open middleware infrastructure institute uk

University of Southampton

myGrid

MANCHESTER 1824 The University of Manchester

myexperiment

A MySpace for the Self-serving Bioinformatician.

Carole Goble

University of Manchester, UK

carole.goble@manchester.ac.uk

from the myGrid people who brought you Taverna

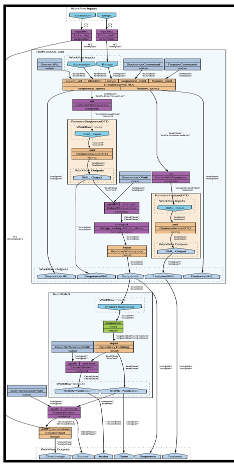
EPSRC

JISC

myexperiment

Talk Roadmap

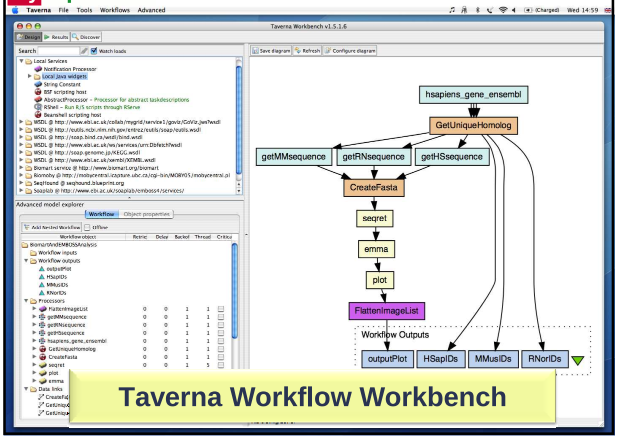
- Workflows (and, in particular, Taverna workflows)
- Workflows are hard won assets and valuable commodities
- myExperiment
  - – Social: networking and shopping.
  - – Repository: from workflows to Research Objects.
  - – Platform: federation, hosting and mashing.
- The Self Serving Selfish Scientist.
- Closing remarks



E. Science laboris

- Workflows are the new rock and roll.
- Machinery for coordinating the execution of (scientific) services and linking together (scientific) resources.
- The era of Service Oriented Applications
- Repetitive and mundane boring stuff made easier.
- The challenge for biology is complexity and heterogeneity, not so much compute.

myexperiment

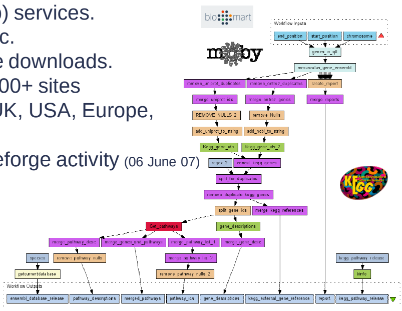


Taverna Workflow Workbench

myexperiment

Taverna Workflow Workbench

- myGrid project http://www.mygrid.org.uk
- Open Middleware Infrastructure Institute-UK.
- Open Source.
- Dataflow over (web) services.
- Application agnostic.
- 37,754 sourceforge downloads.
- 1,500 per month, 200+ sites
- Users throughout UK, USA, Europe, and SE Asia
- Ranked 210 sourceforge activity (06 June 07)



myexperiment

e-Services and Service Providers in the CLOUD

- Independent third party world-wide service providers of applications, tools and data sets. In the Cloud.
- My local applications, tools and datasets. In the Enterprise. In the laboratory.
- Easily incorporate new service without coding. So even more services from the cloud and enterprise.

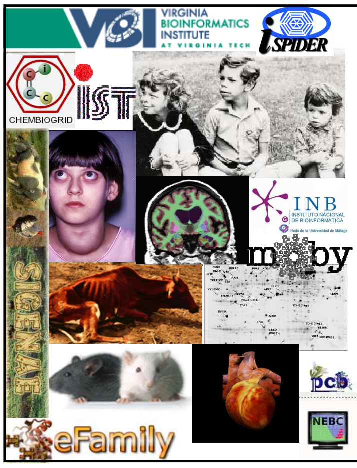
- Open Services
- 3500+ service operations

myexperiment

e-Scientists in the CLOUD

- Individual life scientists, in under-resourced labs, who use other people's applications, with little systems support.
- 200+ groups, ~1000 individuals.
- Exploratory workflows
- The developers are (often) the users. Consumers are providers too.

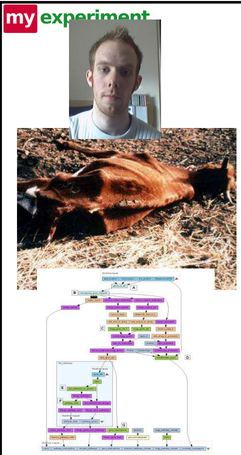
- A distributed, disconnected community of scientists.
- Decoupled suppliers and consumers of services and workflows.
- Scientists in an enterprise and in large projects
- Scientists out of the enterprise, in small projects or sole traders.


VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH
iSPIDER
CHEMBIOGRID
IST
INB INSTITUTO NACIONAL DE BIOINFORMATICA
moby
SIGENAE
pcb
NEBC
eFamily

- Systems biology
- Proteomics
- Gene/protein annotation
- Microarray data analysis
- Medical image analysis
- Heart simulations
- High throughput screening
- Phenotypical studies
- Phylogeny
- Text mining
- Plants, Mouse, Human
- Astronomy

myexperiment

Trypanosomiasis in Cattle



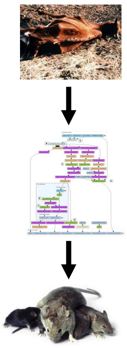
- QTL, microarray analysis, text mining, annotation mining.
- By a PhD student. Paul.
- Identified a pathway for which its correlating gene (Daxx) is believed to play a role in trypanosomiasis resistance.
- Manual analysis on the microarray and QTL data failed to identify this gene as a candidate.
- Repetitive, unbiased analysis.

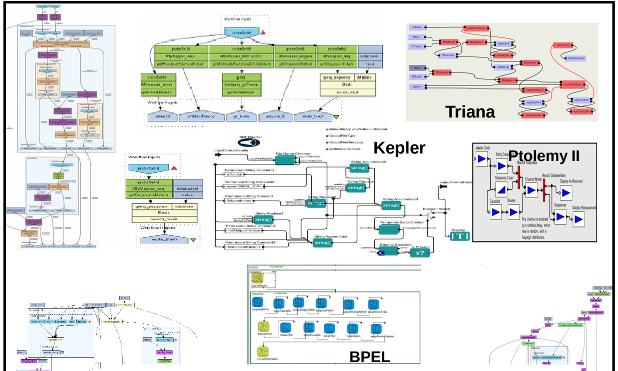
NAR paper

myexperiment

Recycling, Reuse, Repurposing

- Paul meets Jo.
- Trypanosomiasis cattle workflow reused without change.
- Identified the biological pathways involved in sex dependence in the mouse model, previously believed to be involved in the ability of mice to expel the parasite.
- Previously a manual two year study, by Jo, of candidate genes had failed to do this.





Scientific memes. Scientific viruses
Accompany their published outcomes
400+ Scufl workflows in the Web Cloud

myexperiment

Workflow baggage

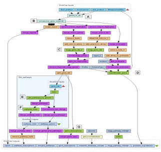
Design
- Workflow design history
- Experiment purpose
- Scientist

Log Book
- Workflow run log
- Data lineage tracking
- Results interpretation view

Data
- Input data, configuration
- Outcomes

Publications
- Notes and hints
- Powerpoint decks
- Articles

Components
- Services
- Applications



Why stop at workflows?

What about SBML models?

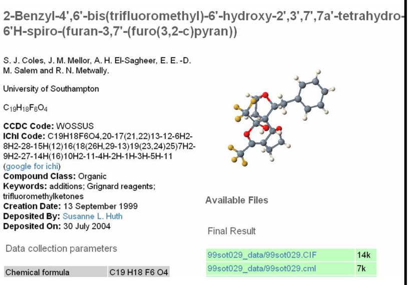
Wet lab protocols?

Publication transparency

The end to Black Box Science, by publishing workflows alongside publications and data.

Publishing Principles: The CombeChem experience

- Focus on making it easy to publish information
- Discovering and sharing experimental artefacts
- Publishing results to standard community repositories
- Publishing scholarly output.
- Publication at source.



2-Benzyl-4',6'-bis(trifluoromethyl)-6'-hydroxy-2',3',7',7a'-tetrahydro-6'H-spiro-(furan-3,7'-(furo(3,2-c)pyran))

S. J. Coles, J. M. Mellor, A. H. El-Sagheer, E. E. -D. M. Salem and R. N. Metwally.

University of Southampton

$C_{19}H_{18}F_6O_4$

CCDC Code: WOSSUS
IChI Code: C19H18F6O4,20-17(21,22)13-12-6H2-8H2-28-15H(12)16(18(26H,29-13)19(23,24)25)7H2-9H2-27-14H(16)10H2-11-4H-2H-1H-3H-5H-11
(google for ichi)
Compound Class: Organic
Keywords: additions; Grignard reagents; trifluoromethylketones
Creation Date: 13 September 1999
Deposited By: Susanne L. Huth
Deposited On: 30 July 2004

Data collection parameters

Chemical formula	C19 H18 F6 O4

Available Files

Final Result

99sot029_data/99sot029.CIF	14k
99sot029_data/99sot029.cml	7k

Workflows are more than just plumbing….

- Workflows are protocols and records.
  - Explicit and precise descriptions of a scientific protocol.
  - Scientific transparency. Easier to explain, share, relocate, reuse and repurpose and remember.
  - Provenance of results for credibility.
- Workflows are know-how.
  - Specialists create applications; experts design and set parameters; inexperienced punch above their weight with sophisticated protocols.
- Workflows are collaborations.
  - Multi-disciplinary workflows promote even broader collaborations.

Workflows are hard work…

- Often complex.
  - Need intelligent steering and analysis.
  - Need explanations to ensure used properly and safely.
- Challenging and expensive to develop.
  - Development assistance. Don't start from scratch.
  - Take a long time to build good ones & a lot of know-how.
- You can still build bad workflows.
  - Enable scientists to be scientists, not programmers.
  - Enable scientists to be creative yet sound.

Workflows are commodities…

- Valuable first class assets in their own right.
  - To be pooled and shared and traded and reused.
  - Within communities and across communities.
  - Of pieces, of wholes, of when and how to.
  - Pattern books. Validated community workflow packs.
  - Publish workflows
  - Peer review workflows.
  - Components of published and repeatable experiment objects to accompany their published outcomes, like papers.
- But….Reusability often confined to the project it was conceived.
- Social and technical challenges for sharing and reuse.

my experiment

Aerospace Engine Design Analogy

90% design Is variant design

70% of information taken from previous designs

Source: Silvia Wong, Uni Southampton

my experiment

Questions

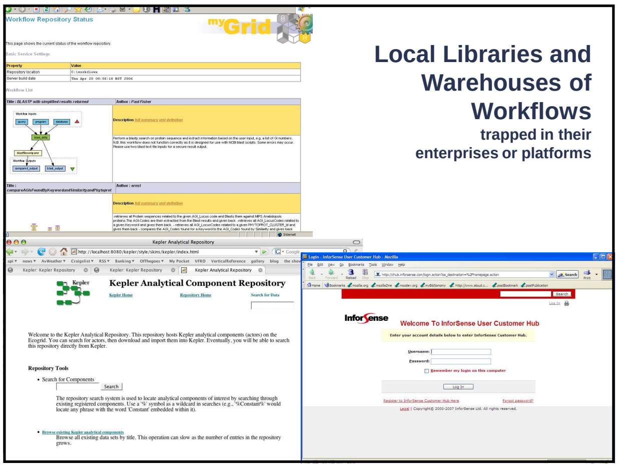
- Which workflow do other people I respect recommend for microarray normalisation?
- How do I use this service effectively?
- What does this workflow do?
- Is there a workflow that does close to what I want? Or need?
- Can I interpret the provenance of this data item properly?
- How do I run and steer this workflow?
- Can I substitute my service/database ok?
- Can I run this alien workflow?
- Did someone else run this and have a problem? Or get some different result?
- What other workflows were used with this workflow?
- Is this workflow reliable? Does it still work? How can I make it work when one of the services doesn't run anymore?

?

my experiment

e-Science is empowering-Scientists

- Enabling scientists to be (more) creative.
- Enabling scientists to be scientists. And not programmers.
- Enabling mediocre scientists to become better and thus have better science.
- Enabling smart scientists to be smarter and propagate their smartness.
- Accelerate dissemination, pooling, insight.
- Encouraging sanctioned plagiarism.
- For a community that are IN THE CLOUD and not in an ENTERPRISE.



Local Libraries and Warehouses of Workflows

trapped in their enterprises or platforms

Workflow Repository Status

This page shows the current status of the workflow repository.

Basic Service Settings

Property	Value
Repository location	C:\workflows
Server build date	Thu Apr 20 00:56:18 BST 2006

Workflow List

Title : BLASTP with simplified results returned — Author : Paul Fisher

Description full summary xml definition

Perform a blastp search on protein sequence and extract information based on the user input, e.g. a list of GI numbers. N.B. this workflow does not function correctly as it is designed for use with NCBI blast scripts. Some errors may occur. Please use two blast text file inputs for a secure result output.

Title : compareAGIsFoundByKeywordandSimilarityandPhytoprot — Author : ernst

Description full summary xml definition

-retrieves all Protein sequences related to the given AGI_Locus code and Blasts them against MIPS Arabidopsis proteins. The AGI-Codes are then extracted from the Blast results and given back. -retrieves all AGI_LocusCodes related to a given Keyword and gives them back. - retrieves all AGI_LocusCodes related to a given PHYTOPROT_CLUSTER_Id and gives them back - compares the AGI_Codes found for a Keyword to the AGI_Codes found by Similarity and gives back

Kepler Analytical Repository

http://localhost:8080/kepler/style/skins/kepler/index.html

Kepler Analytical Component Repository

Kepler Home Repository Home Search for Data

Welcome to the Kepler Analytical Repository. This repository hosts Kepler analytical components (actors) on the Ecogrid. You can search for actors, then download and import them into Kepler. Eventually, you will be able to search this repository directly from Kepler.

Repository Tools

- Search for Components
  Search
  The repository search system is used to locate analytical components of interest by searching through existing registered components. Use a '%' symbol as a wildcard in searches (e.g., '%Constant%' would locate any phrase with the word 'Constant' embedded within it).
- Browse existing Kepler analytical components
  Browse all existing data sets by title. This operation can slow as the number of entries in the repository grows.

Login - InforSense User Customer Hub - Mozilla

http://chub.inforsense.com/login.action?os_destination=%2Fhomepage.action

Welcome To InforSense User Customer Hub

Enter your account details below to enter InforSense Customer Hub.

Username:

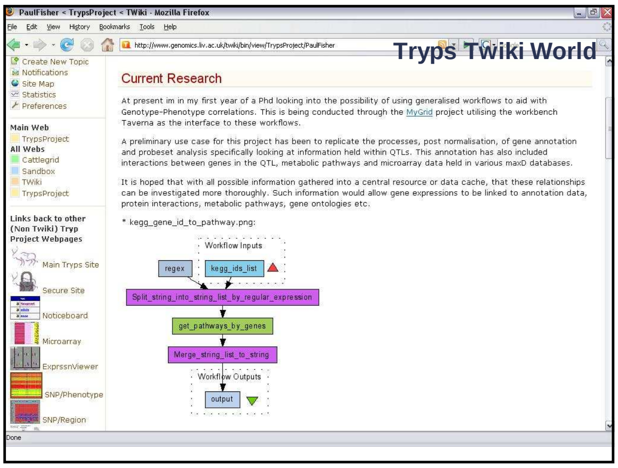
Password:

□ Remember my login on this computer

Log In

Register to InforSense Customer Hub Here — Forgot password?

Legal | Copyright© 2000-2007 InforSense Ltd. All rights reserved.



Tryps Twiki World

PaulFisher < TrypsProject < TWiki - Mozilla Firefox

http://www.genomics.liv.ac.uk/twiki/bin/view/TrypsProject/PaulFisher

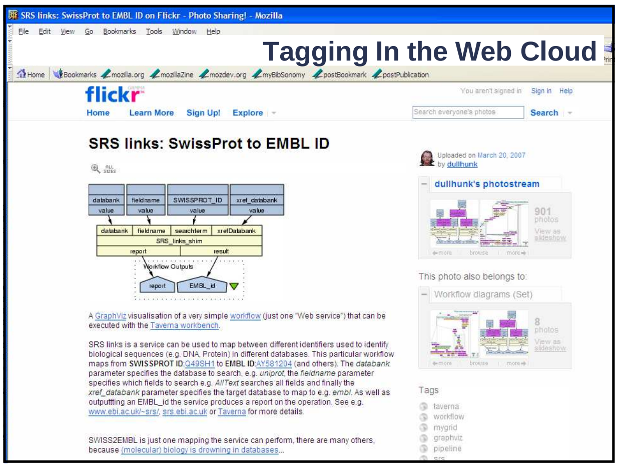
Current Research

At present im in my first year of a Phd looking into the possibility of using generalised workflows to aid with Genotype-Phenotype correlations. This is being conducted through the MyGrid project utilising the workbench Taverna as the interface to these workflows.

A preliminary use case for this project has been to replicate the processes, post normalisation, of gene annotation and probeset analysis specifically looking at information held within QTLs. This annotation has also included interactions between genes in the QTL, metabolic pathways and microarray data held in various maxD databases.

It is hoped that with all possible information gathered into a central resource or data cache, that these relationships can be investigated more thoroughly. Such information would allow gene expressions to be linked to annotation data, protein interactions, metabolic pathways, gene ontologies etc.

* kegg_gene_id_to_pathway.png:



Tagging In the Web Cloud

SRS links: SwissProt to EMBL ID on Flickr - Photo Sharing! - Mozilla

SRS links: SwissProt to EMBL ID

A GraphViz visualisation of a very simple workflow (just one "Web service") that can be executed with the Taverna workbench.

SRS links is a service can be used to map between different identifiers used to identify biological sequences (e.g. DNA, Protein) in different databases. This particular workflow maps from **SWISSPROT ID**:Q49SH1 to **EMBL ID**:AY581204 (and others). The *databank* parameter specifies the database to search, e.g. *uniprot*, the *fieldname* parameter specifies which fields to search e.g. *AllText* searches all fields and finally the *xref_databank* parameter specifies the target database to map to e.g. *embl*. As well as outputtting an EMBL_id the service produces a report on the operation. See e.g. www.ebi.ac.uk/~srs/, srs.ebi.ac.uk or Taverna for more details.

SWISS2EMBL is just one mapping the service can perform, there are many others, because (molecular) biology is drowning in databases...

Uploaded on March 20, 2007 by dullhunk

MySpace or FaceBook for Workflow e-Scientists

my experiment beta

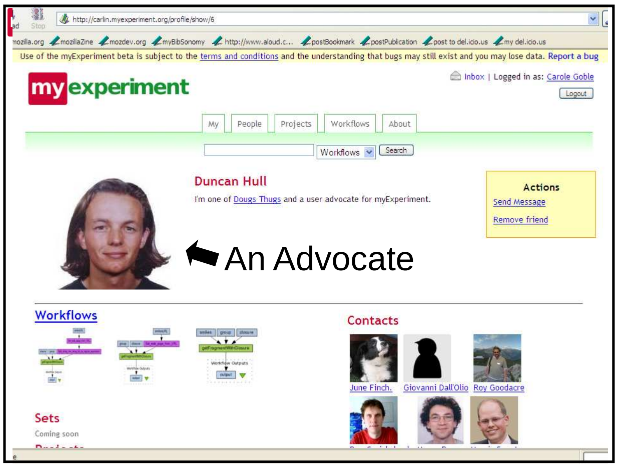
from the myGrid team who brought you Taverna

myExperiment makes it really easy for the next generation of scientists to contribute to a pool of scientific workflows, build communities and form relationships. myExperiment enables scientists to share, re-use and repurpose workflows and reduce time-to-experiment, share expertise and avoid reinvention.

> *"Their kids may have got there first but scientists will soon have their very own version of MySpace, where they will be able to share preliminary results, ideas and research tools."*
> — New Scientist Tech, October 2006.

myExperiment introduces the concept of a workflow bazaar; a collaborative environment where scientists can safely publish their creations, share them with a wider group and find the workflows of others. Workflows can now be swapped, sorted and searched like photos and videos on the web.

myexperiment.org



http://carlin.myexperiment.org/profile/show/6

Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

Inbox | Logged in as: Carole Goble

Duncan Hull

I'm one of Dougs Thugs and a user advocate for myExperiment.

Actions

Send Message

Remove friend

An Advocate

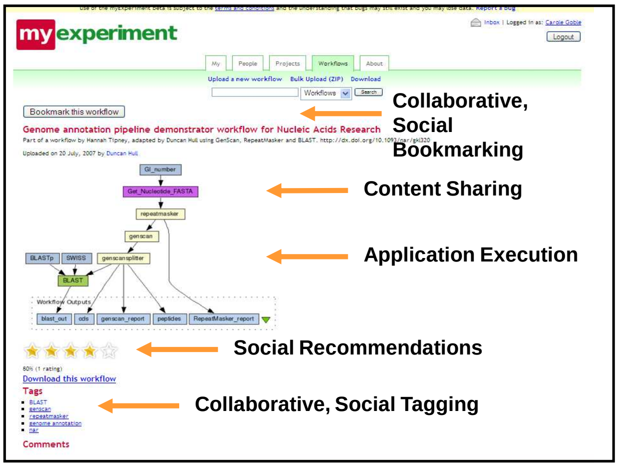
Workflows

Contacts

June Finch. Giovanni Dall'Olio Roy Goodacre

Sets

Coming soon



Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

Inbox | Logged in as: Carole Goble

Bookmark this workflow

Genome annotation pipeline demonstrator workflow for Nucleic Acids Research

Part of a workflow by Hannah Tipney, adapted by Duncan Hull using GenScan, RepeatMasker and BLAST. http://dx.doi.org/10.1093/nar/gkl320

Uploaded on 20 July, 2007 by Duncan Hull

80% (1 rating)

Download this workflow

Tags

- BLAST
- genscan
- repeatmasker
- genome annotation
- nar

Comments

Collaborative, Social Bookmarking

Content Sharing

Application Execution

Social Recommendations

Collaborative, Social Tagging

my experiment

Demo

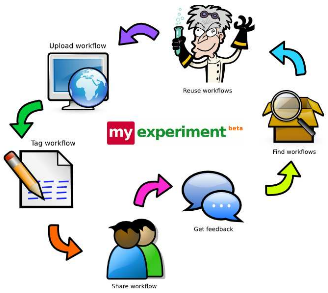
my experiment

Key evaluator and inspiration

- Cat De Roure
- Her idea.
- 15 years old.
- This is one of her MySpace pictures.
- Immediately understood it.
- Immediately could use it.
- "It's like MySpace for science homework."

myExperiment.org is…

- A market place.
- A community social network.
- A gateway to other publishing environments.
- A federated repository
- A platform for launching workflows.
- Publishing self-describing encapsulated myExperiment Objects.
- Mindful publication.
- Started February 2007.



myExperiment.org principles…

- Make it easy to publish, easy to participate, easy to add value to though mash-ups
- Use familiar techniques
  - Shopping, Social networking, gaming
- Use off the shelf, open source web tooling, not restrictive portals. Keep it funky, keep it flexible, keep it extensible. Assume other people will add functionality.
  - Ruby on Rails, Facebook platform
- Aim it at young people. Make it fun and attractive. Say no to 1970s library interfaces!

Front End.

A market place.
A community social network.

User Participation.

A Market Place: Shoe Shop?

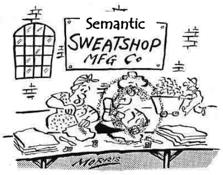
- Shopping for Workflows and Services and Data should be as easy as shopping for shoes.
- Don't need to train people.
- Fuel for diagnostics. Find a similar workflow.
- Organic growth good and bad.
- We need good, organised metadata for automated use.
- Impediance mismatch
- Identity and Ontology Authority

A Social Network and Collective Intelligence

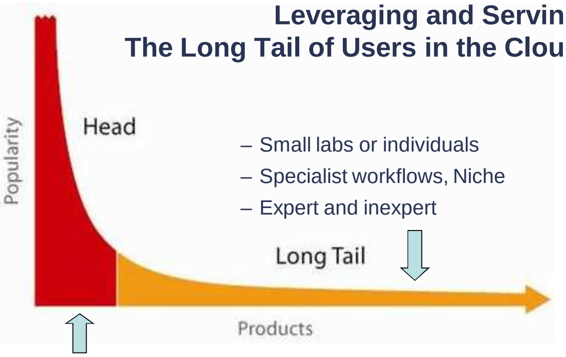
- Source of large amount of metadata.
- Open tagging, folksonomies, blogging, profiles, recommendations.
- Social network analysis and e-tracking are valuable intelligence.
- How do we avoid being deafened by the shouting?
- What are the incentive models for scientists?

Semantic Sweatshop

- Service vocabularies and curation in the wild
  - Keywords and tagging only goes some of the way.
  - Curation
- Workflow metadata
  - Find a workflow that is similar to what I want or need; similar to this one. Or hers. Hmmmm.
- Recommendations, people profiles
- Fuel for discovery and diagnostics
- Pipeline from tag cloud to shop



Leveraging and Serving
The Long Tail of Users in the Cloud



- Small labs or individuals
- Specialist workflows, Niche
- Expert and inexpert

- Big labs & big groups
- Common *de facto* workflows

myexperiment

Back End.

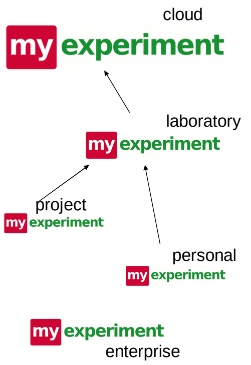
Federated repositories.
A gateway to other publishing environments.
A platform for launching workflows.

User Participation.

myexperiment

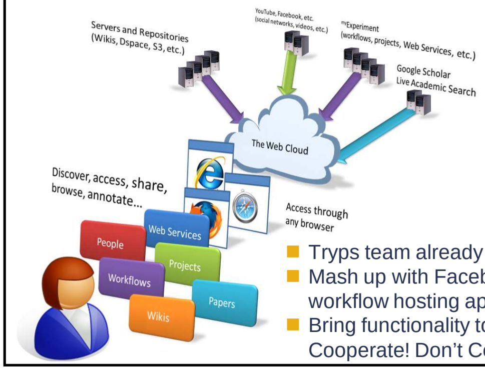
Warehouse or Federation

- Community web site, federated repository.
- Multiple and My.
- Publish what I want when I want within the group I want.
- Mixed identity regimes: an identity authority
- Open Archives Initiative.
  http://www.openarchives.org/
- The CombeChem project.
  http://www.combechem.org/



myexperiment

A gateway to other publishing services


Servers and Repositories
(Wikis, Dspace, S3, etc.)
YouTube, Facebook, etc.
(social networks, videos, etc.)
myExperiment
(workflows, projects, Web Services, etc.)
Google Scholar
Live Academic Search
The Web Cloud
Discover, access, share, browse, annotate...
Access through any browser
People
Web Services
Projects
Workflows
Papers
Wikis

- Tryps team already has a wiki
- Mash up with Facebook and workflow hosting apps.
- Bring functionality to the user. Cooperate! Don't Control.

myexperiment

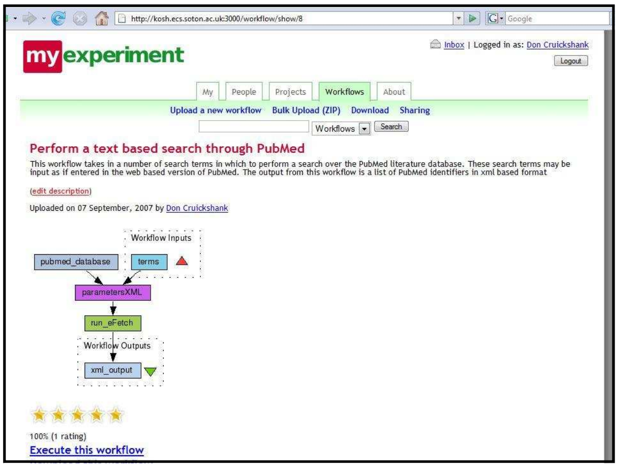
We Need You! Cooperate, Don't Control

- A network of cooperating data services with simple interfaces which make it easy to work with content.
- Provides services
- Reuses the service of others.
- Support lightweight programming models so that it can easily be part of loosely coupled systems.
- You add a mashup!
- Workflows as content syndication?

myexperiment

A platform – popular request!

- See a workflow? Launch it! Coming soon….
  - The ability to execute workflows from myExperiment + integration of tools such as Taverna achieves increased value through usage – network effects.
  - And lets do that with Kepler, Triana etc
  - The WHIPs project for a workflow hosting portal environment. Funded by OMII-UK. Run by Triana!
  - Handling of results? Wiki pages?
- Workflows available through a variety of interfaces supporting workflow execution.



Perform a text based search through PubMed

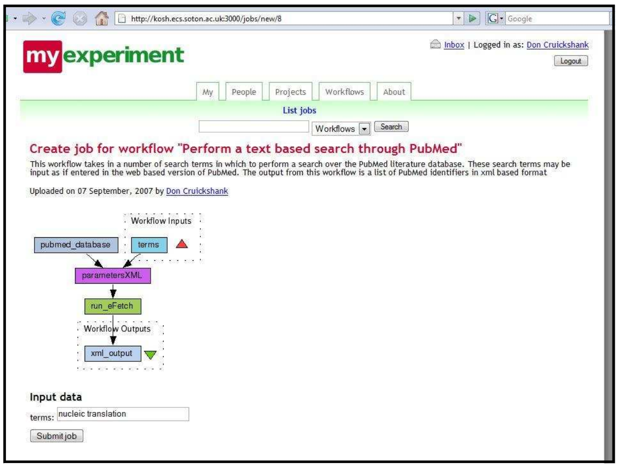
This workflow takes in a number of search terms in which to perform a search over the PubMed literature database. These search terms may be input as if entered in the web based version of PubMed. The output from this workflow is a list of PubMed identifiers in xml based format

(edit description)

Uploaded on 07 September, 2007 by Don Cruickshank

100% (1 rating)

Execute this workflow



http://kosh.ecs.soton.ac.uk:3000/jobs/new/8

Inbox | Logged in as: Don Cruickshank

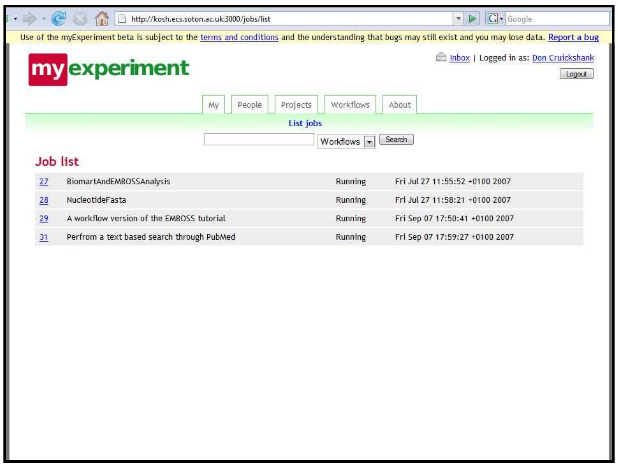
Create job for workflow "Perform a text based search through PubMed"

This workflow takes in a number of search terms in which to perform a search over the PubMed literature database. These search terms may be input as if entered in the web based version of PubMed. The output from this workflow is a list of PubMed identifiers in xml based format

Uploaded on 07 September, 2007 by Don Cruickshank

Input data

terms: nucleic translation



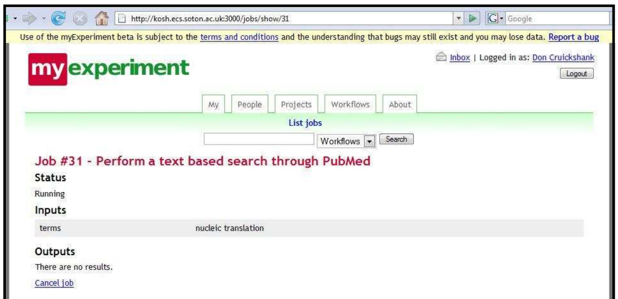
http://kosh.ecs.soton.ac.uk:3000/jobs/list

Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

Inbox | Logged in as: Don Cruickshank

Job list

27	BiomartAndEMBOSSAnalysis	Running	Fri Jul 27 11:55:52 +0100 2007
28	NucleotideFasta	Running	Fri Jul 27 11:58:21 +0100 2007
29	A workflow version of the EMBOSS tutorial	Running	Fri Sep 07 17:50:41 +0100 2007
31	Perfrom a text based search through PubMed	Running	Fri Sep 07 17:59:27 +0100 2007



http://kosh.ecs.soton.ac.uk:3000/jobs/show/31

Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

Inbox | Logged in as: Don Cruickshank

Job #31 - Perform a text based search through PubMed

Status

Running

Inputs

terms	nucleic translation

Outputs

There are no results.

Cancel job

What to do with the outputs?

myexperiment

The Selfish Scientist

The Self Serving Bioinformatician

myexperiment

User Participation for Content and Functionality

- Adoption depends on lots of shared services and workflows
- and enabling Scientists to add value through applications and collaborative tagging
- The Selfish Scientist – e-Science is me-Science
- Incentive models for Scientists to share?

myexperiment

The Selfish Scientist

“A biologist would rather share their toothbrush than their gene names”

Mike Ashburner
Professor Genetics
University of Cambridge, UK

And others

my experiment

What motivates Scientists?

Ego! Competitive advantage.
To be the first with the Nature paper.
To get my service / tool / foo / bar used.
To get added value to my service / tool for low or no cost.
Protect my turf
Credit, credibility, fame, acclaim, recognition, respect.
And more funding.
And not to be misinterpreted or misrepresented.
Or look stupid.

my experiment

“We are Naughty” – Helen Hulme

Low cost of adoption. Surmountable obstacles.
Fit into my world and what I do.
Don’t give me pre-defined recipes or tie me down.
Freedom and me in control
Jam today and more and better jam tomorrow
Just enough, just in time when I think its worth it.
Gimme content that I need and want.
Its my idea.

science commons

Hell is other people’s workflows

- Some Rights Reserved.
- Privacy, IP, Authorship attribution and guarantees.
- Copyright has to be sorted. Scientists need protection as well as sharing.
- Curation and Policing. Quality, Reliability, Validation, Safety, Persistence, Longevity, Palpability?
- Pollution and viral infections.
- Licences for services. Can I run this workflow?
- Provenance.
- Can we mix up personal results safely with public resources?

my experiment

Workflow Publishing and Mashing

- Workflow Publishing
  - What does it mean to publish a workflow? Expose vs export vs publish?
  - Encapsulate a workflow with its provenance
  - Design for hackability and remixability. Is that possible?
- Workflow Design Mashing
  - "hackable" and "remixable“…so….
  - When does your workflow stop being your workflow? If your workflow is combined with another are you still responsible if it is now crap?
  - Provenance mashing?
  - Mashing tracking?

my experiment

The Perpetual Beta

- Continually evolve in response to its users and by its users
- Design for user participation mash ups
- Developers embedded in the user community.
- Pilot with local friends and family.
- Advocates for supported of communities

Freeplay >120 users
First pilot July 07
Series of User Trials

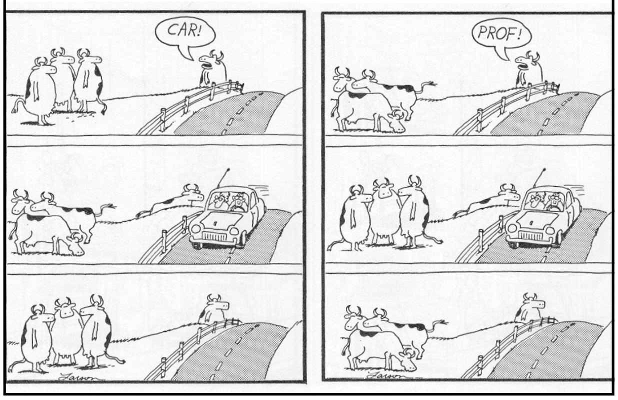
my experiment

Our Laboratory Guinea Pigs

- Chemists at the University of Southampton, UK
- Blogging data, publishing lab books

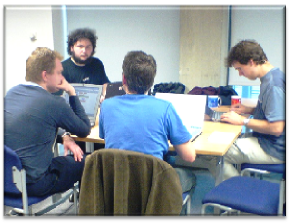
- Bioinformaticians at the University of Manchester, UK
- Publishing workflows and models

Embedding Outcomes in Doug Kell's Lab….



The Perpetual Beta

- A collection of online services
- Continually evolve in response to its users and by its users
- Design for user participation mash ups
- Developers embedded in the user community.
- Hackfests. Do, don't think!
- New code base in 4 weeks over summer
- Microsoft version



From me-Science to we-Science

- Tribal bonding and sharing
- Crossing Tribal Boundaries
- Across communities and disciplines (MIT)
- "Intellectual Fusion" & "Swarming"; breaking down silos
- Understanding outside my expertise. E.g. sources of error
- Metadata challenges.
- Social challenges.

Closing myExperiment.org

- A Web 2.0 Environment for Scientists to share experiments.
- The collective benefits of participation arise not only from the users but also from the developers – ease of use and ease of development.
- Innovation is rarely technology driven. The challenge is the social.

Acknowledgements

- David De Roure
- Matt Lee, Don Cruickshank, David Withers, Mark Borkum, Jiten Bhagat (Developers)
- Duncan Hull, Paul Fisher, Katy Wolstencroft, Marco Roos (Pioneering Life Science Users)
- Jeremy Frey, Simon Coles (Chemistry pilot)
- Alex Voss, Rob Proctor (National e-Social Science Centre)
- Nosh Contractor (Social networks)
- The myGrid, CombeChem and Taverna teams
- EPSRC, OMII-UK, JISC
- http://www.mygrid.org.uk
- http://taverna.sourceforge.net
- http://myexperiment.org

Want to play?

http://myexperiment.org
special word: dolphin

For beta release

my experiment

Extras