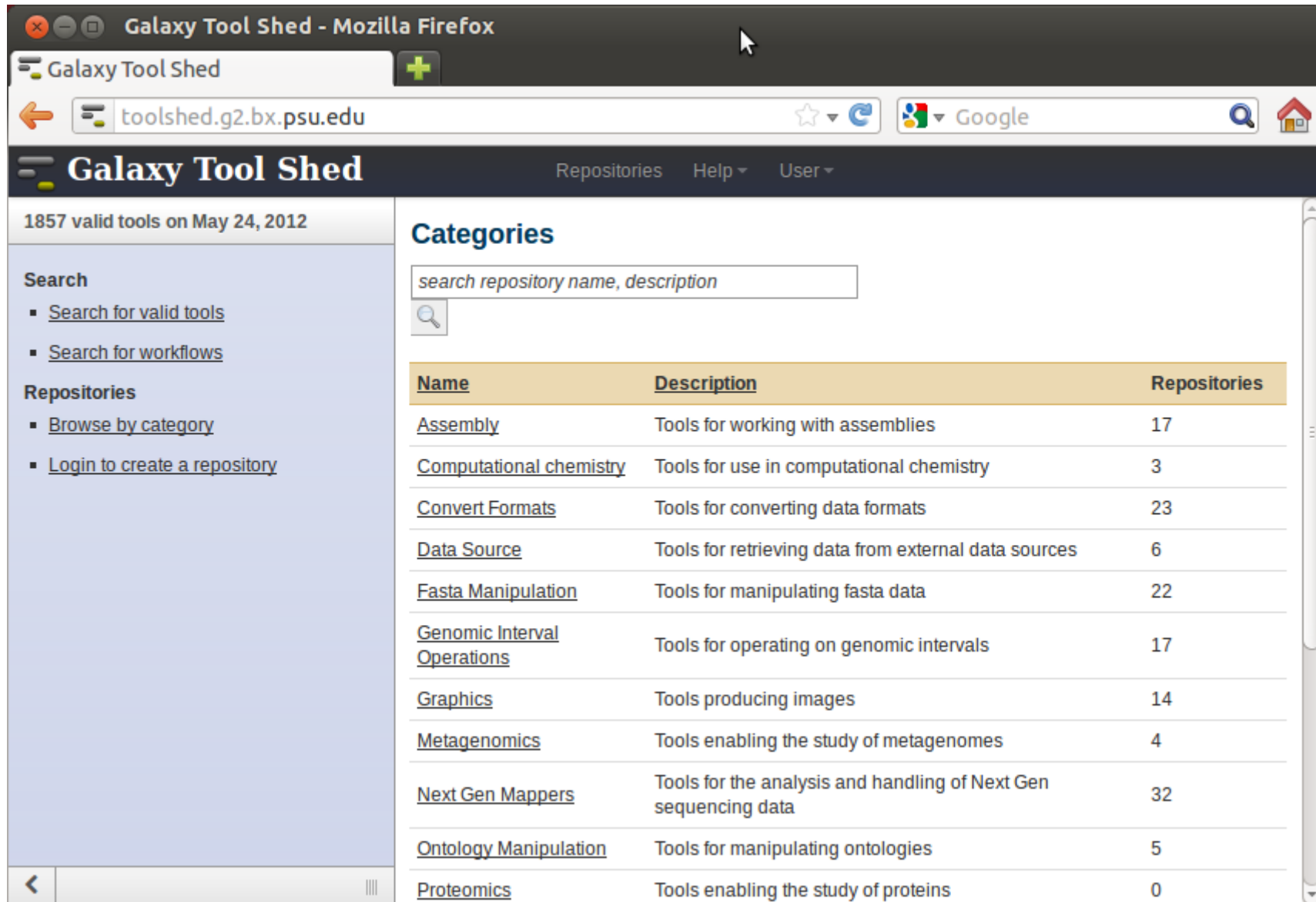
The screenshot displays the Galaxy workflow editor interface. The main canvas shows a workflow titled "Clone of 'metagenomic analysis' shared by 'anton@bx.psu.edu'". The workflow consists of several interconnected tools:

- Input dataset** (two instances) feeds into **Select high quality segments** and **FASTA-to-Tabular**.
- Select high quality segments** outputs **Reads** and **Quality scores**, which feed into **Megablast**.
- FASTA-to-Tabular** outputs **output (tabular)**, which feeds into **Add column**.
- Add column** outputs **out_file1**, which feeds into **Tabular-to-FASTA**.
- Tabular-to-FASTA** outputs **output (fasta)**, which feeds into **Megablast**.
- Megablast** (two instances) outputs **output1 (tabular)**, which feeds into **Concatenate queries**.
- Concatenate queries** outputs **out_file1**, which feeds into **Join two Queries**.
- Join two Queries** outputs **out_file1**, which feeds into **Filter**.
- Filter** outputs **out_file1**, which feeds into **Fetch taxonomic representation**.
- Fetch taxonomic representation** outputs **out_file1 (taxonomy)**, which feeds into **Find lowest diagnostic rank**.
- Find lowest diagnostic rank** outputs **out_file1 (taxonomy)**, which feeds into **Summarize taxonomy**.
- Summarize taxonomy** outputs **out_file1 (tabular)**, which feeds into **Draw phylogeny**.
- Draw phylogeny** outputs **out_file1 (pdf)**.

The right sidebar contains the **Details** panel for the selected **Filter** tool:

- Tool: Filter**
- Filter**
- Data input:** 'Input' (tabular)
- With following condition:** $c5/c15 > 0.5$
- Edit Step Attributes**
- Annotation / Notes:** (empty text area)
- TIP:** Double equal signs, ==, must be used as 'equal to' (e.g., $c1 == 'chr22'$)
- TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".
- TIP:** If your data is not TAB delimited, use **Text Manipulation -> Convert**
- Syntax**
- The filter tool allows you to restrict the dataset using simple conditional statements.
- Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of the tabular file.

Repozytorium narzędzi - toolshed



Galaxy Tool Shed - Mozilla Firefox

Galaxy Tool Shed

toolshed.g2.bx.psu.edu

Google

Galaxy Tool Shed

Repositories Help User

1857 valid tools on May 24, 2012

Search

- Search for valid tools
- Search for workflows

Repositories

- Browse by category
- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	17
Computational chemistry	Tools for use in computational chemistry	3
Convert Formats	Tools for converting data formats	23
Data Source	Tools for retrieving data from external data sources	6
Fasta Manipulation	Tools for manipulating fasta data	22
Genomic Interval Operations	Tools for operating on genomic intervals	17
Graphics	Tools producing images	14
Metagenomics	Tools enabling the study of metagenomes	4
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	32
Ontology Manipulation	Tools for manipulating ontologies	5
Proteomics	Tools enabling the study of proteins	0

Możliwość tworzenia własnych “opakowań” dla istniejących narzędzi

- Pisząc dość krótki plik “opakowujący” (Ang. Wrapper) można dołączyć dowolny program działający na serwerze do własnego galaxy
- Łatwo też można zmodyfikować istniejące wrappery, aby dostosować istniejący program do własnych potrzeb
- Jest to dość proste, choć jeśli dotyczy aplikacji graficznych, może być bardziej skomplikowane

Podsumowanie

- Wiele projektów pomyślanych jako narzędzie do “reproducible research”
- W zasadzie jedyny model finansowania to granty badawcze
- Bardzo powolna adopcja rozwiązań, które wymagają dodatkowego wysiłku, lub są ograniczone
- Duży sukces galaxy w środowisku bioinformatycznym, bo tu może stanowić ułatwienie
- Nadal niełatwe zdobywanie popularności wśród specjalistów