MANCHESTER 1824

The University of Manchester

myGrid

myexperiment

Taverna

Scientific Workflow Management System

BioCatalogue beta

"The Life Science Web Services Registry"

Taverna, Biocatalogue, myExperiment, and the provenance of it all: forward-looking while looking back

Dr. Paolo Missier, Prof. Carole Goble

Information Management Group

School of Computer Science, University of Manchester, UK

with additional material kindly provided by:

Prof. Dave DeRoure, Prof. Luc Moreau, Univ. of Southampton, UK

Andrea Wiggins, Syracuse University, NY

Outline

Part I: models and technology for e-science

1. Addressing the needs of the e-scientist:
   - Workflow as a model of experimental science
     - Taverna
     - Services as building blocks
     - Biocatalogue
2. Scaling up along the social dimension:
     - towards open, collaborative science
     - myExperiment

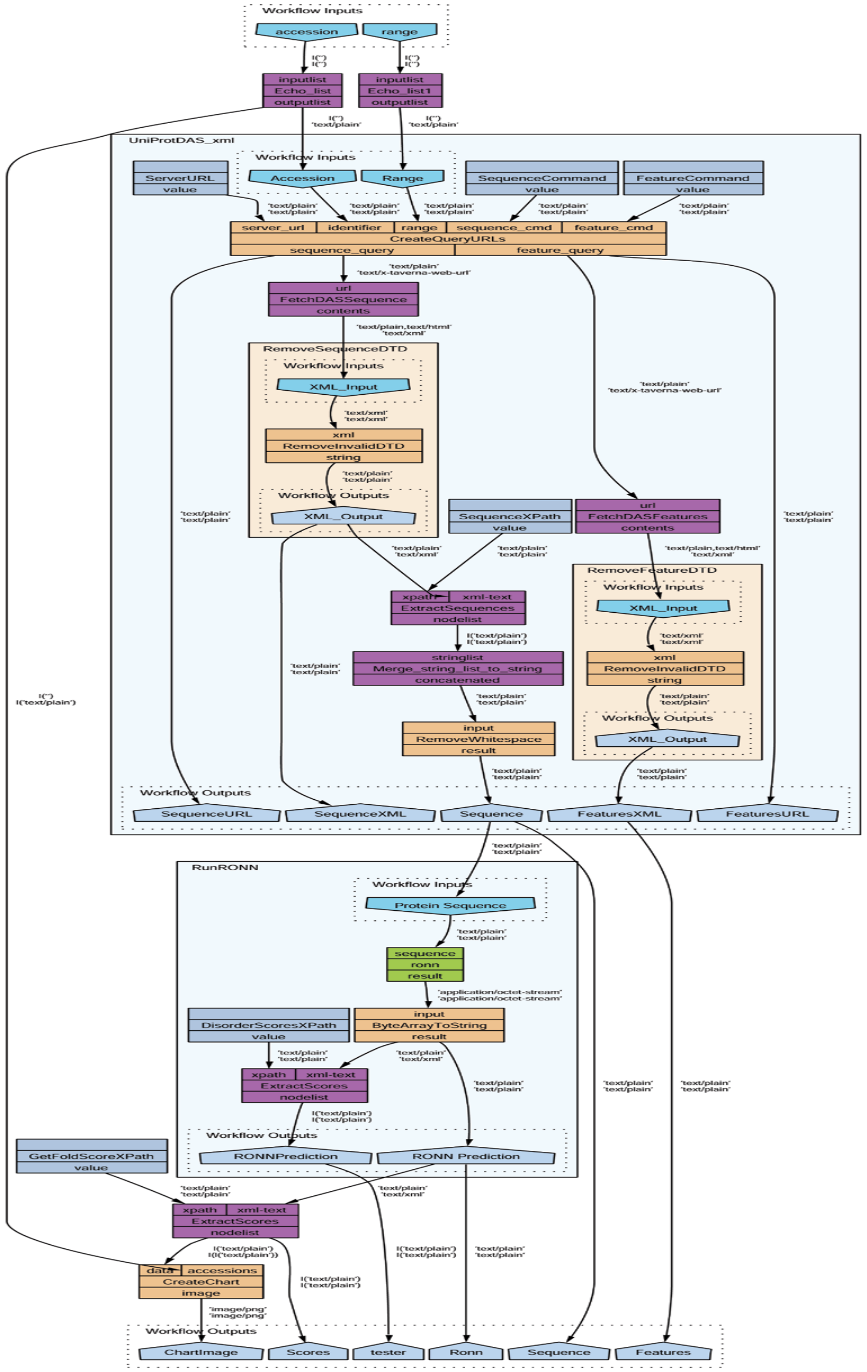
Part II: Explaining and Preserving experimental outcomes

- Data provenance support in Taverna
- provenance for open science: the OPM vision

What is the myGrid Project?

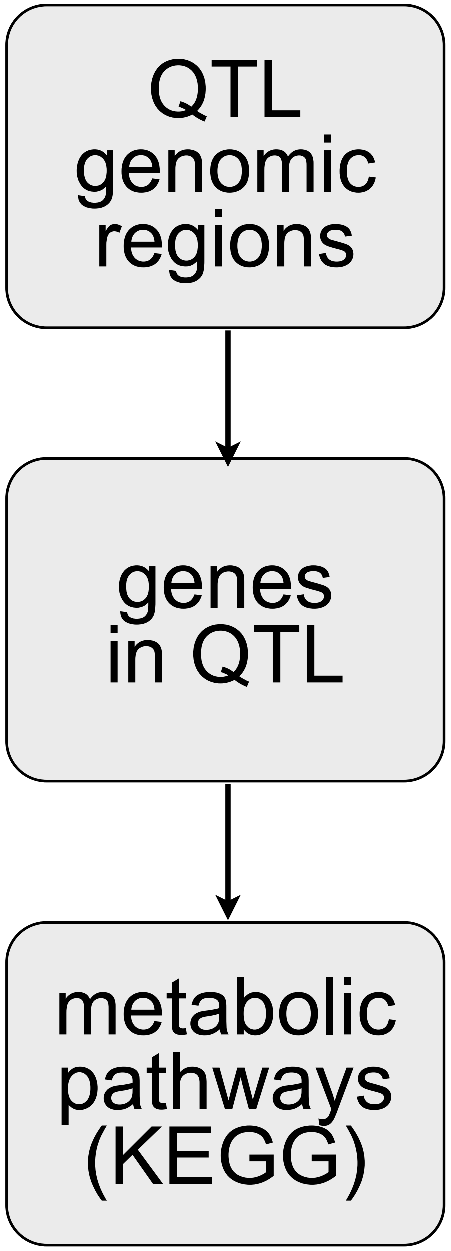
- UK e-Science pilot project since 2001.
- Centred at Manchester, Southampton and the EMBL-EBI
- Part of Open Middleware Infrastructure Institute UK http://www.omii.ac.uk.
- Mixture of developers, bioinformaticians and researchers
- An alliance of contributing projects and partners
- Open source development and content LGPL or BSD
- Infrastructure
- We don’t own any resources (apart from catalogues)
- Or a Grid.

Workflows: *E. Science laboris*



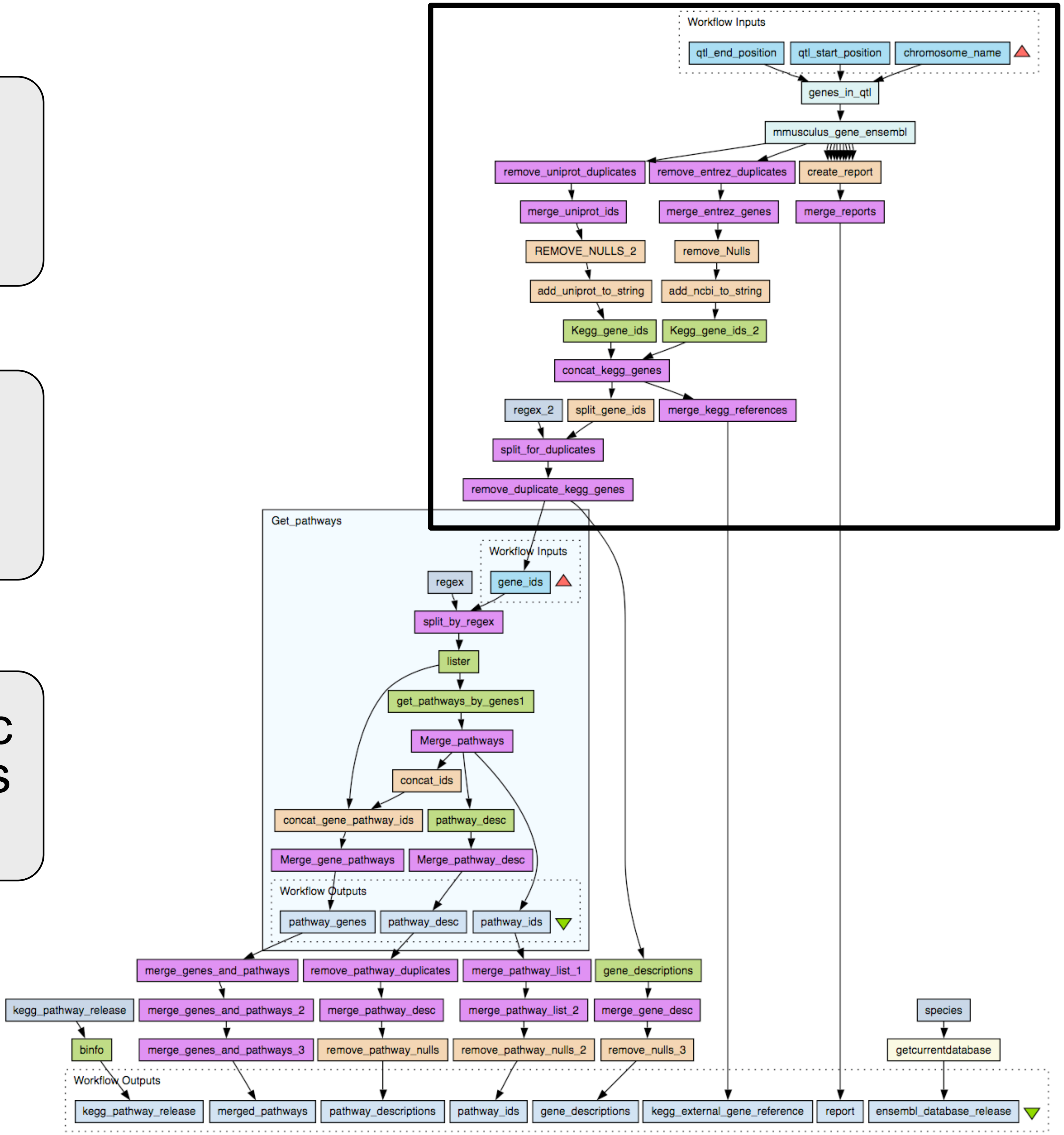
- Pipeline processing
- Automated processing
- Repetitive and mundane boring stuff made easier, reliable and adaptable.
- Shield interoperability horror
- Trackable results
- Agile software development
- Big science, small science & collaborative science

Workflow as data integrator

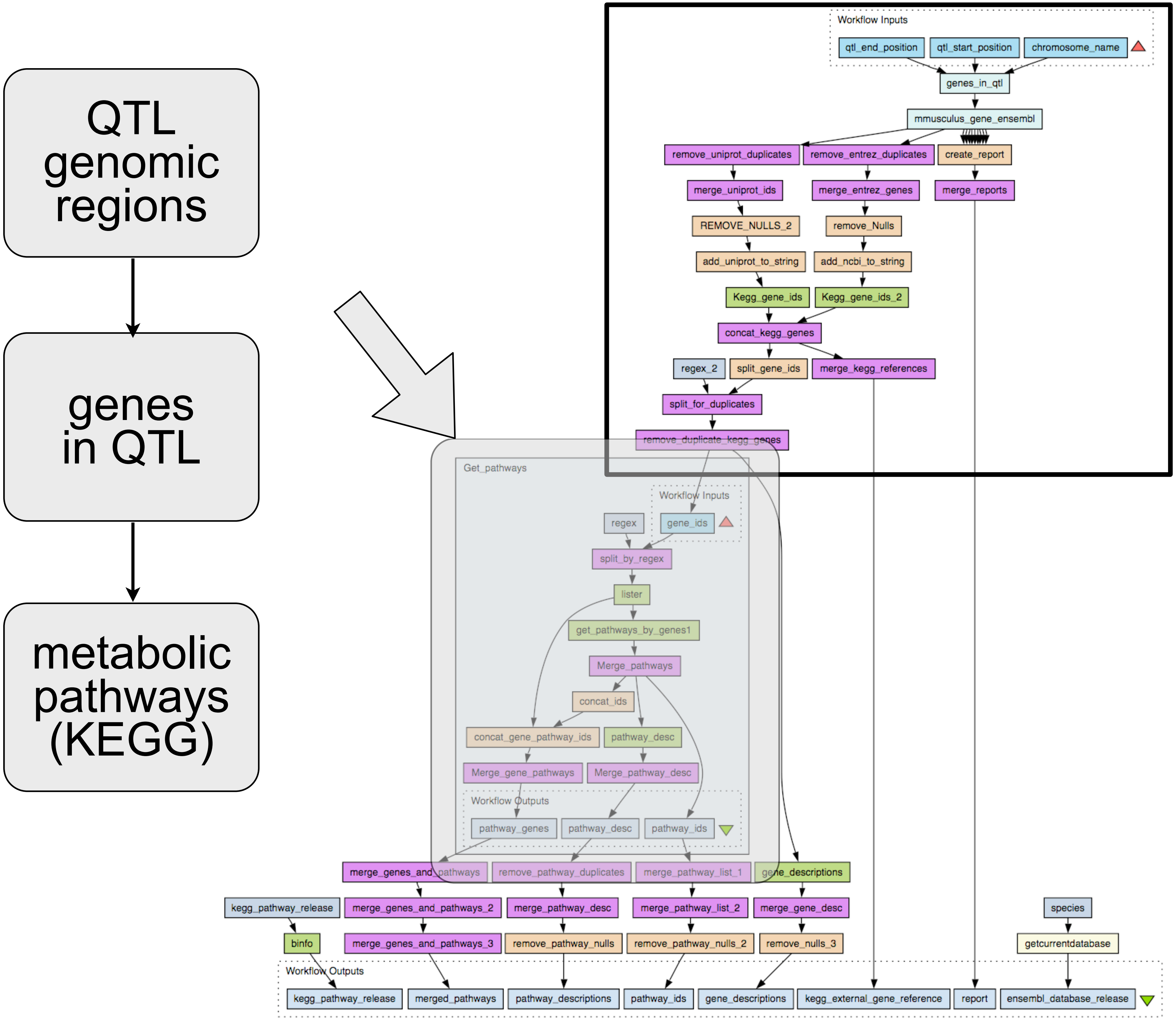

QTL genomic regions

genes in QTL

metabolic pathways (KEGG)



Workflow as data integrator



QTL genomic regions

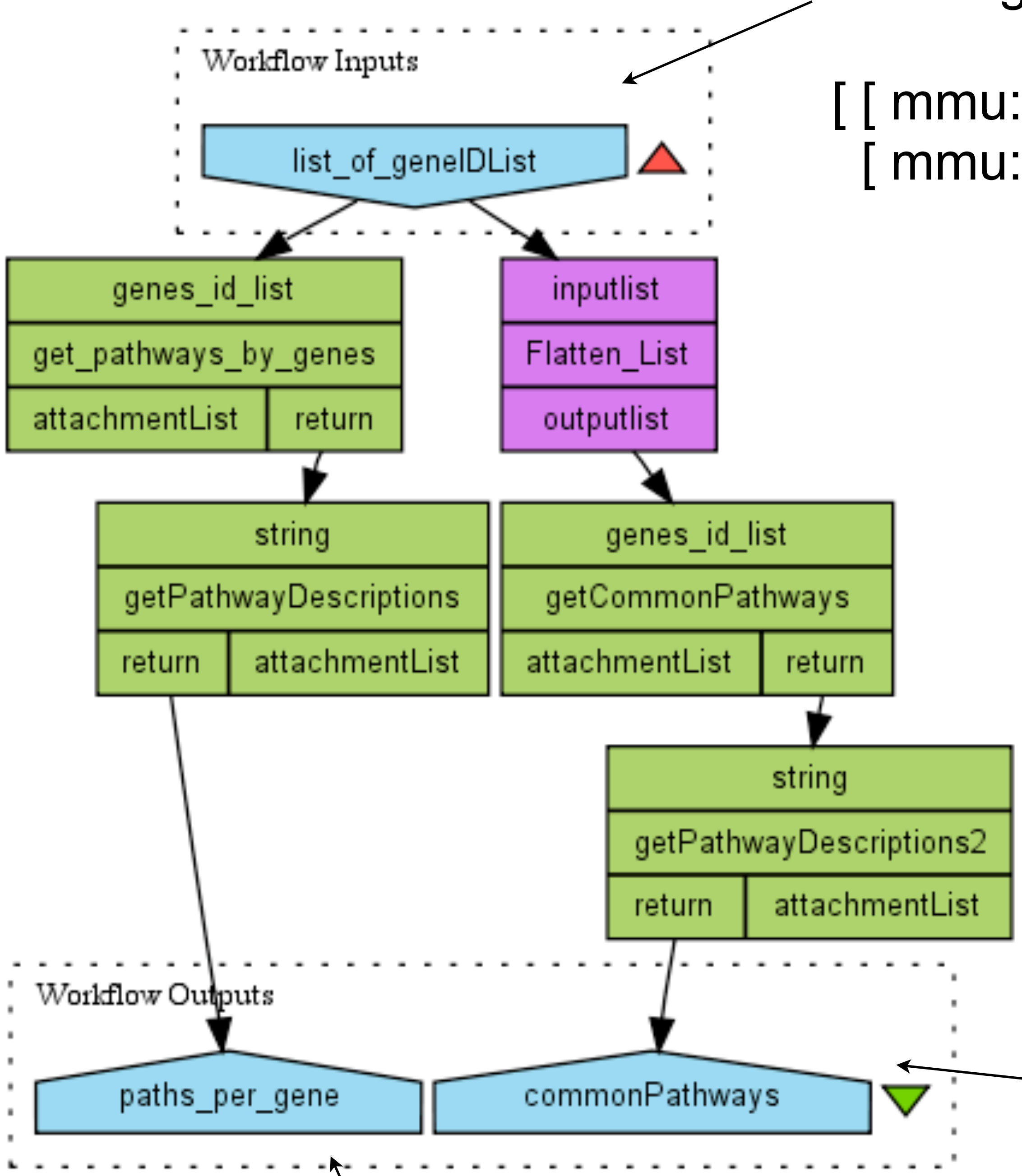
genes in QTL

metabolic pathways (KEGG)

Data-driven computation in Taverna

List-structured KEGG gene ids:

[[mmu:26416],
[mmu:328788]]



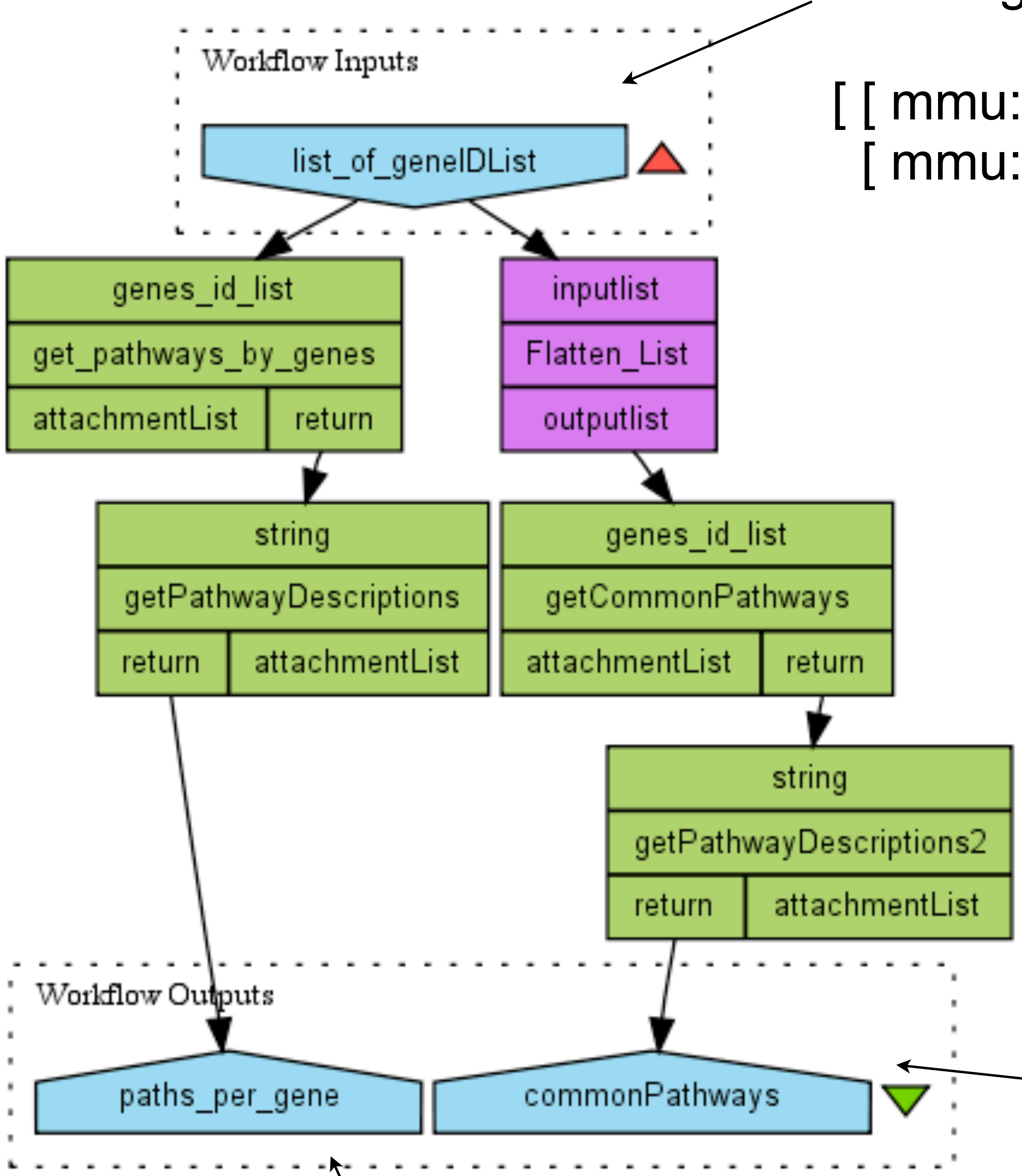
[path:mmu04010 MAPK signaling,
path:mmu04370 VEGF signaling]

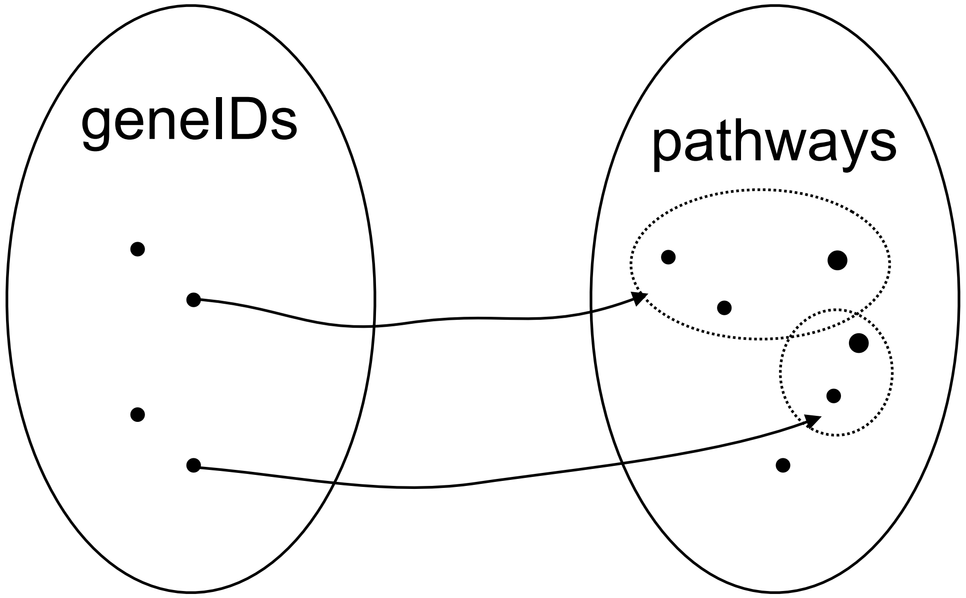
[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
[path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Data-driven computation in Taverna

List-structured KEGG gene ids:

[[mmu:26416],
  [mmu:328788]]

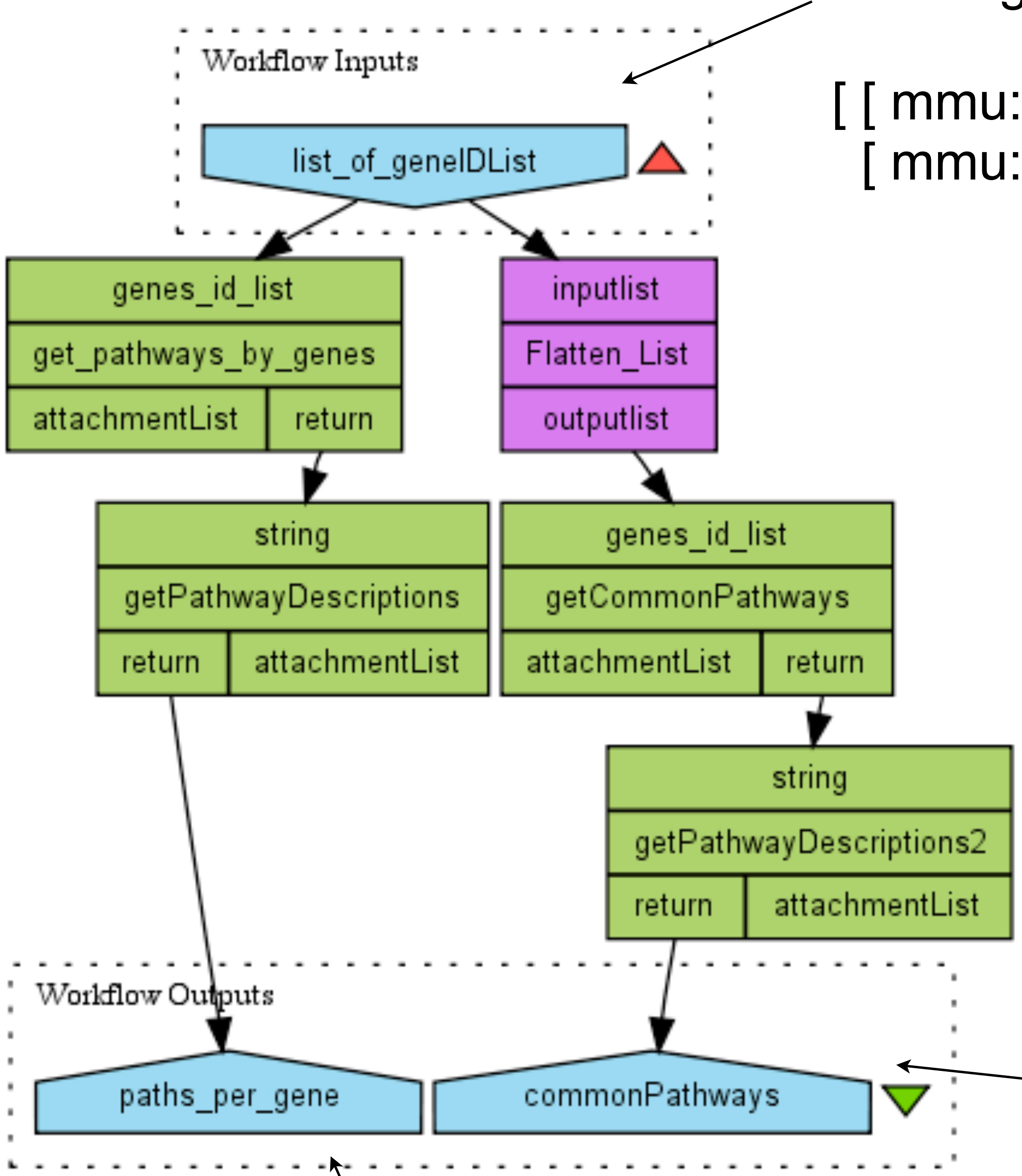




[path:mmu04010 MAPK signaling,
  path:mmu04370 VEGF signaling]

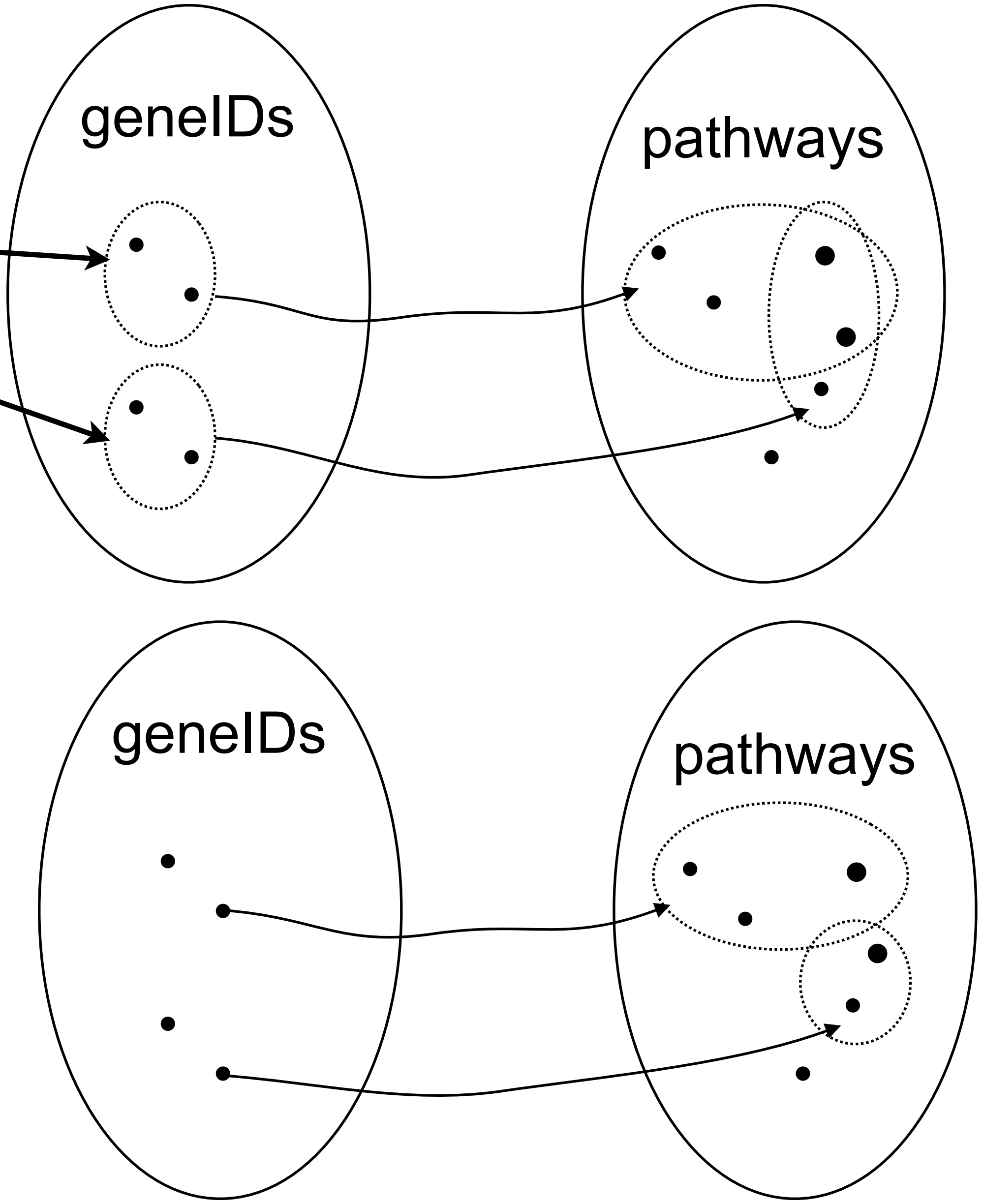
[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
  [path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Data-driven computation in Taverna



List-structured KEGG gene ids:

[[mmu:26416],
  [mmu:328788]]



[path:mmu04010 MAPK signaling,
  path:mmu04370 VEGF signaling]

[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
  [path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

What do Scientists use Taverna for?

Systems biology model building
Proteomics
Sequence analysis
Protein structure prediction
Gene/protein annotation
Microarray data analysis
QTL studies
QSAR studies
Medical image analysis
Public Health care epidemiology
Heart model simulations
High throughput screening
Phenotypical studies
Phylogeny
    Statistical analysis
    Text mining

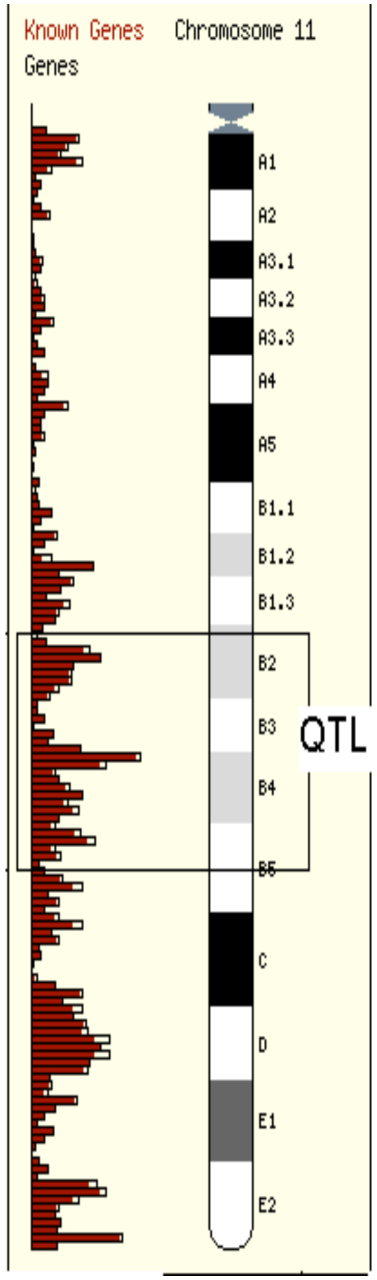
Astronomy, Music, Meteorology

Netherlands Bioinformatics Centre
Genome Canada Bioinformatics Platform
BioMOBY
US FLOSS social science program
RENCI
SysMO Consortium
French SIGENAE farm animals project
ThaiGrid
CARMEN Neuroscience project
SPINE consortium
EU Enfin, EMBRACE, BioSapian, Casimir
EU SysMO Consortium
NERC Centre for Ecology and Hydrology
Bergen Centre for Computational Biology
Max-Planck institute for Plant Breeding Research
Genoa Cancer Research Centre
AstroGrid

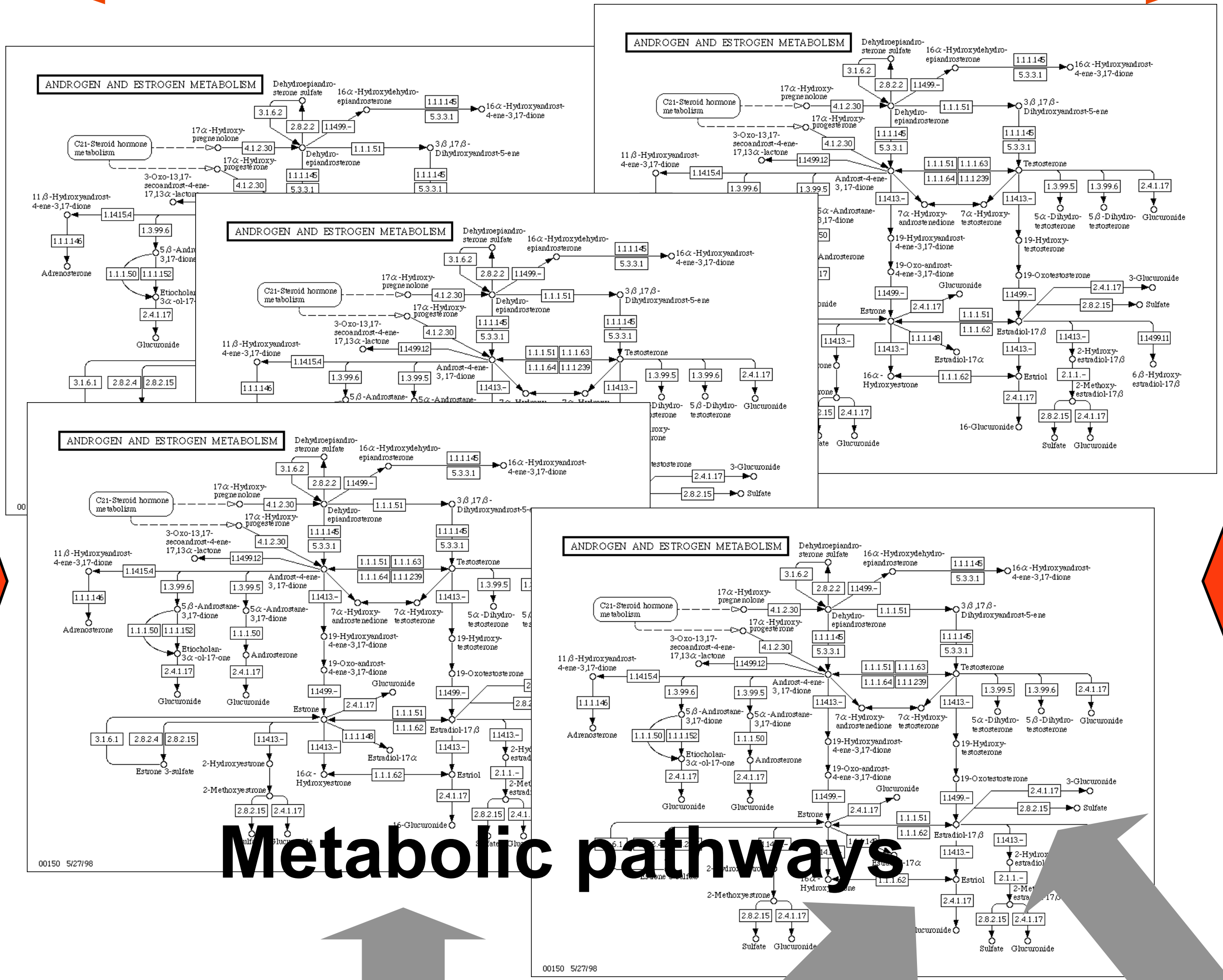
30 USA academic and research institutions

Genotype

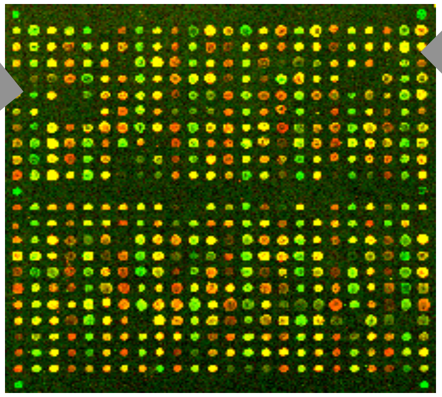
Phenotype

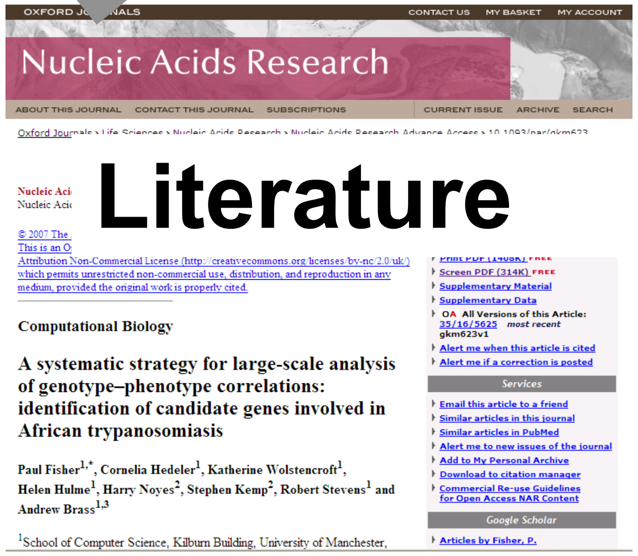


200



Metabolic pathways





OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences > Nucleic Acids Research > Nucleic Acids Research Advance Access > 10.1093/nar/gkm623

Nucleic Aci
Nucleic Acic

© 2007 The
This is an O
Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Computational Biology

A systematic strategy for large-scale analysis of genotype–phenotype correlations: identification of candidate genes involved in African trypanosomiasis

Paul Fisher[1,*], Cornelia Hedeler[1], Katherine Wolstencroft[1], Helen Hulme[1], Harry Noyes[2], Stephen Kemp[2], Robert Stevens[1] and Andrew Brass[1,3]

[1]School of Computer Science, Kilburn Building, University of Manchester,

- Print PDF (1408K) FREE
- Screen PDF (314K) FREE
- Supplementary Material
- Supplementary Data
- OA All Versions of this Article: 35/16/5625 most recent gkm623v1
- Alert me when this article is cited
- Alert me if a correction is posted

Services

- Email this article to a friend
- Similar articles in this journal
- Similar articles in PubMed
- Alert me to new issues of the journal
- Add to My Personal Archive
- Download to citation manager
- Commercial Re-use Guidelines for Open Access NAR Content

Google Scholar

- Articles by Fisher, P.

Literature

[Paul Fisher]

Workflows operating over Grid Infrastructure

KnowARC integrated with Taverna" application prototype to use Taverna as direct interface to Grid resources running ARC.

http://www.knowarc.eu

Open source grid software infrastructure aimed at enabling multi-institutional data sharing and analysis. Underpins caBIG. Taverna links together caGrid resources.

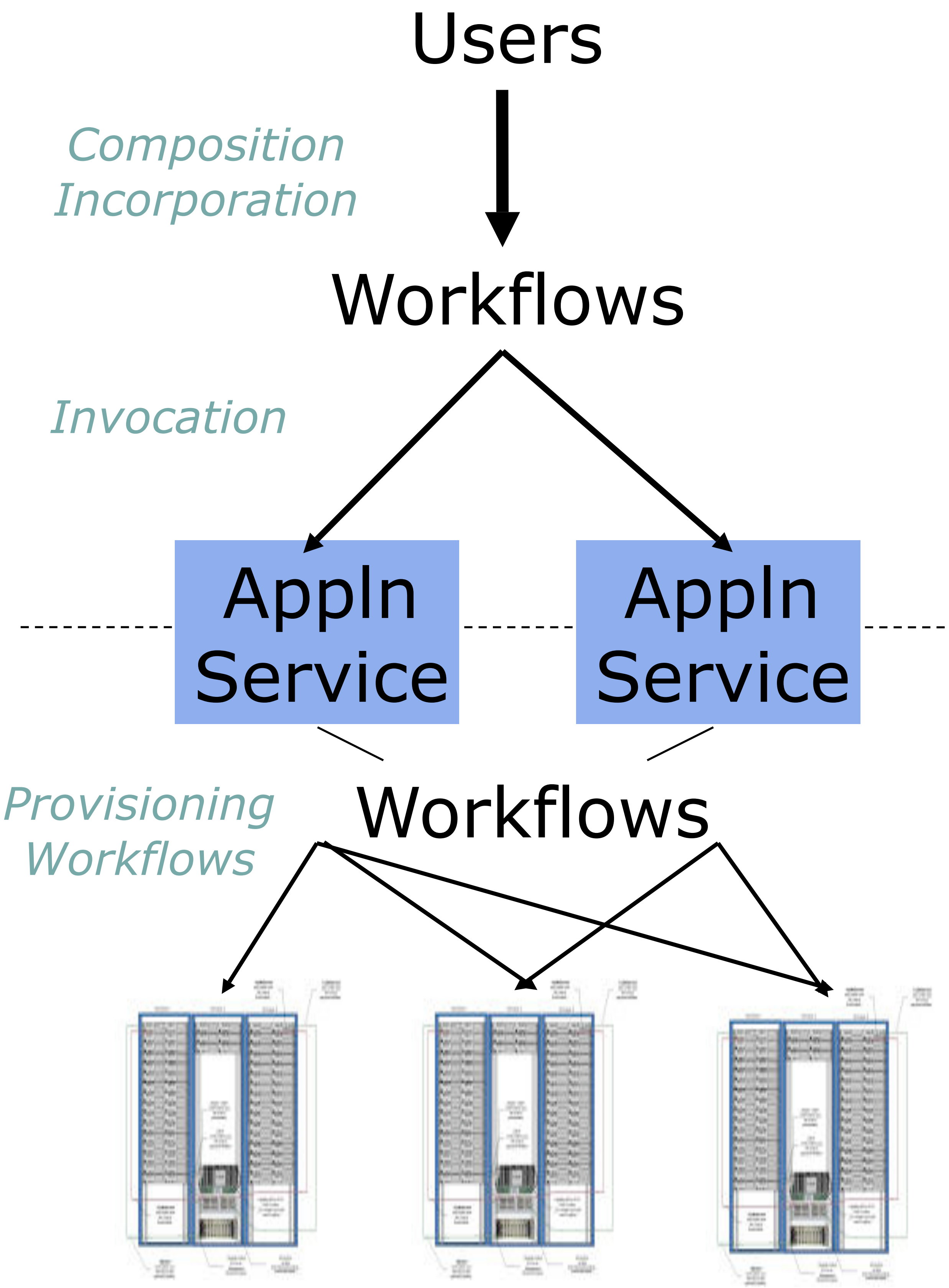
http://cagrid.org/

Europe's leading grid computing project, Piloted Taverna over EGEE gLite services

http://www.eu-egee.org/

[Foster 2005]

- Service-oriented **applications**
  - Applications components of workflows
  - Compose applications into workflows
  - Incorporate workflows into applications



Users

Composition
Incorporation

Workflows

Invocation

Appln Service

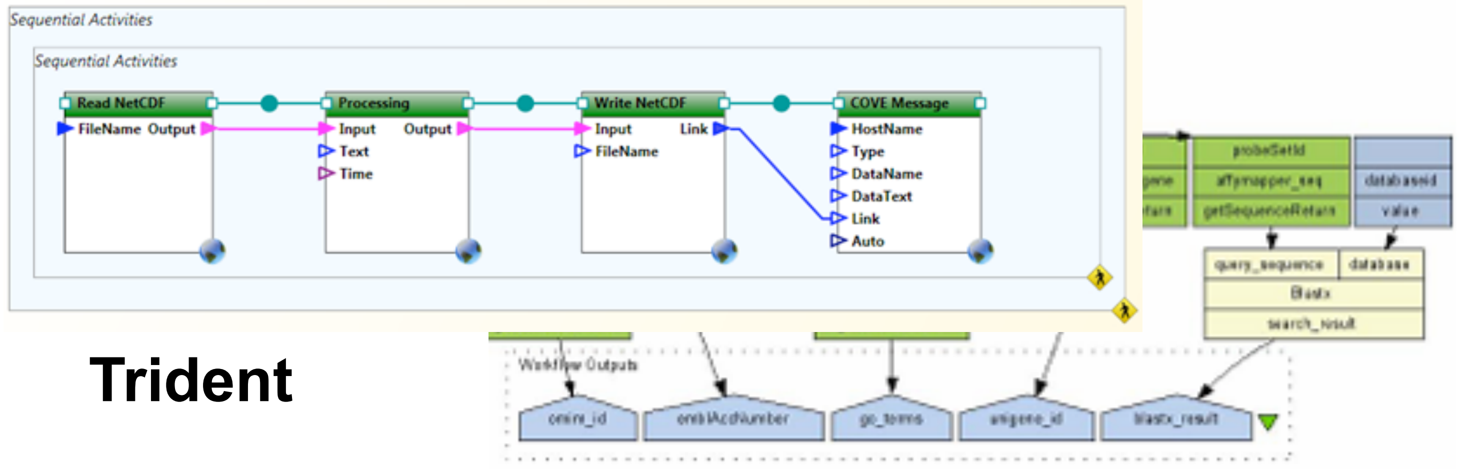
Appln Service

Provisioning
Workflows

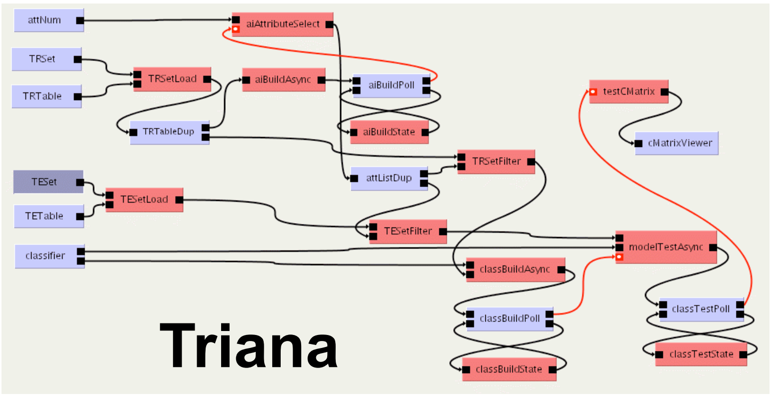
Workflows

- Service-oriented **Grid Infrastructure**
  - Provision physical resources to support application workflows
  - Coordinate resources through workflows

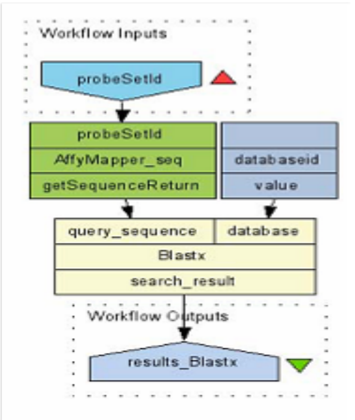
Who else is in this space?

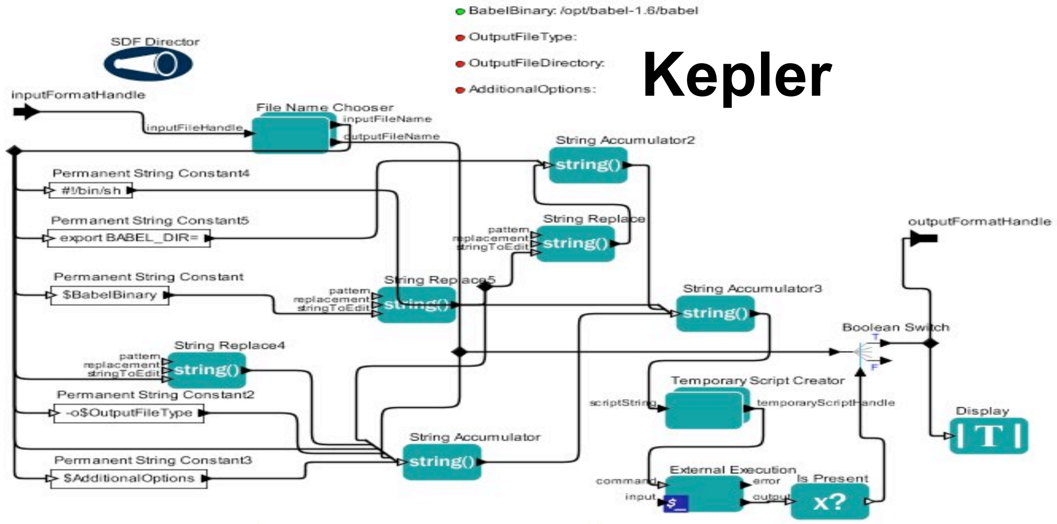


Trident

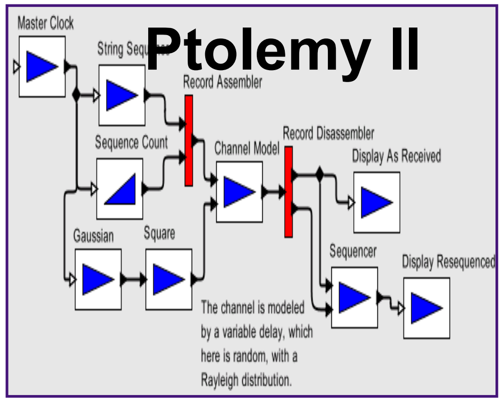


Triana



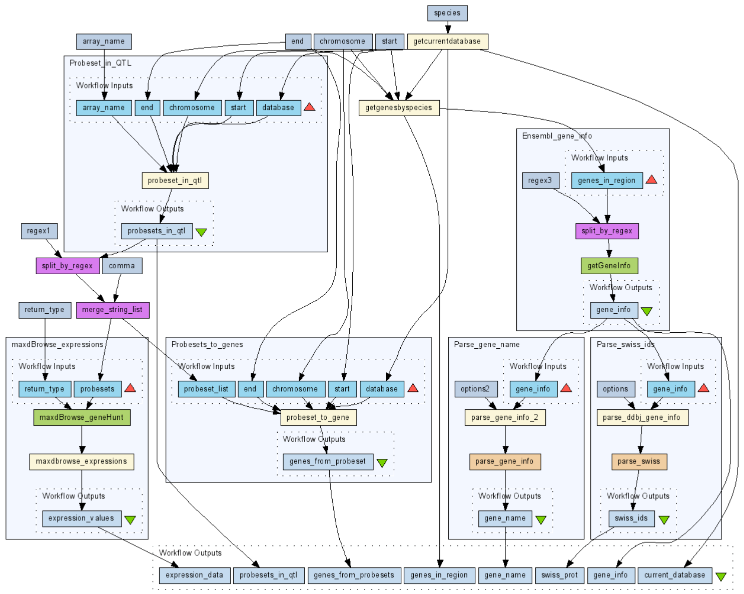


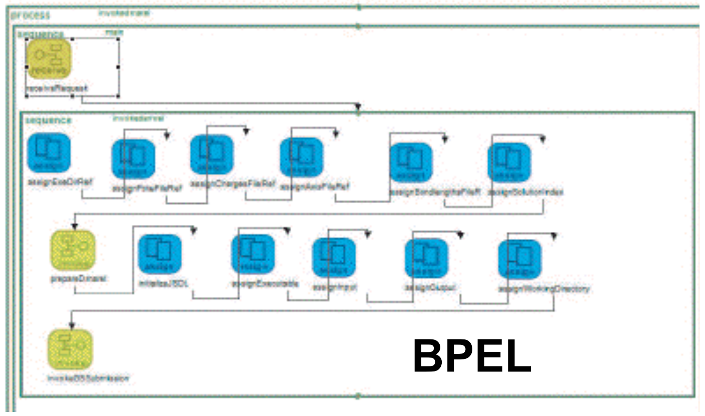
Kepler



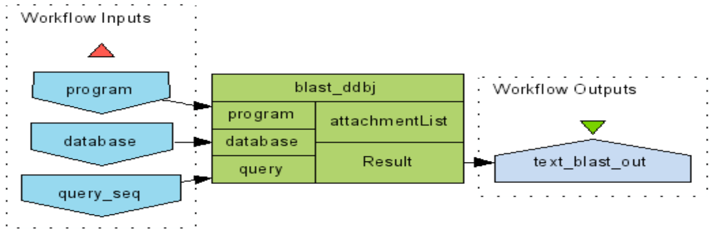
Ptolemy II

Taverna

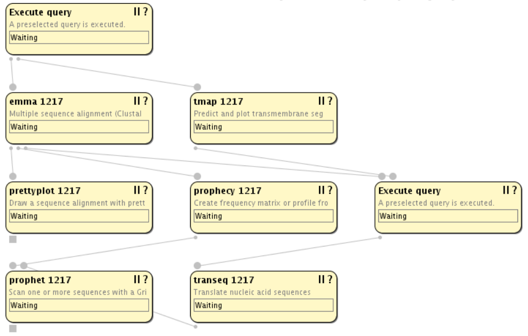




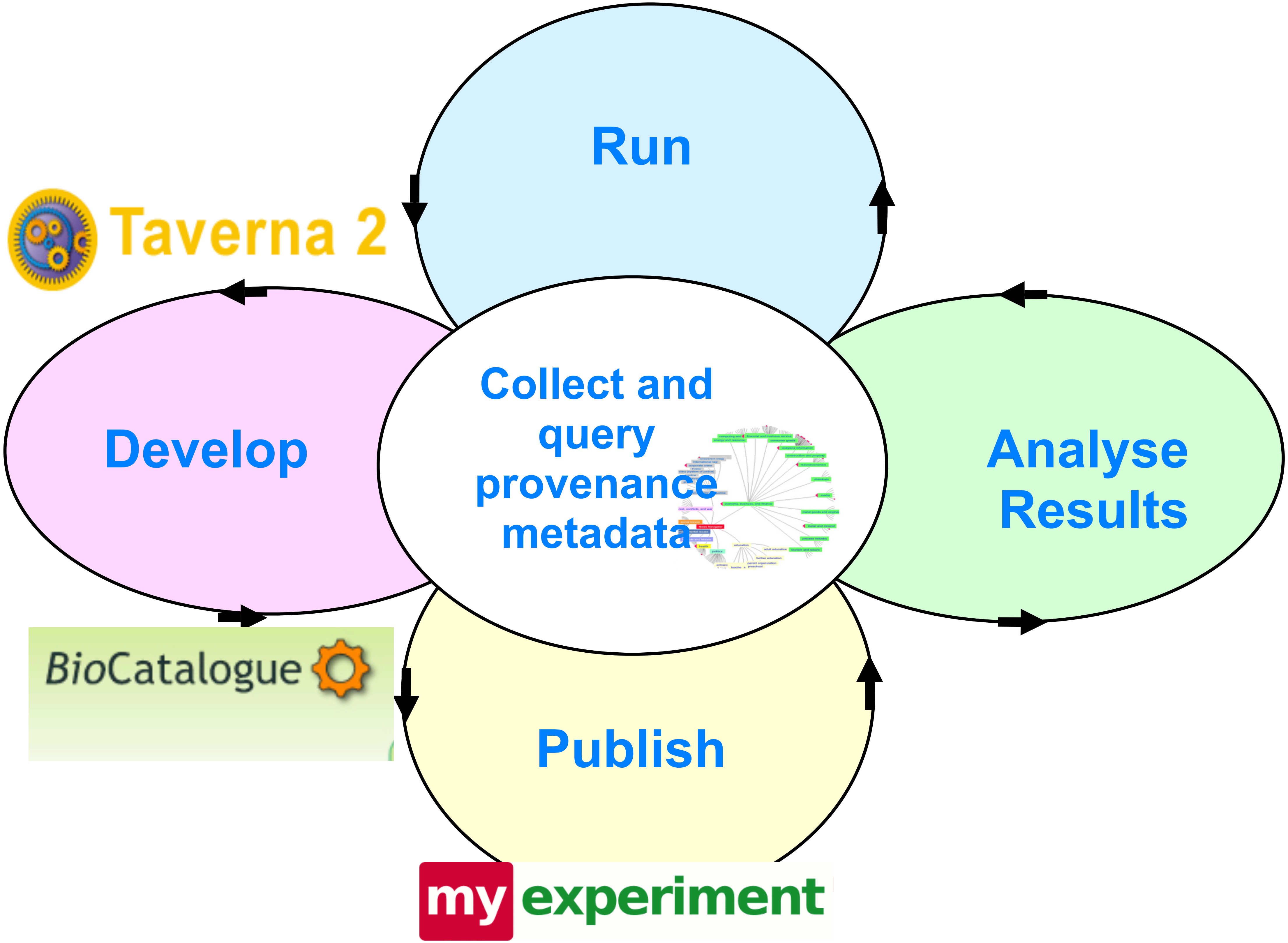
BPEL



BioExtract



Workflow-based experimentation lifecycle



Run

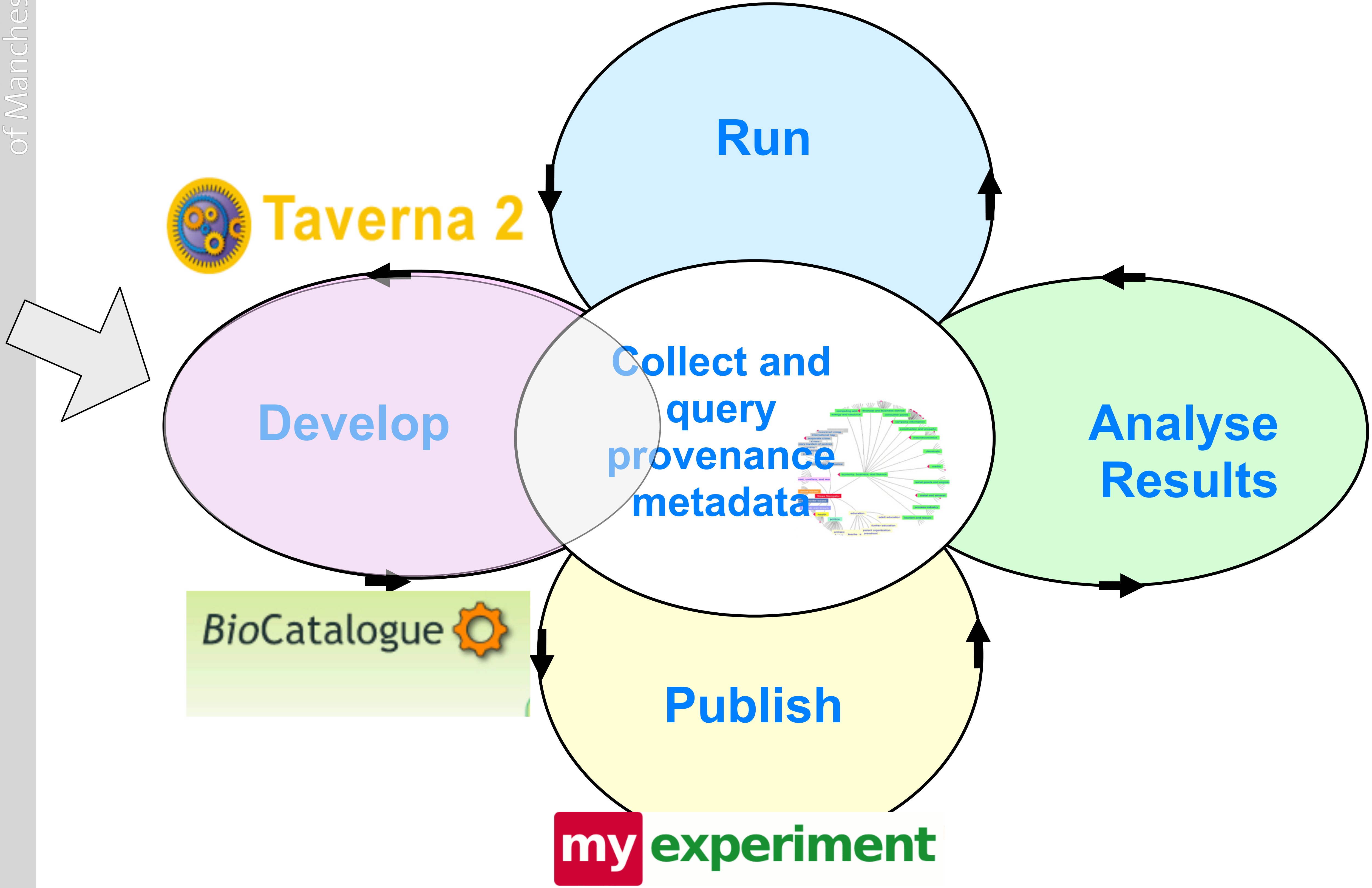
Develop

Collect and query provenance metadata

Analyse Results

Publish

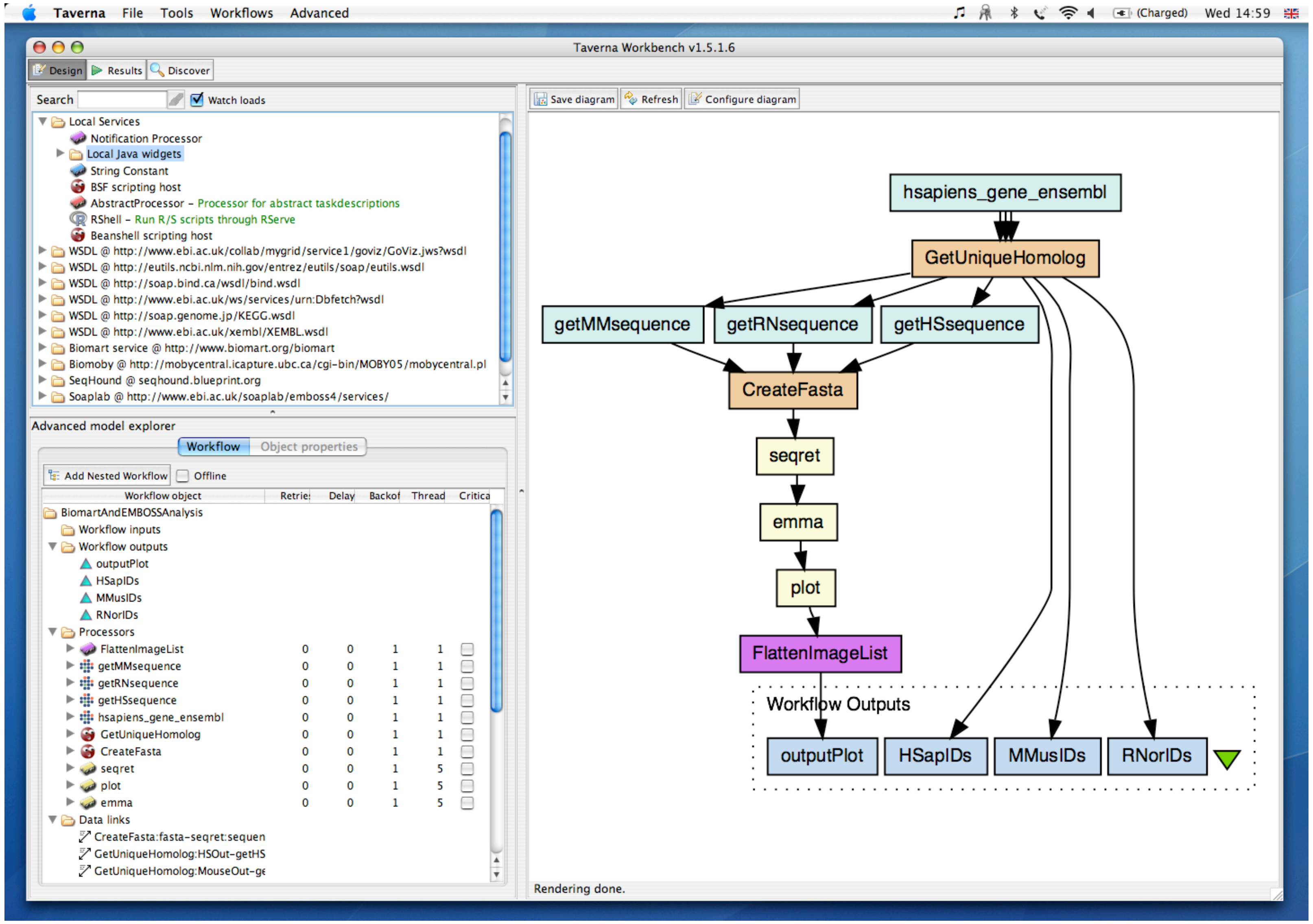
Workflow-based experimentation lifecycle



Taverna

Graphical Workbench For Professionals

Plug-in architecture
Nested Workflows
Drag and Drop
Wiring together



- 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/wsdl/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/

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	□
plot	0	0	1	5	□
emma	0	0	1	5	□
Data links					
CreateFasta:fasta-seqret:sequen					
GetUniqueHomolog:HSOut-getHS					
GetUniqueHomolog:MouseOut-ge					

Rapidly incorporate new service without coding.

Not restricted to predetermined services

Access to local and remote resources and analysis tools

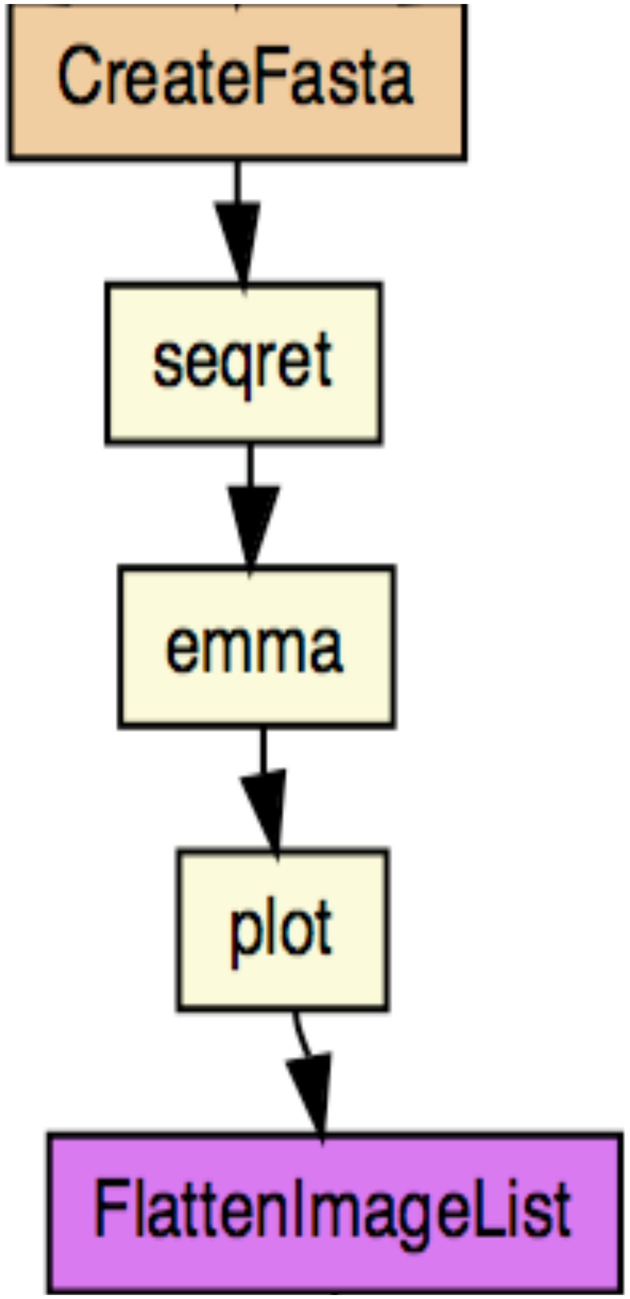
3500+ service operations available when start up

Services Mutability

implications for

sustainability, accountability and reproducability

- Reliability and robustness of workflows depends on the reliability and robustness of the components
- In house service support
- Services in constant and (silent) change.
- Versioning.
- Workflow Decay
- Monitoring and Repair of wrappers, shims and service substitutions.



BioCatalogue

BioCatalogue beta

"The Life Science Web Services Registry"

http://www.biocatalogue.org

http://beta.biocatalogue.org

Professor Carole Goble
University of Manchester, UK
Director myGrid Consortium

28 April 2009, Boston MA

Data curation + process curation=data integration + science

Briefings in Bioinformatics, doi:10.1093/bib/bbn034 (Dec., 2008)

Carole Goble, Robert Stevens, Duncan Hull, Katy Wolstencroft and Rodrigo Lopez

The short story

- Public, Curated Catalogue of Life Science Web Services
- Register, Find, Curate Web Services
- Community-sourced annotation, expert oversee
- Open content

- Open platform with open REST interfaces
- Web 2.0 site and development.
- Open source code base.

- Started June 2008. In first beta phase.
- Launch**ed** June 2009 at ISMB.
- www.biocatalogue.org

BioCatalogue beta
"The Life Science Web Services Registry"

The short story

- Public, Curated Catalogue of Life Science Web Services
- Register, Find, Curate Web Services
- Community-sourced annotation, expert oversee
- Open content

- Open platform with open REST interfaces
- Web 2.0 site and development.
- Open source code base.

- Started June 2008. In first beta phase.
- Launch**ed** June 2009 at ISMB.
- www.biocatalogue.org

BioCatalogue beta

"The Life Science Web Services Registry"

The short story

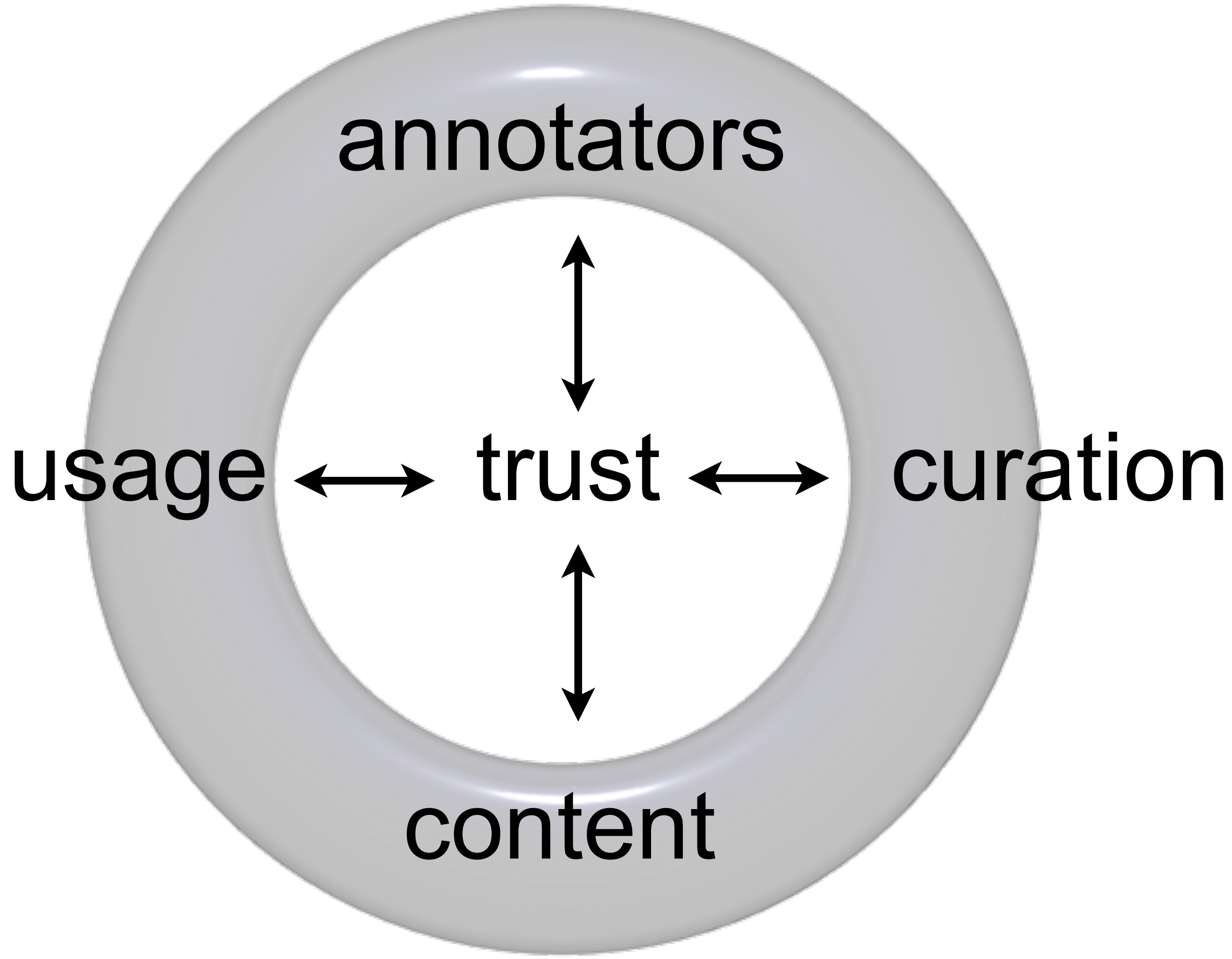
- Public, Curated Catalogue of Life Science Web Services
- Register, Find, Curate Web Services
- Community-sourced annotation, expert oversee
- Open content

- Open platform with open REST interfaces
- Web 2.0 site and development.
- Open source code base.

- ~~Started June 2008. In first beta phase.~~
- Launch**ed** June 2009 at ISMB.
- www.biocatalogue.org

A virtual circle

- curation involves substantial human effort why would it happen at all?



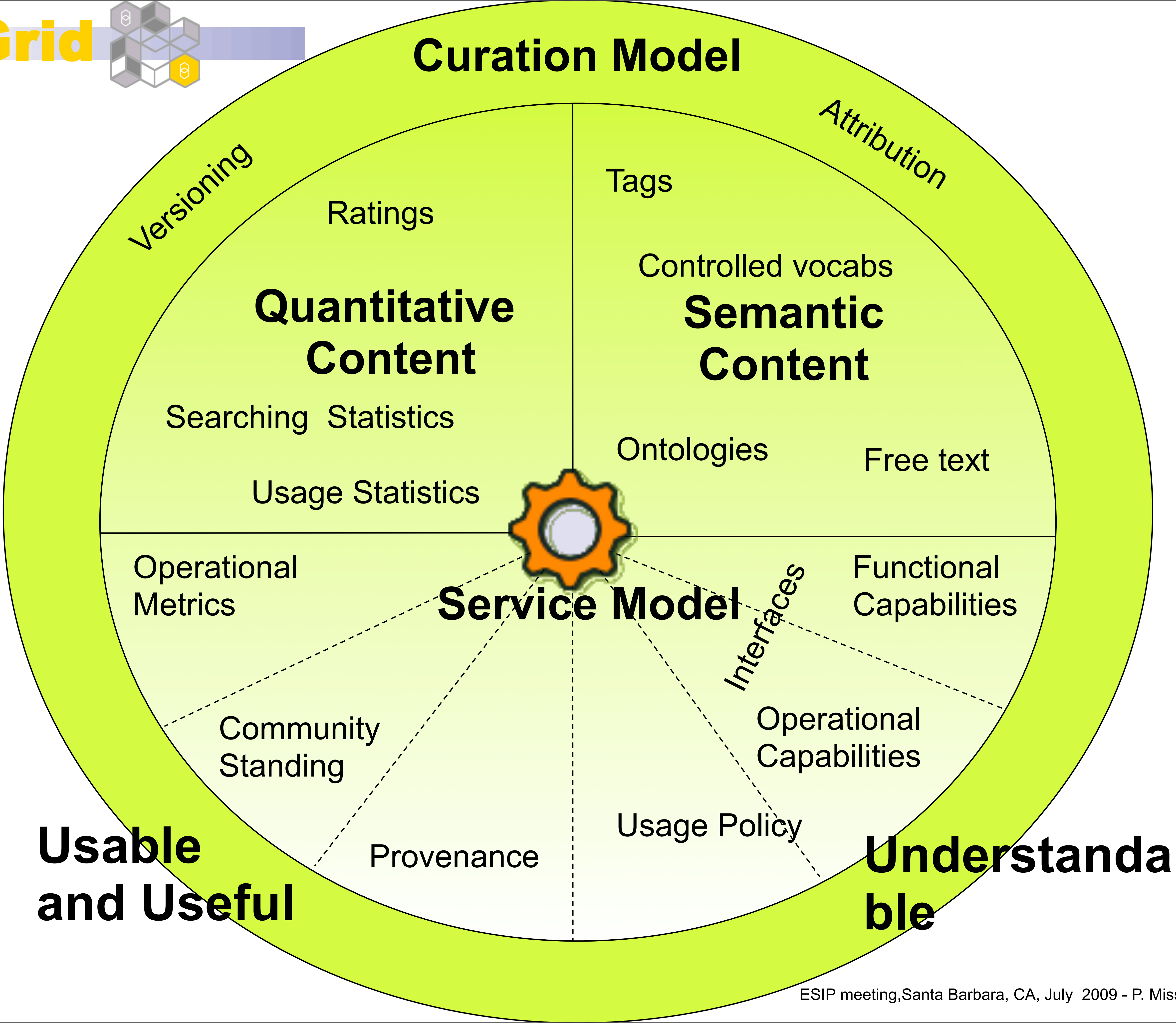
annotators

usage ⟷ trust ⟷ curation

content

myGrid

Curation Model



Versioning

Attribution

Ratings

Quantitative Content

Searching Statistics

Usage Statistics

Tags

Controlled vocabs

Semantic Content

Ontologies

Free text

Service Model

Operational Metrics

Community Standing

Provenance

Usage Policy

Operational Capabilities

Interfaces

Functional Capabilities

Usable and Useful

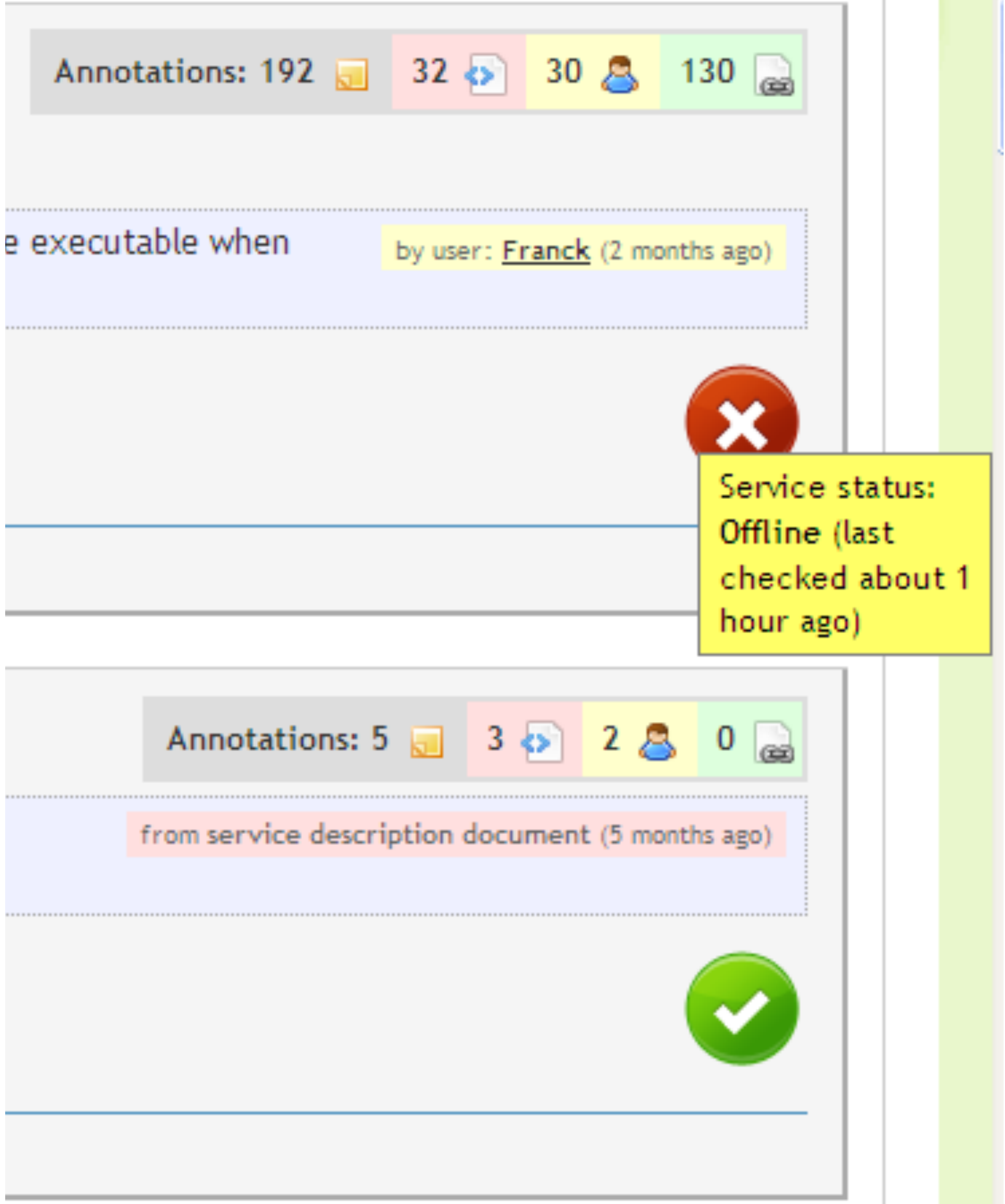
Understandable

BioCatalogue beta

"The Life Science Web Services Registry"

Curation

- Just enough just in time
- Universal annotation scheme
- Mixed: Free text, Tags, controlled vocabs, community ontologies
- Community sourced tags, comments, recommendations
- Expert curation ontology-based annotation. myGrid OWL Ontology
- Automated WSDL ripping and analytics
- Automated monitoring & testing
- Partner feeds (e.g. myExperiment)
- Update feeds to users

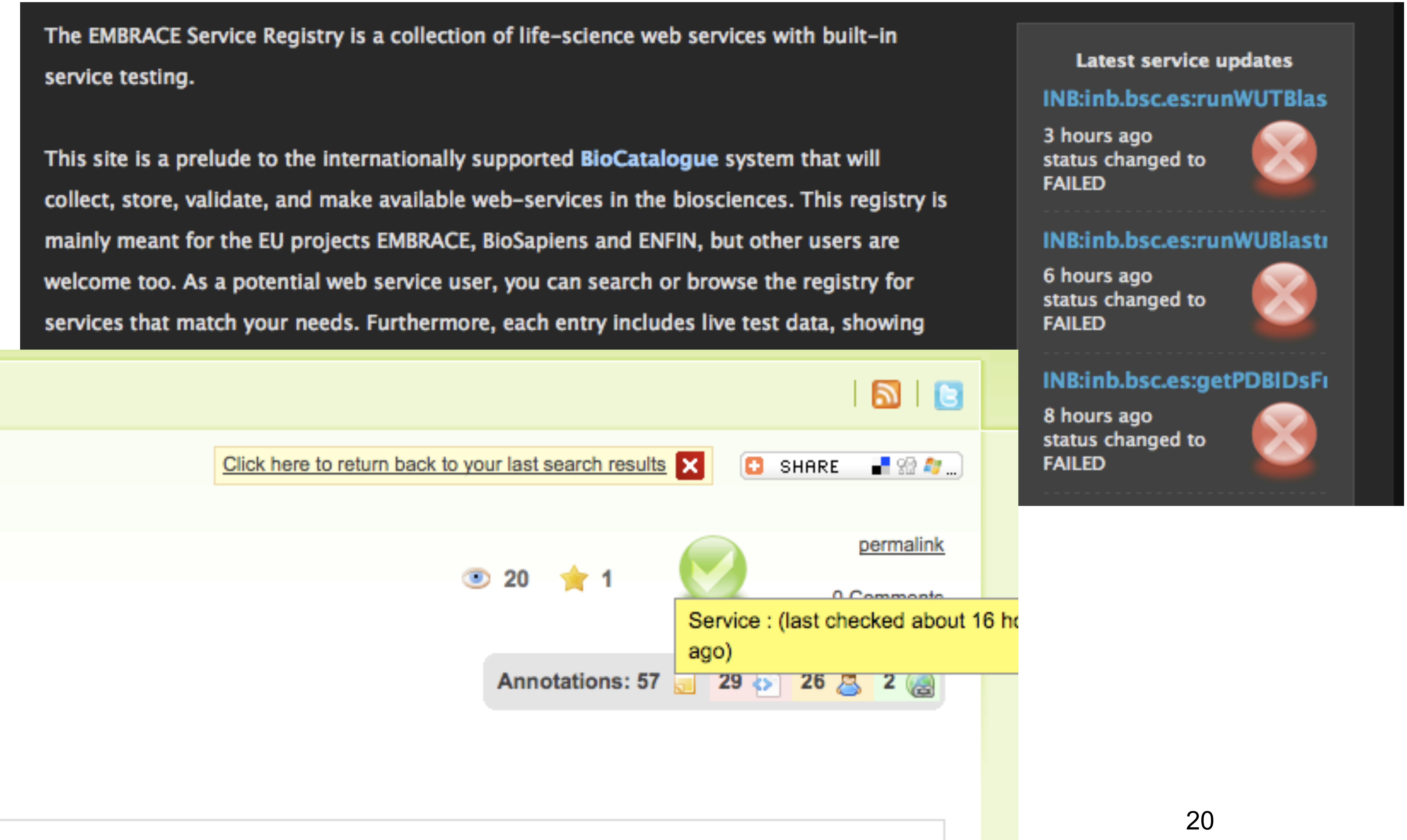

Annotations: 192
32
30
130
e executable when
by user: Franck (2 months ago)
Annotations: 5
3
2
0
from service description document (5 months ago)

Service status: Offline (last checked about 1 hour ago)

Today: 14902 annotations (provider, user, registries)
KEGG: 1433 annotations

BioCatalogue beta
"The Life Science Web Services Registry"

Service monitoring



The EMBRACE Service Registry is a collection of life-science web services with built-in service testing.

This site is a prelude to the internationally supported BioCatalogue system that will collect, store, validate, and make available web-services in the biosciences. This registry is mainly meant for the EU projects EMBRACE, BioSapiens and ENFIN, but other users are welcome too. As a potential web service user, you can search or browse the registry for services that match your needs. Furthermore, each entry includes live test data, showing

Latest service updates

INB:inb.bsc.es:runWUTBlas
3 hours ago status changed to FAILED

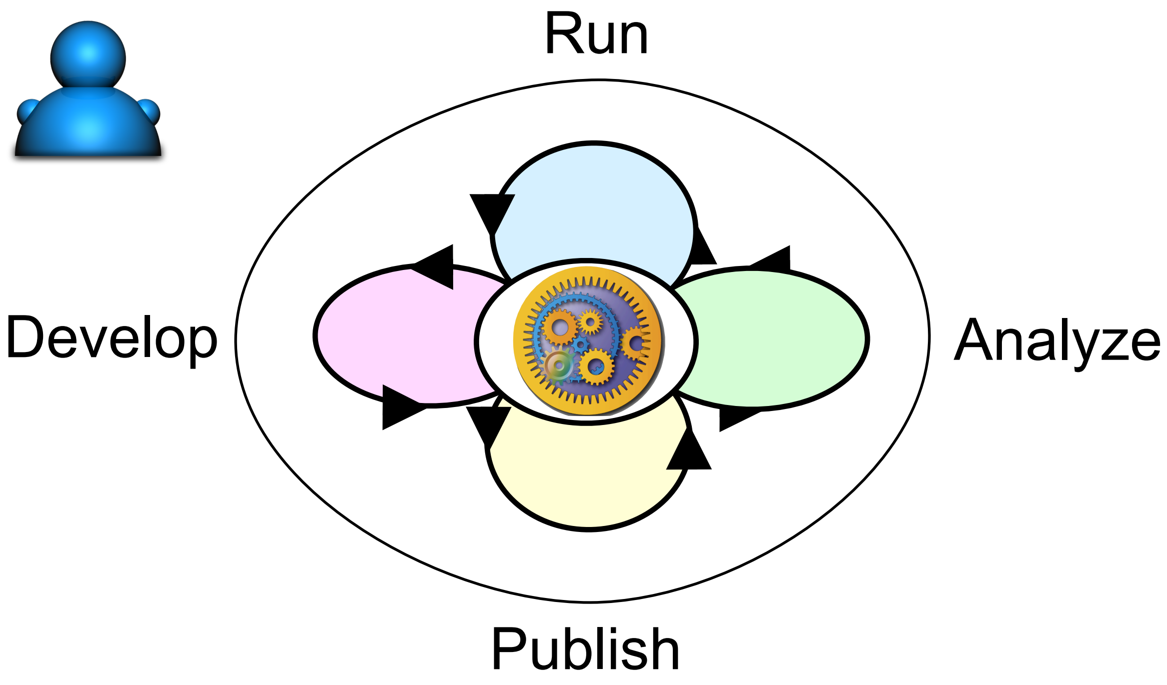
INB:inb.bsc.es:runWUBlast
6 hours ago status changed to FAILED

INB:inb.bsc.es:getPDBIDsF
8 hours ago status changed to FAILED

Service : (last checked about 16 h ago)

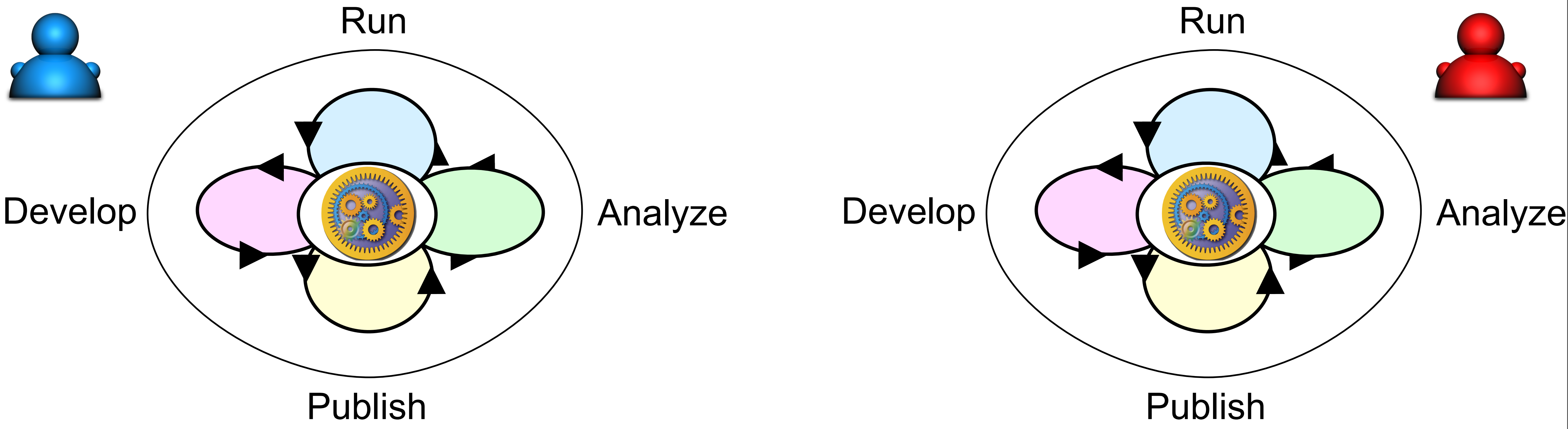
Scaling up along the social dimension

Crossing the boundaries of individual investigation



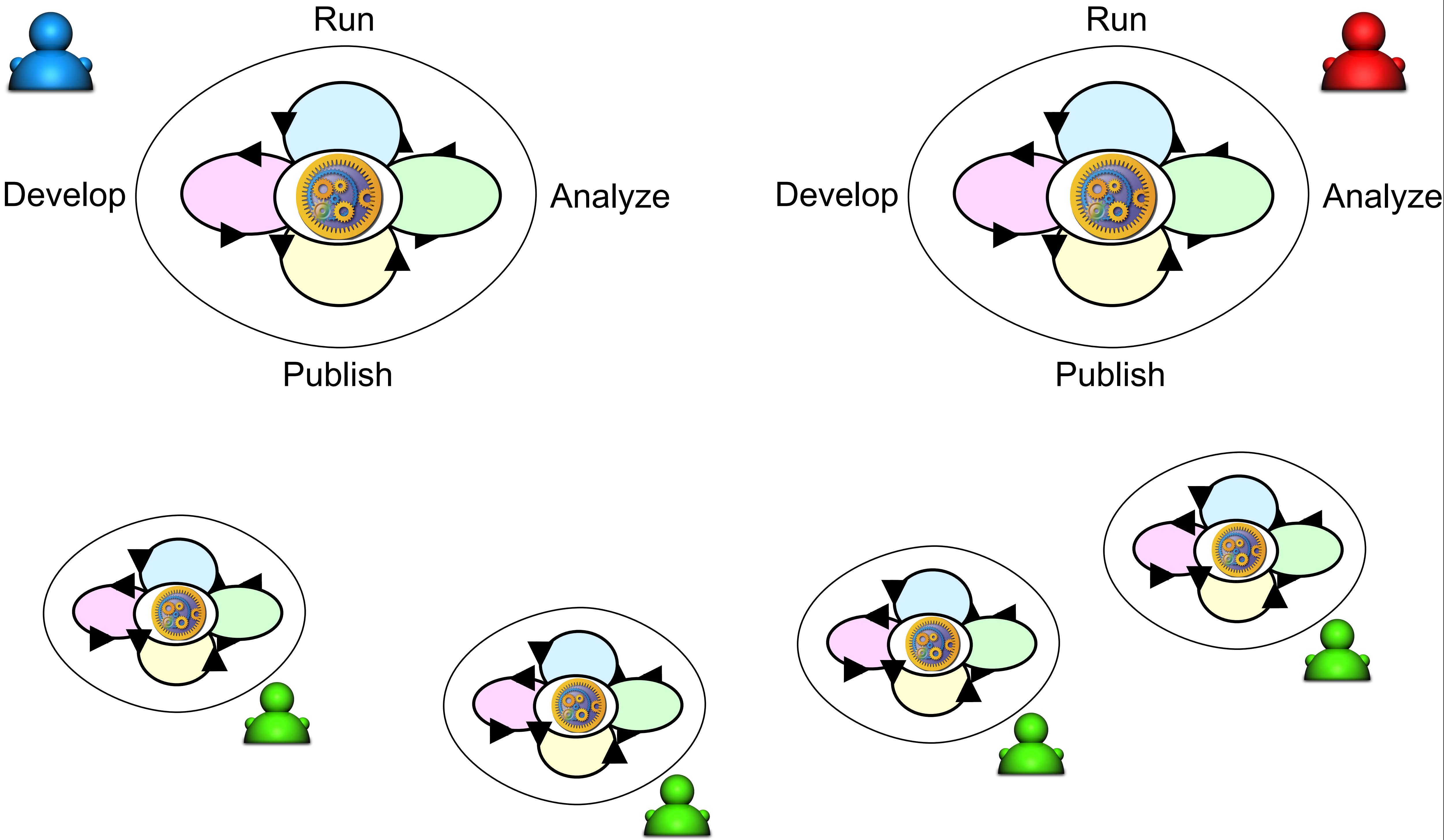
Scaling up along the social dimension

Crossing the boundaries of individual investigation



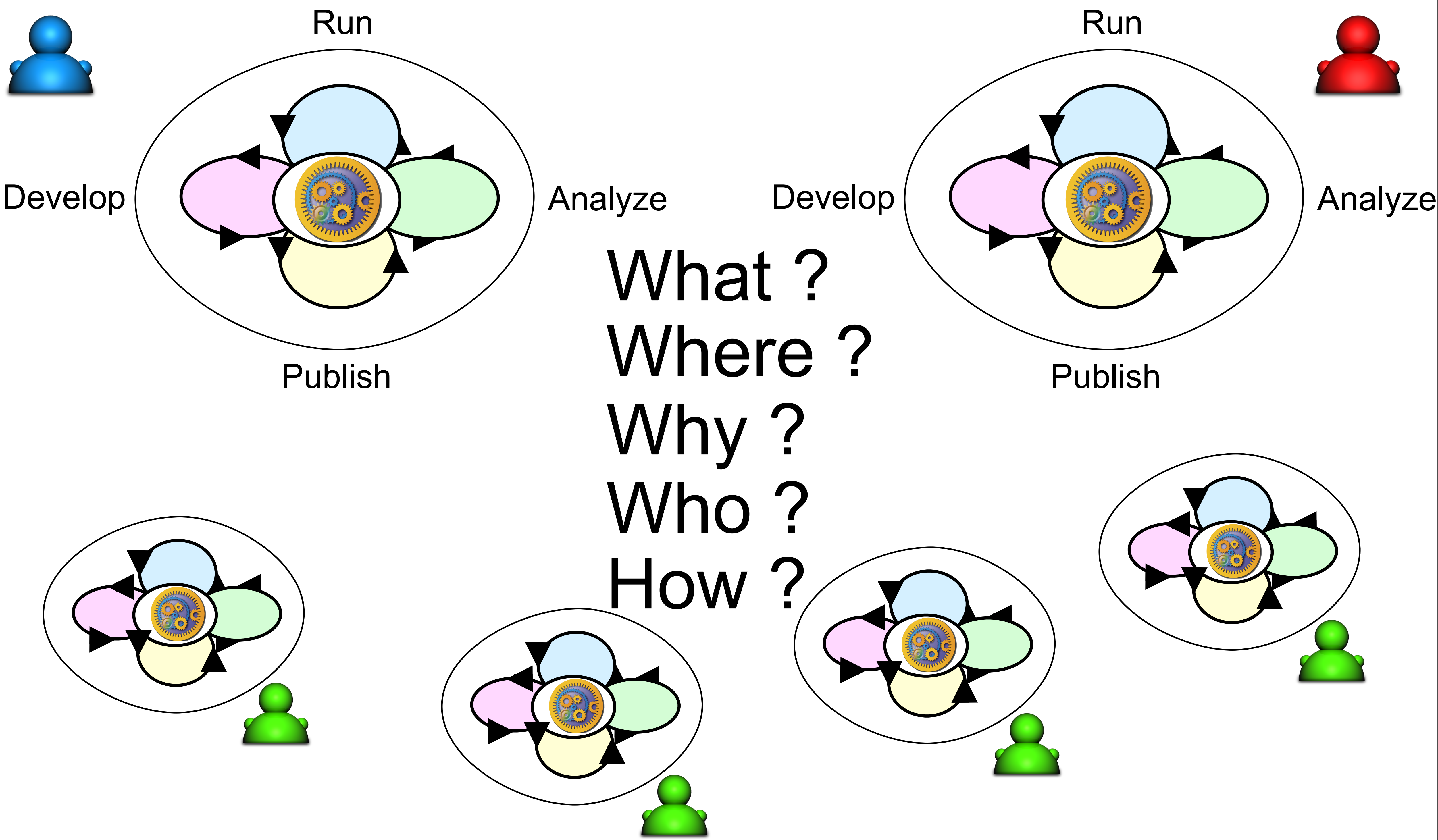
Scaling up along the social dimension

Crossing the boundaries of individual investigation



Scaling up along the social dimension

Crossing the boundaries of individual investigation



What ?
Where ?
Why ?
Who ?
How ?

Scientific Collaboration Requirements

- Shared goals
  - Establishes focus of research
- Shared research resources
  - Both social and artifactual
  - Social aspects include training and community socialization

http://www.flickr.com/photos/ryanr/142455033/

Source: **Andrea Wiggins**, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009

Historical Research Artifacts

- Letters, Books, Journals, Lectures
- Also technologies: methods, instrumentation
- Sharing?
  - Recordkeeping is not always a researcher's main priority
  - Without records, there's not much to share except the research outputs

http://www.flickr.com/photos/smailtronic/1535870363/

Source: **Andrea Wiggins**, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009

Today’s Research Artifacts

- Large scale datasets, scripts, software, workflows, papers, images, video, audio, annotations, ephemera, web sites...
  - “Research objects” - bundling all the pieces together
  - Hybrids of boundary objects and touchstones
- Technologies -> scientific revolution!
  - Open science

http://www.flickr.com/photos/smiteme/2379630899/

Source: **Andrea Wiggins**, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009

The Selfish (or Self-interested) Scientist

"A biologist would rather share their toothbrush than their gene name"

Mike Ashburner and others
Professor Genetics,
University of Cambridge,\UK

"Data mining: my data's mine and your data's mine"

The potential for collaboration

- What:
  - processes: “materials and methods” → workflows
  - data: unlikely, and certainly not until published
  - metadata (annotations, provenance traces...): ??
- When:
  - for contributors: part of publication process
    - some publishers demand public data and repeatable experiments
  - for consumers: reuse as part of experiment design
- Where and how:
  - a meeting point for a virtual community
  - Web 2.0 style of interaction
  - voluntary, incentive-based contributions

Asymmetric vs symmetric sharing

Traditional sharing is asymmetric:

- Producer-consumer:
  - from service providers to workflow designers
  - Biocatalogue

Open science is symmetric:

- Peer-based
  - sharing of workflows as complex processes
  - myExperiment

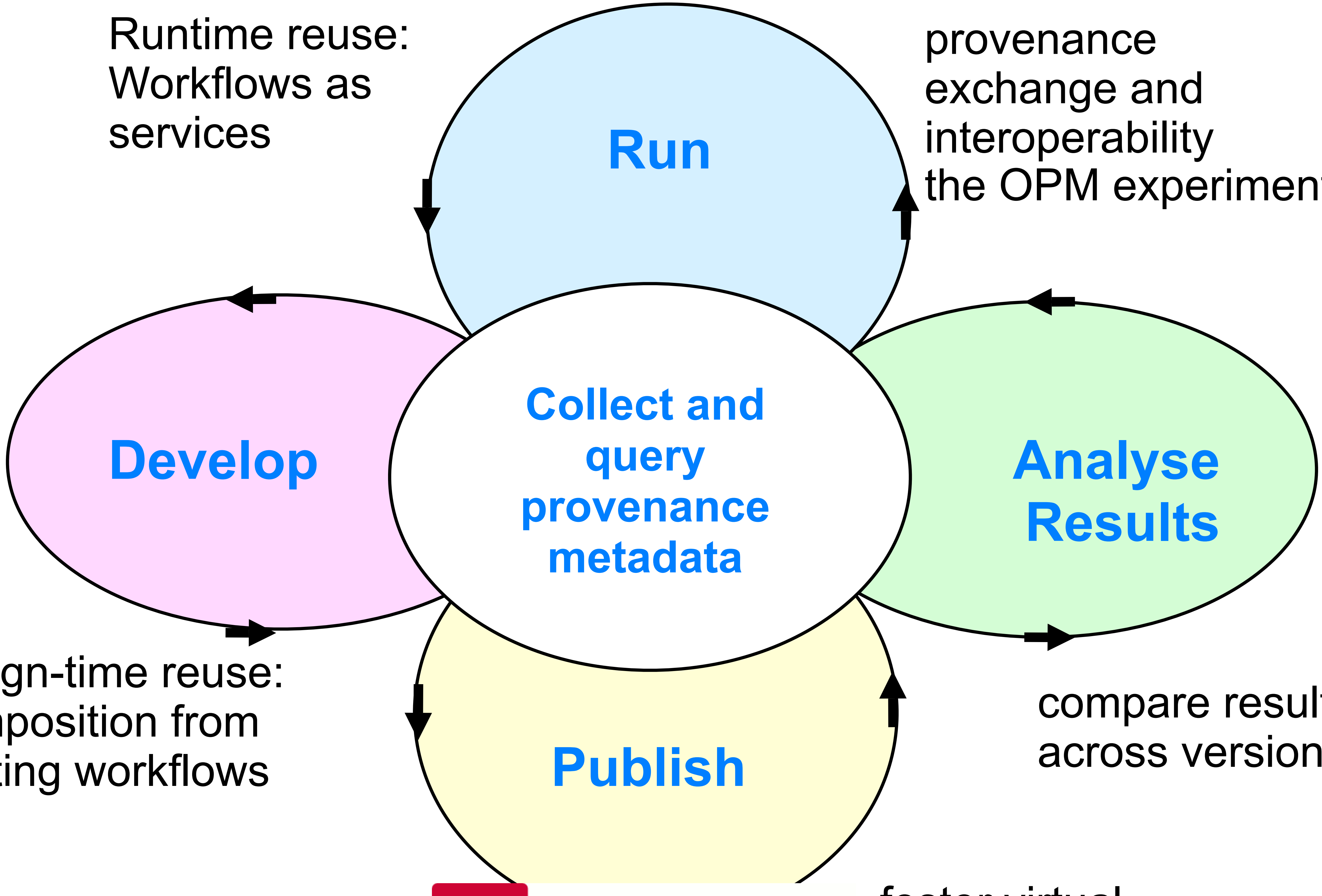
myGrid combines both paradigms:

- Service space “closed under composition”:
  - workflows are compositions of services
  - ... and they are services themselves
- Scientists become providers
  - of conceptual process models
  - and of executable services, as well!

Publishing for collaboration

Runtime reuse: Workflows as services

provenance exchange and interoperability the OPM experiment



Design-time reuse: Composition from existing workflows

compare results across versions

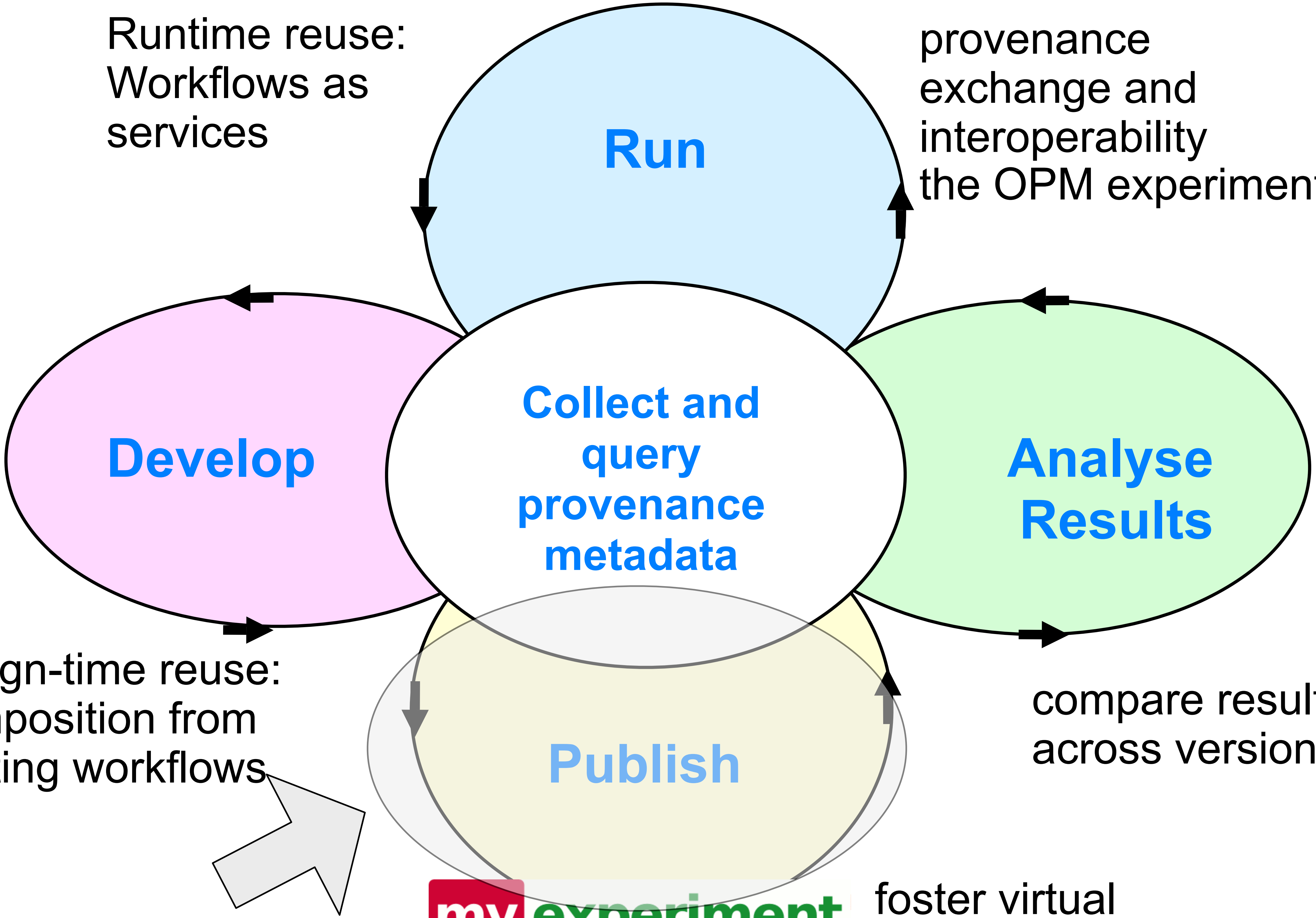
myexperiment

foster virtual scientific communities

Publishing for collaboration

Runtime reuse:
Workflows as
services

provenance
exchange and
interoperability
the OPM experiment



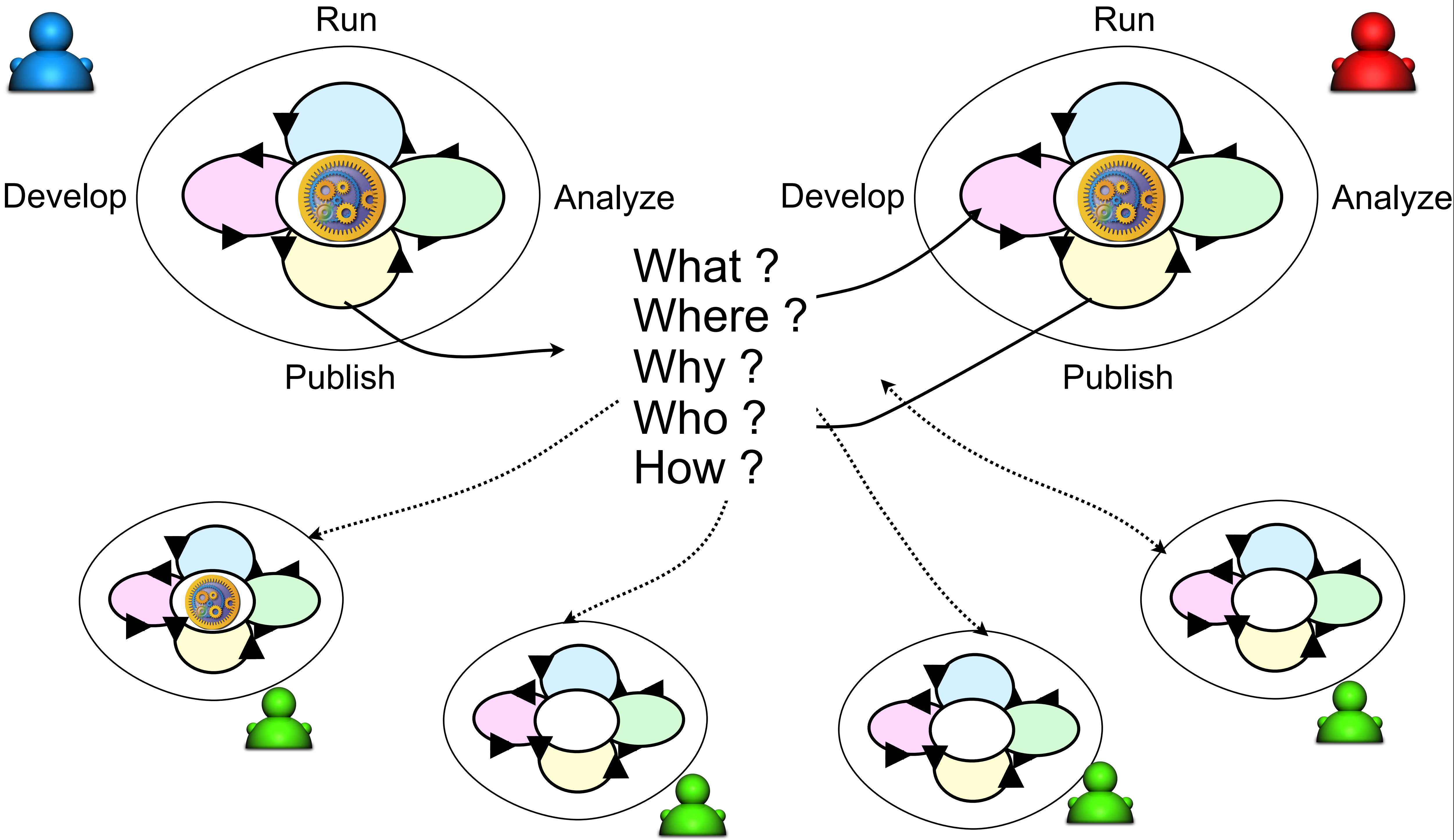
Design-time reuse:
Composition from
existing workflows

compare results
across versions

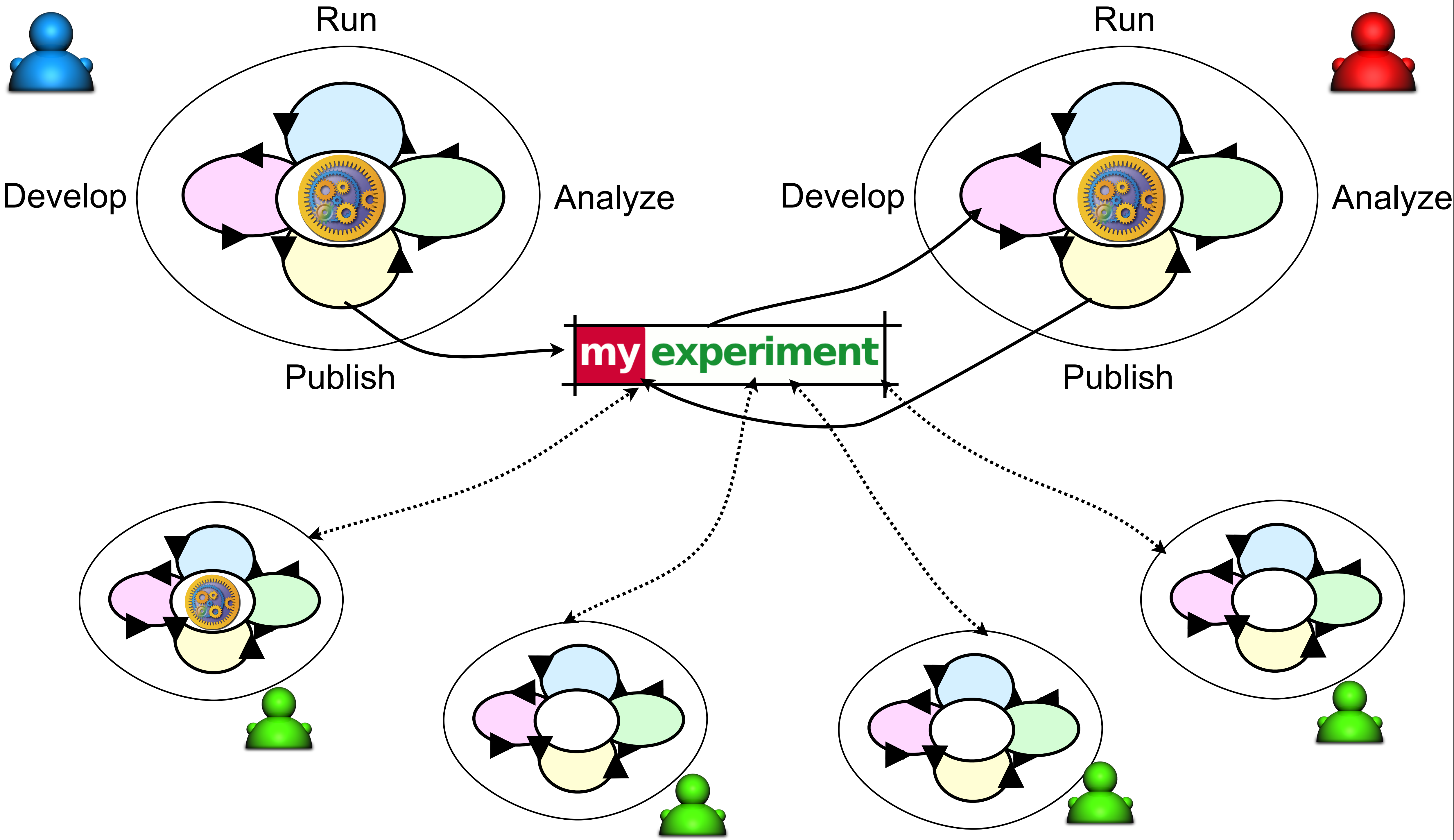
myexperiment

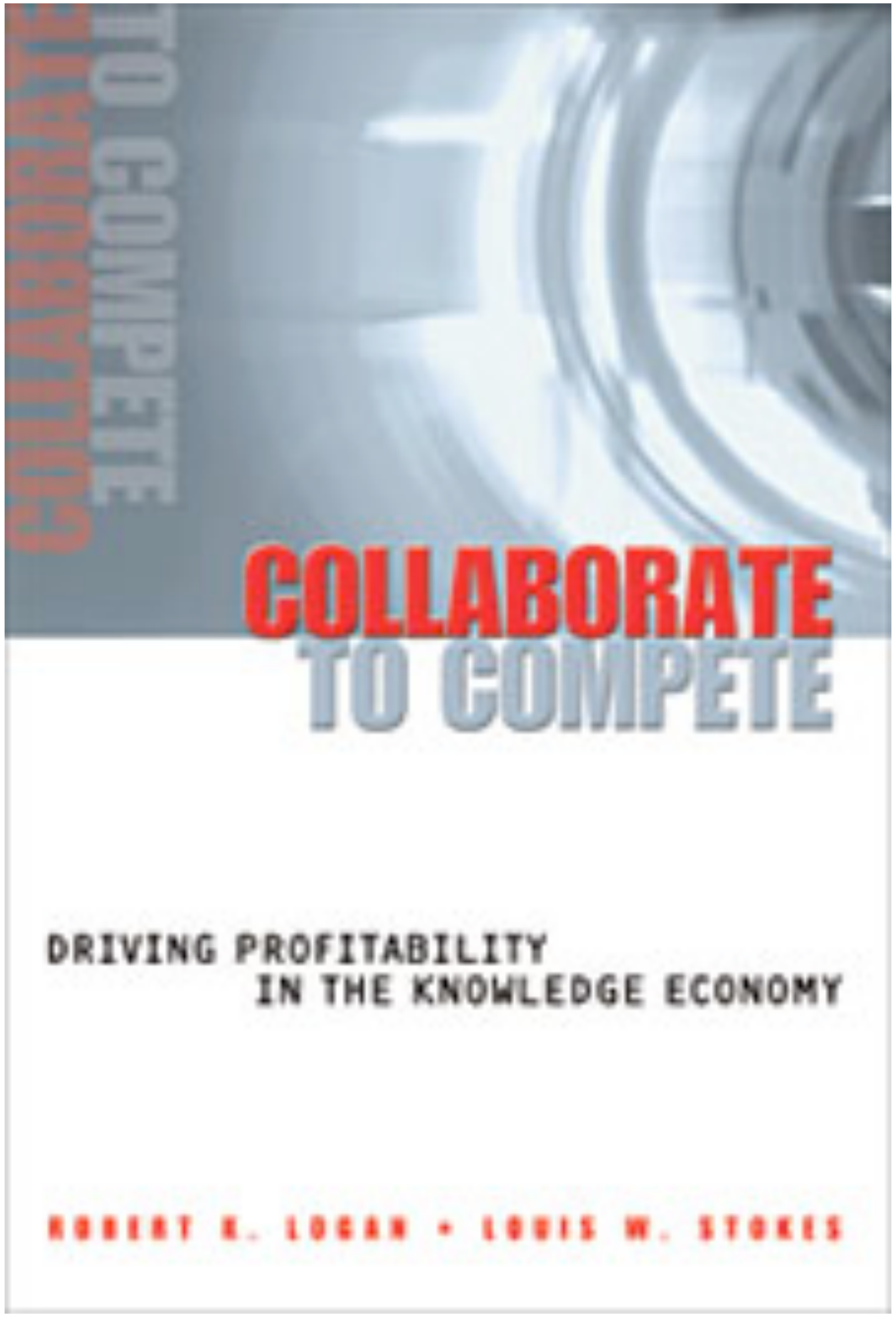
foster virtual
scientific
communities

Collaboration in the workflow space



Collaboration in the workflow space




COLLABORATE TO COMPETE
COLLABORATE
TO COMPETE
DRIVING PROFITABILITY IN THE KNOWLEDGE ECONOMY
ROBERT K. LOGAN • LOUIS W. STOKES

Rewards

Competitive advantage.

Academic vanity.

Adoption.

Reputation.

Fears

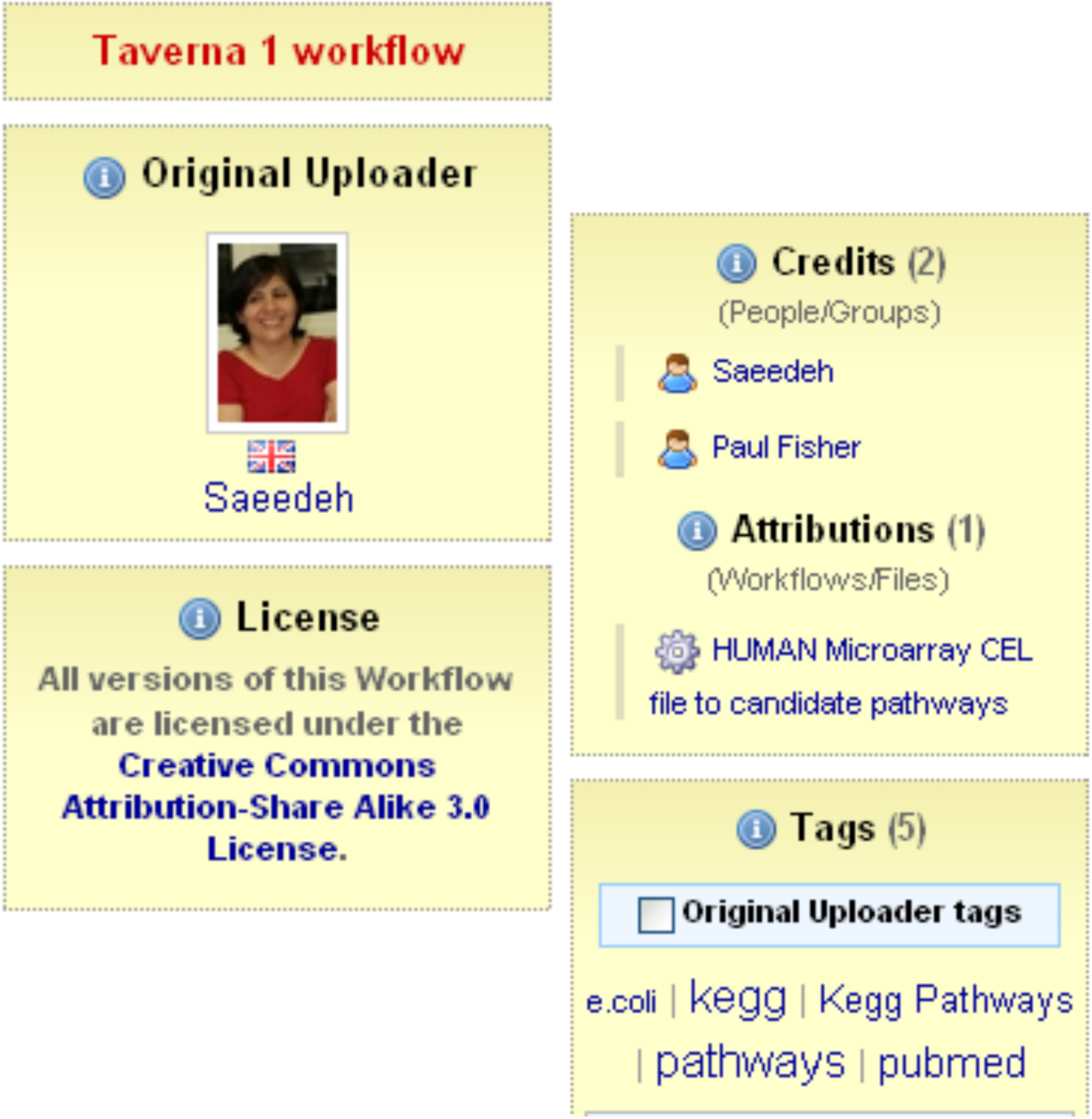
Scrutiny.

Being scooped.

Misinterpretation.

Reputation.

my experiment


Taverna 1 workflow

Original Uploader

Saeedeh

License

All versions of this Workflow are licensed under the Creative Commons Attribution-Share Alike 3.0 License.

Credits (2)
(People/Groups)

- Saeedeh
- Paul Fisher

Attributions (1)
(Workflows/Files)

- HUMAN Microarray CEL file to candidate pathways

Tags (5)

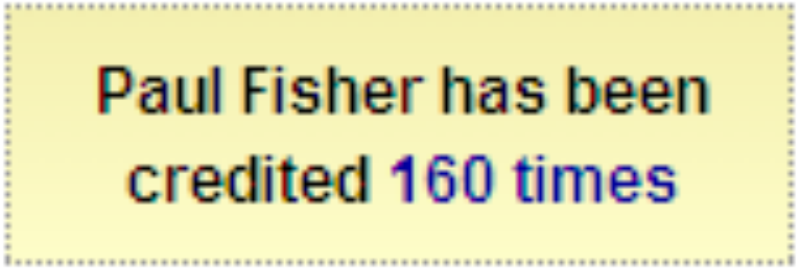
□ Original Uploader tags

e.coli | kegg | Kegg Pathways | pathways | pubmed

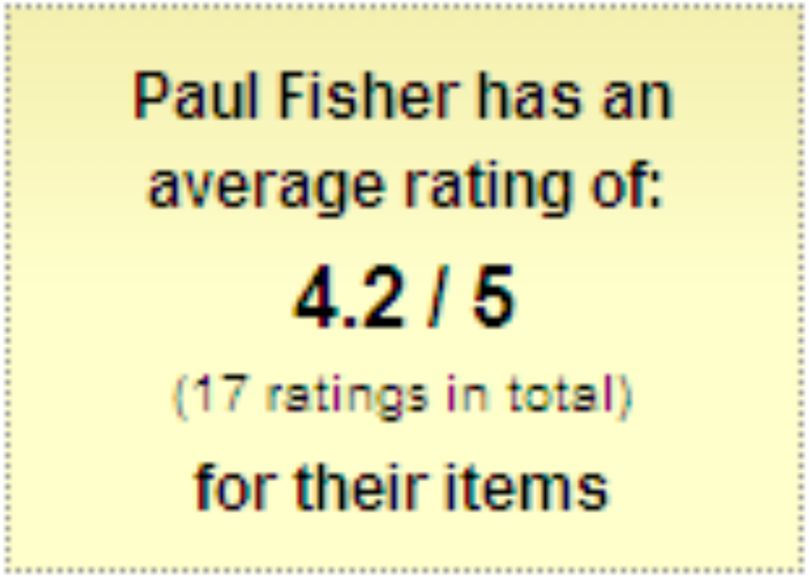
- Getting author to take credit!
- Creating a culture of attribution.
- Attribution and credit chains.
- Licensing and rights protection

Incentive and reputation

- Strong sense of persistent identity.
- Building reputation and boasting opportunities.
- Cult of the individual.
- High visibility to the participant and the community.
- Downloads & Views.
- Instrumentation and automated analysis.
- Feedback.
- Liability policy.



Paul Fisher has been credited 160 times

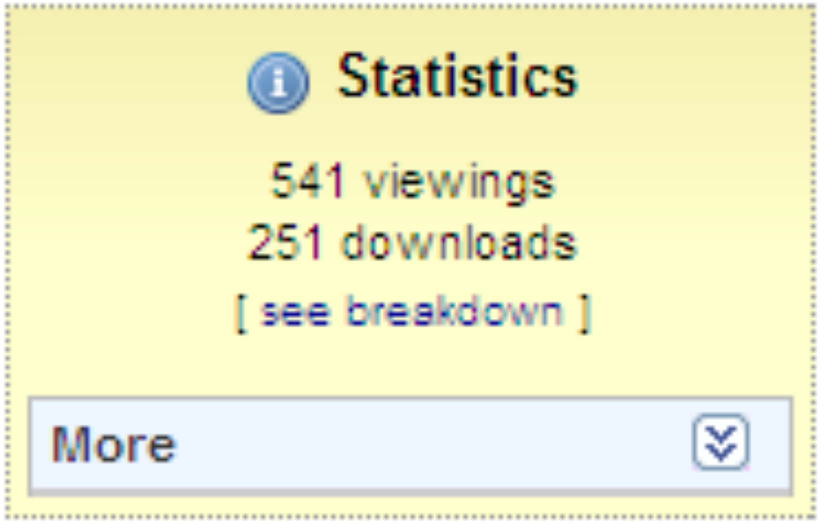


Paul Fisher has an average rating of:

4.2 / 5

(17 ratings in total)

for their items



Statistics

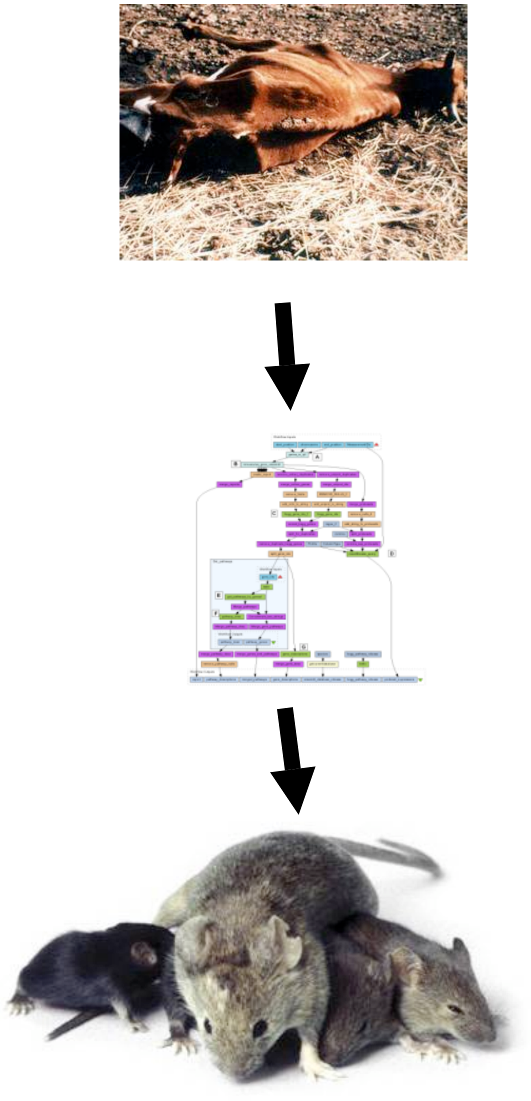
541 viewings

251 downloads

[see breakdown]

Reuse, Recycling, Repurposing Cross-fertilization

- Paul writes workflows for identifying biological pathways implicated in resistance to Trypanosomiasis in cattle
- Paul meets Jo. Jo is investigating Whipworm in mouse.
- Jo reuses one of Paul's workflow without change.
- Jo identifies the biological pathways involved in sex dependence in the mouse model, believed to be involved in the ability of mice to expel the parasite.
- Previously a manual two year study by Jo had failed to do this.

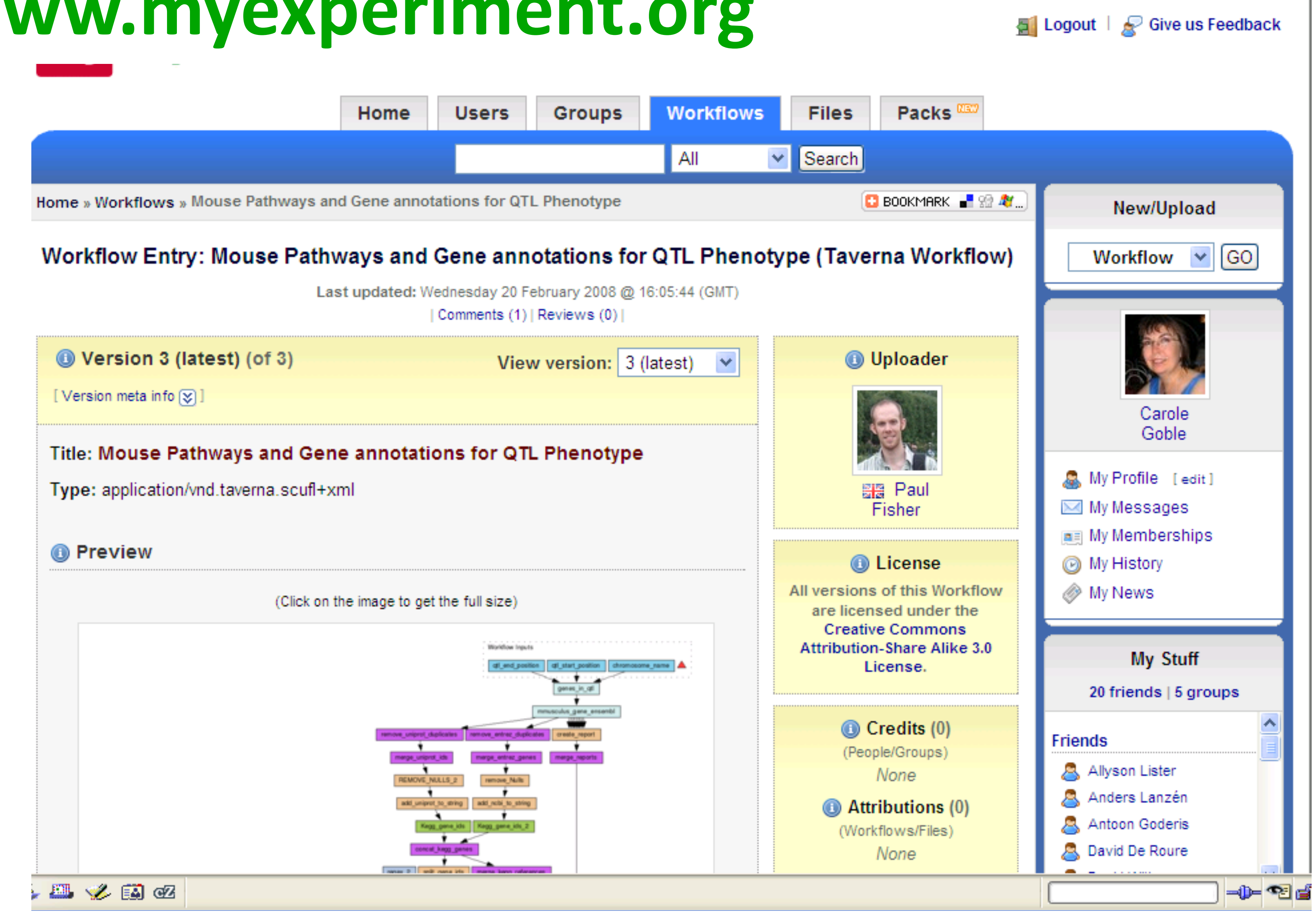


my experiment

www.myexperiment.org

Socially share, discover and reuse workflows and other methods.

Cooperative bazaar.



Logout | Give us Feedback

Home | Users | Groups | Workflows | Files | Packs NEW

All | Search

Home » Workflows » Mouse Pathways and Gene annotations for QTL Phenotype

BOOKMARK

Workflow Entry: Mouse Pathways and Gene annotations for QTL Phenotype (Taverna Workflow)

Last updated: Wednesday 20 February 2008 @ 16:05:44 (GMT)

| Comments (1) | Reviews (0) |

Version 3 (latest) (of 3)

View version: 3 (latest)

[Version meta info]

Title: Mouse Pathways and Gene annotations for QTL Phenotype

Type: application/vnd.taverna.scufl+xml

Preview

(Click on the image to get the full size)

Uploader

Paul Fisher

License

All versions of this Workflow are licensed under the Creative Commons Attribution-Share Alike 3.0 License.

Credits (0)

(People/Groups)

None

Attributions (0)

(Workflows/Files)

None

New/Upload

Workflow | GO

Carole Goble

My Profile [edit]

My Messages

My Memberships

My History

My News

My Stuff

20 friends | 5 groups

Friends

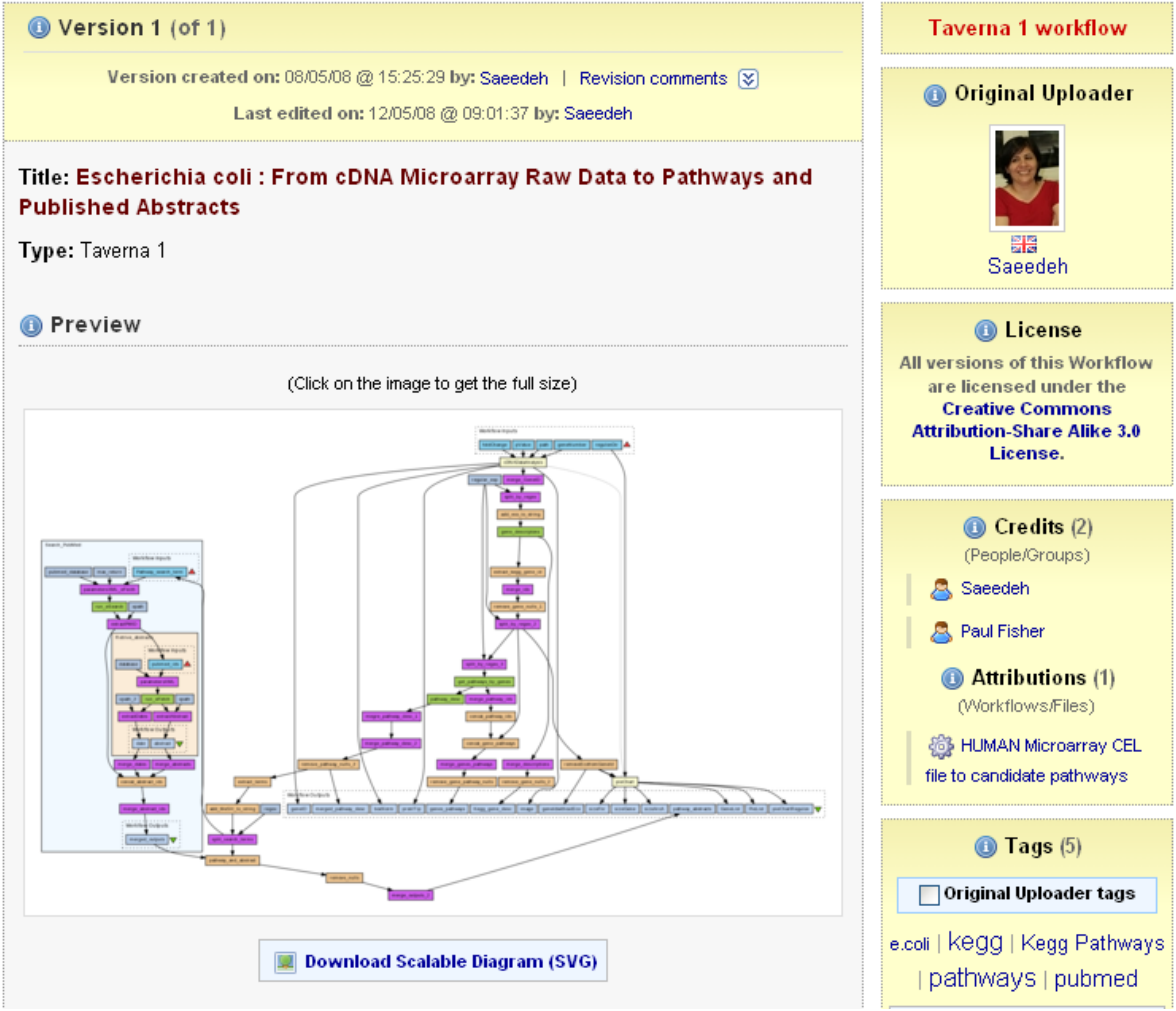
Allyson Lister

Anders Lanzén

Antoon Goderis

David De Roure

Sunday 10th May:
1748 registered users, 143 groups, 669 workflows, 197 files, 52 packs
56 different countries. Top 4: UK, US, The Netherlands, Germany



Version 1 (of 1)

Version created on: 08/05/08 @ 15:25:29 by: Saeedeh | Revision comments

Last edited on: 12/05/08 @ 09:01:37 by: Saeedeh

Title: Escherichia coli : From cDNA Microarray Raw Data to Pathways and Published Abstracts

Type: Taverna 1

Preview

(Click on the image to get the full size)

Taverna 1 workflow

Original Uploader

Saeedeh

License

All versions of this Workflow are licensed under the Creative Commons Attribution-Share Alike 3.0 License.

Credits (2)
(People/Groups)

- Saeedeh
- Paul Fisher

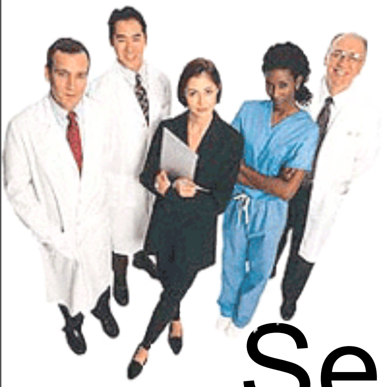
Attributions (1)
(Workflows/Files)

- HUMAN Microarray CEL file to candidate pathways

Tags (5)

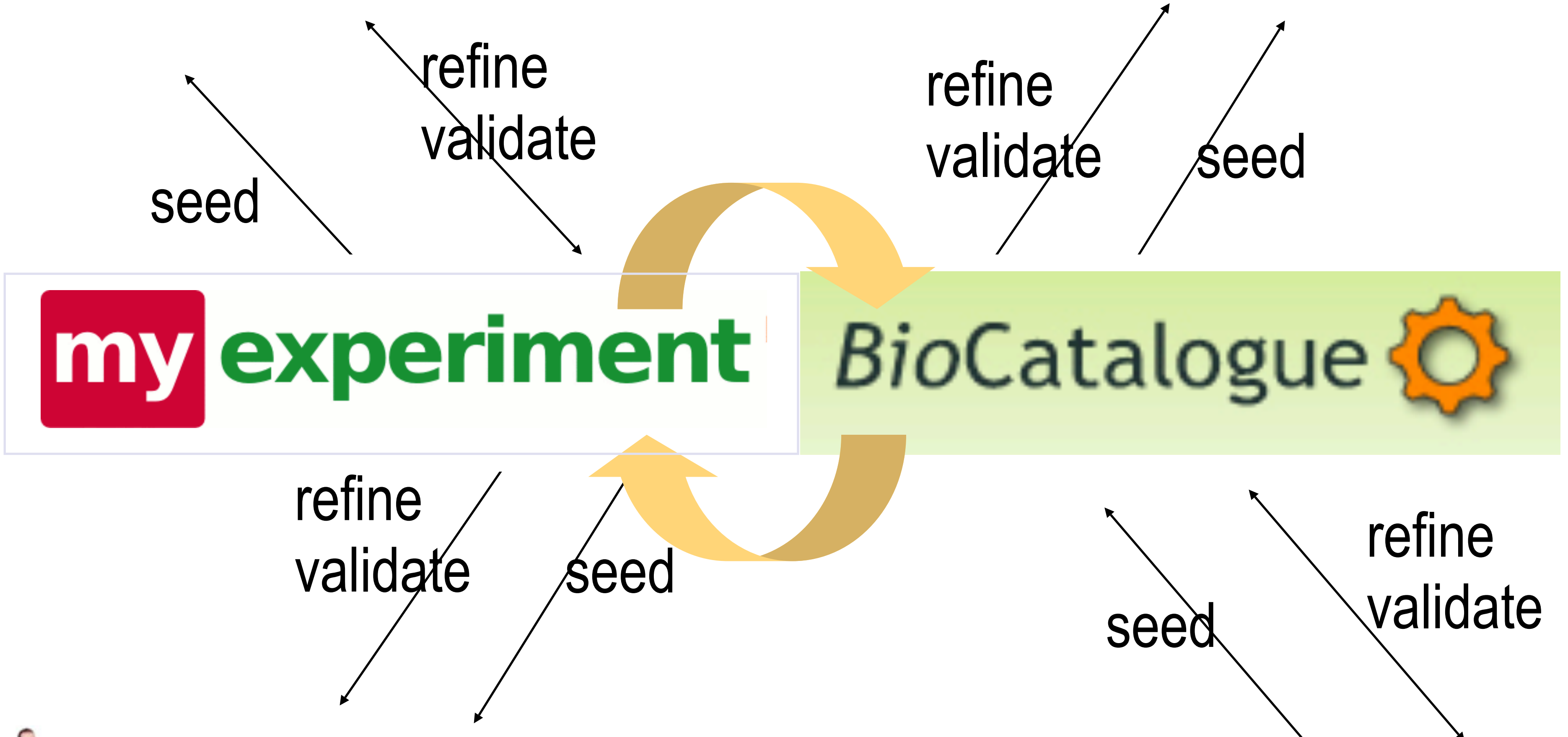
□ Original Uploader tags

e.coli | kegg | Kegg Pathways | pathways | pubmed



Self-Curation
by Contributors

Curation by
Experts

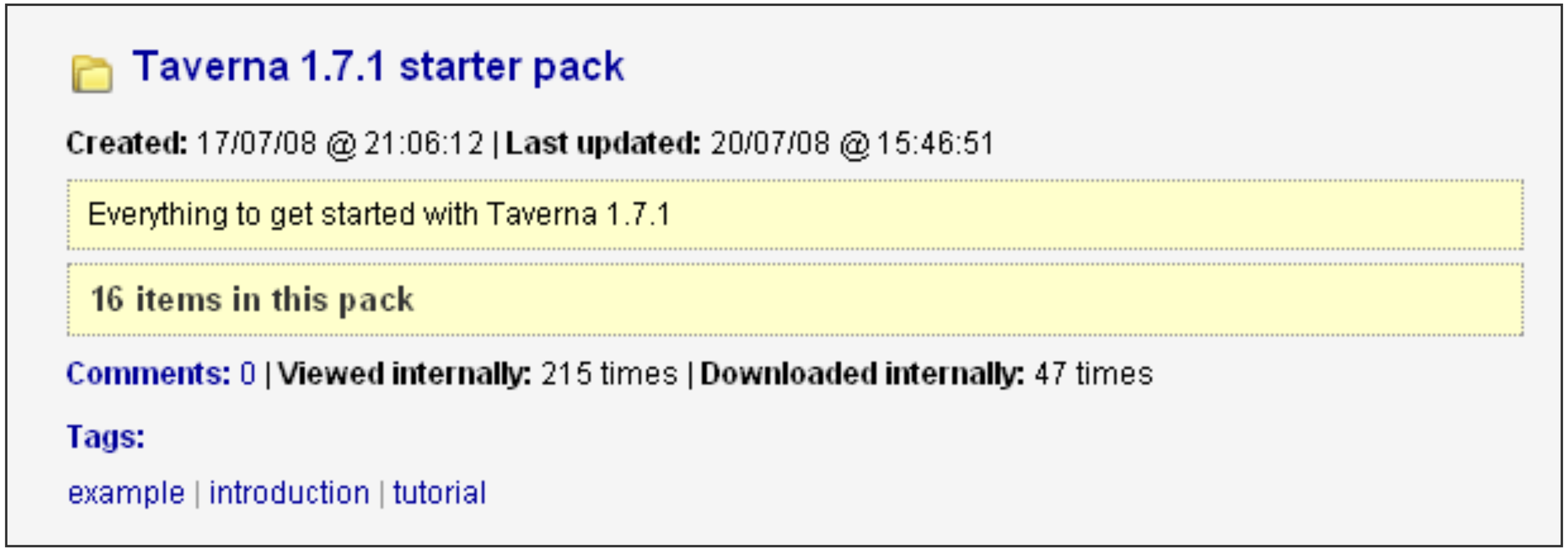


Social Curation
by the Crowd

Automated
Curation

Taverna II - Workflow 2.0
Taverna II - Workflow 2.0
Taverna II - Workflow 2.0

Packs



Taverna 1.7.1 starter pack

Created: 17/07/08 @ 21:06:12 | **Last updated:** 20/07/08 @ 15:46:51

Everything to get started with Taverna 1.7.1

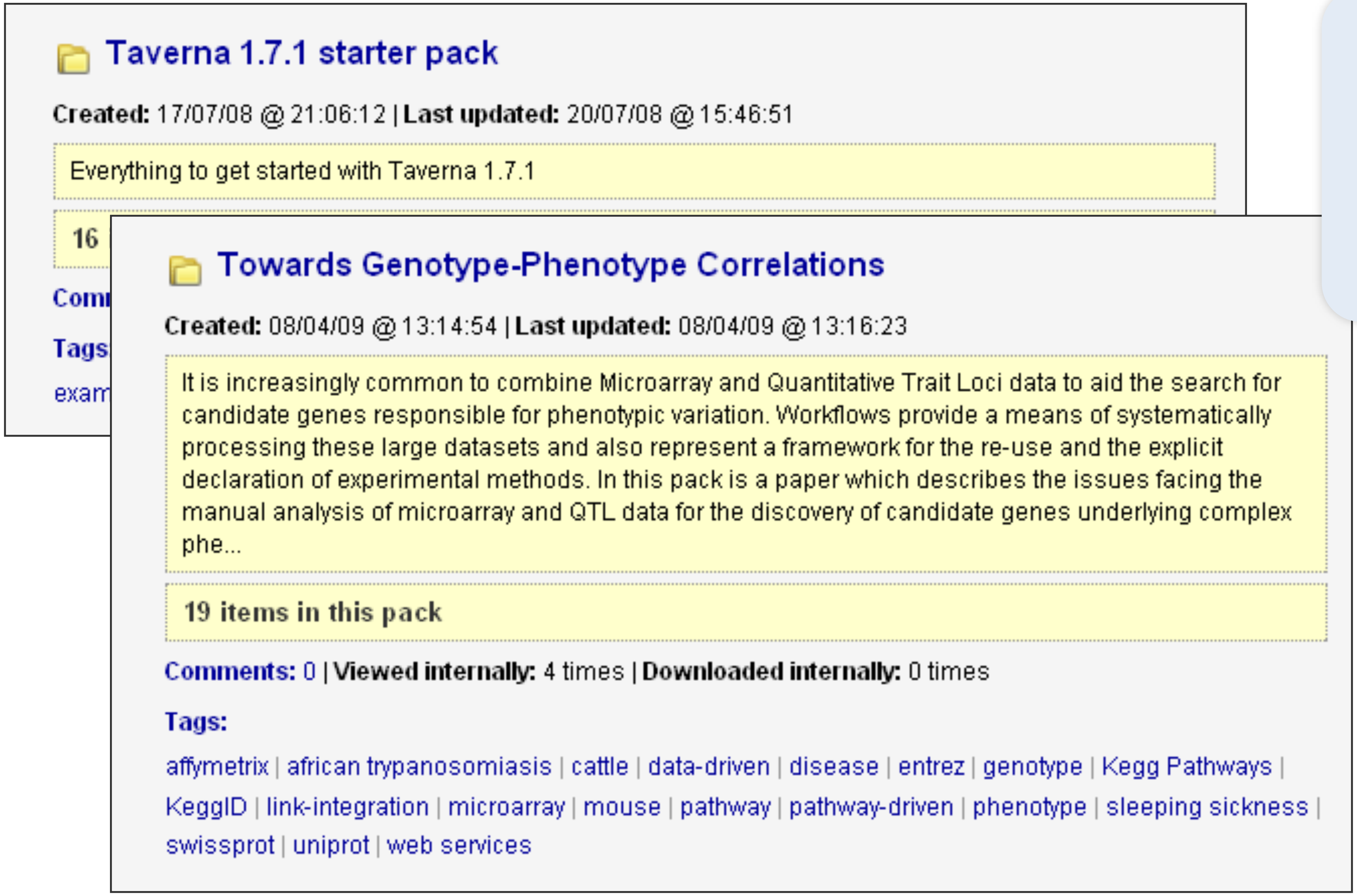
16 items in this pack

Comments: 0 | **Viewed internally:** 215 times | **Downloaded internally:** 47 times

Tags:

example | introduction | tutorial

Packs



Taverna 1.7.1 starter pack

Created: 17/07/08 @ 21:06:12 | **Last updated:** 20/07/08 @ 15:46:51

Everything to get started with Taverna 1.7.1

16

Com

Tags

exam

Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | **Last updated:** 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically processing these large datasets and also represent a framework for the re-use and the explicit declaration of experimental methods. In this pack is a paper which describes the issues facing the manual analysis of microarray and QTL data for the discovery of candidate genes underlying complex phe...

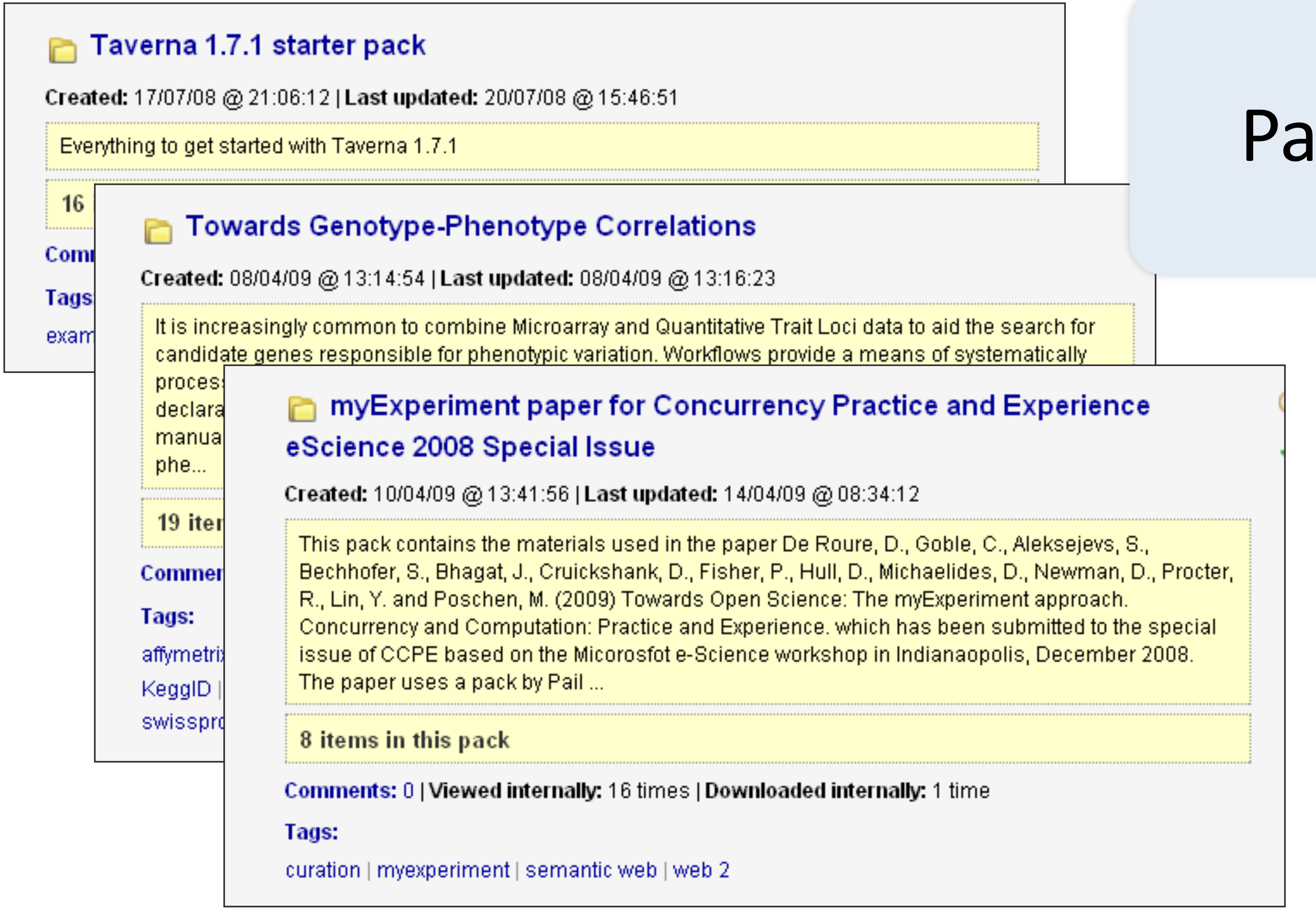
19 items in this pack

Comments: 0 | **Viewed internally:** 4 times | **Downloaded internally:** 0 times

Tags:

affymetrix | african trypanosomiasis | cattle | data-driven | disease | entrez | genotype | Kegg Pathways | KeggID | link-integration | microarray | mouse | pathway | pathway-driven | phenotype | sleeping sickness | swissprot | uniprot | web services

Packs



Taverna 1.7.1 starter pack

Created: 17/07/08 @ 21:06:12 | **Last updated:** 20/07/08 @ 15:46:51

Everything to get started with Taverna 1.7.1

16

Com

Tags

exam

Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | **Last updated:** 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically proces
declara
manua
phe...

19 iter

Commer

Tags:

affymetri

KeggID |

swisspro

myExperiment paper for Concurrency Practice and Experience eScience 2008 Special Issue

Created: 10/04/09 @ 13:41:56 | **Last updated:** 14/04/09 @ 08:34:12

This pack contains the materials used in the paper De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y. and Poschen, M. (2009) Towards Open Science: The myExperiment approach. Concurrency and Computation: Practice and Experience. which has been submitted to the special issue of CCPE based on the Micorosfot e-Science workshop in Indianaopolis, December 2008. The paper uses a pack by Pail ...

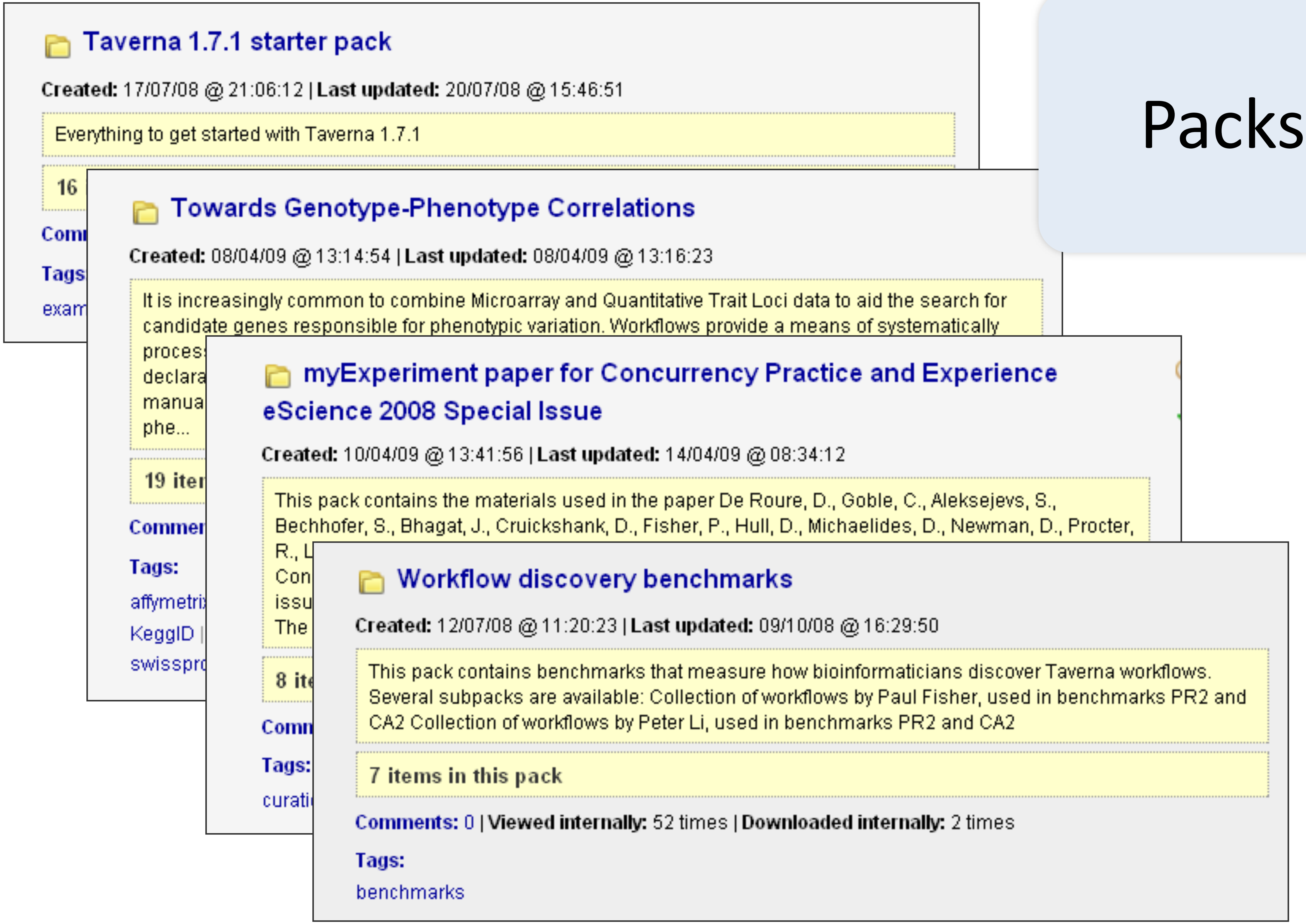
8 items in this pack

Comments: 0 | **Viewed internally:** 16 times | **Downloaded internally:** 1 time

Tags:

curation | myexperiment | semantic web | web 2

Packs



Taverna 1.7.1 starter pack

Created: 17/07/08 @ 21:06:12 | **Last updated:** 20/07/08 @ 15:46:51

Everything to get started with Taverna 1.7.1

16

Com

Tags

exam

Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | **Last updated:** 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically proces
declara
manua
phe...

19 iter

Commer

Tags:

affymetri
KeggID |
swisspro

myExperiment paper for Concurrency Practice and Experience eScience 2008 Special Issue

Created: 10/04/09 @ 13:41:56 | **Last updated:** 14/04/09 @ 08:34:12

This pack contains the materials used in the paper De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., L
Con
issu
The

8 ite

Comm

Tags:

curati

Workflow discovery benchmarks

Created: 12/07/08 @ 11:20:23 | **Last updated:** 09/10/08 @ 16:29:50

This pack contains benchmarks that measure how bioinformaticians discover Taverna workflows. Several subpacks are available: Collection of workflows by Paul Fisher, used in benchmarks PR2 and CA2 Collection of workflows by Peter Li, used in benchmarks PR2 and CA2

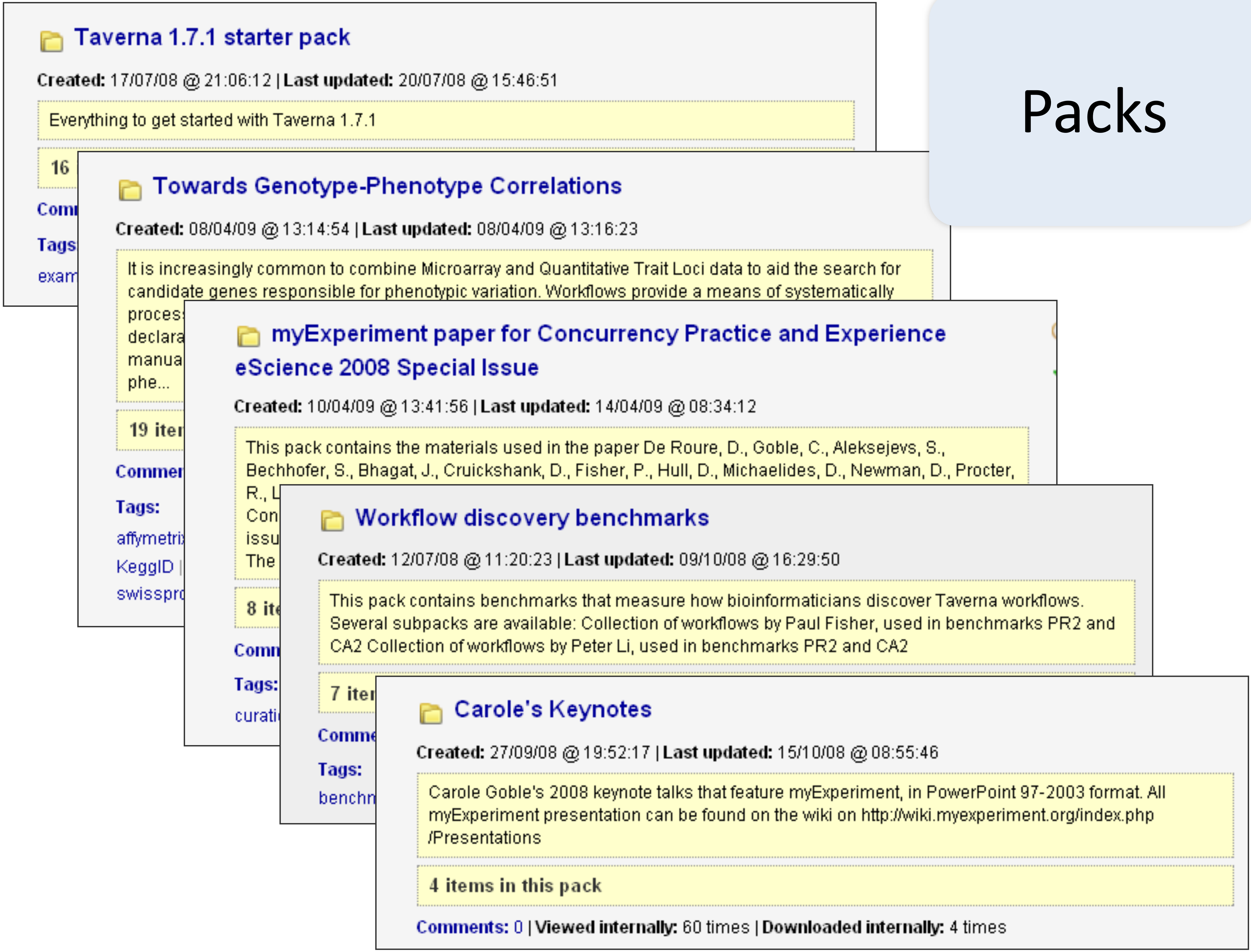
7 items in this pack

Comments: 0 | **Viewed internally:** 52 times | **Downloaded internally:** 2 times

Tags:

benchmarks

Packs



Taverna 1.7.1 starter pack

Created: 17/07/08 @ 21:06:12 | **Last updated:** 20/07/08 @ 15:46:51

Everything to get started with Taverna 1.7.1

16

Comm

Tags

exam

Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | **Last updated:** 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically proces
declara
manua
phe...

19 iter

Commen

Tags:

affymetrix

KeggID |

swisspro

myExperiment paper for Concurrency Practice and Experience eScience 2008 Special Issue

Created: 10/04/09 @ 13:41:56 | **Last updated:** 14/04/09 @ 08:34:12

This pack contains the materials used in the paper De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., L
Con
issu
The

8 ite

Comm

Tags:

curati

Workflow discovery benchmarks

Created: 12/07/08 @ 11:20:23 | **Last updated:** 09/10/08 @ 16:29:50

This pack contains benchmarks that measure how bioinformaticians discover Taverna workflows. Several subpacks are available: Collection of workflows by Paul Fisher, used in benchmarks PR2 and CA2 Collection of workflows by Peter Li, used in benchmarks PR2 and CA2

7 iter

Comme

Tags:

benchm

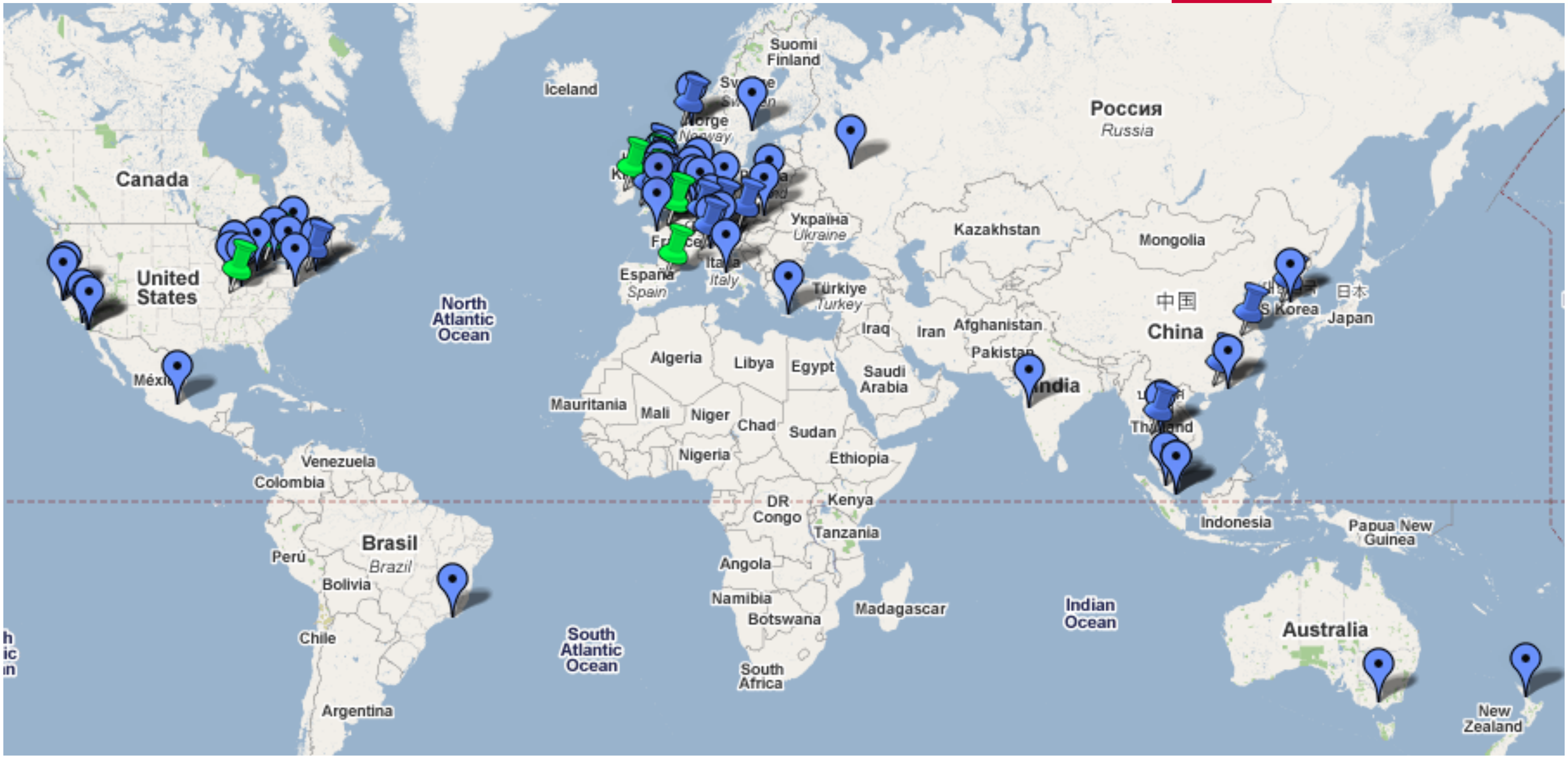
Carole's Keynotes

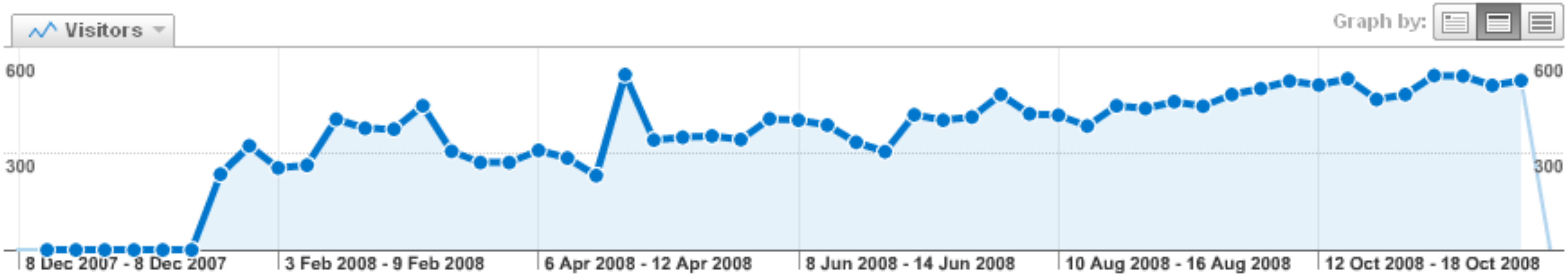
Created: 27/09/08 @ 19:52:17 | **Last updated:** 15/10/08 @ 08:55:46

Carole Goble's 2008 keynote talks that feature myExperiment, in PowerPoint 97-2003 format. All myExperiment presentation can be found on the wiki on http://wiki.myexperiment.org/index.php /Presentations

4 items in this pack

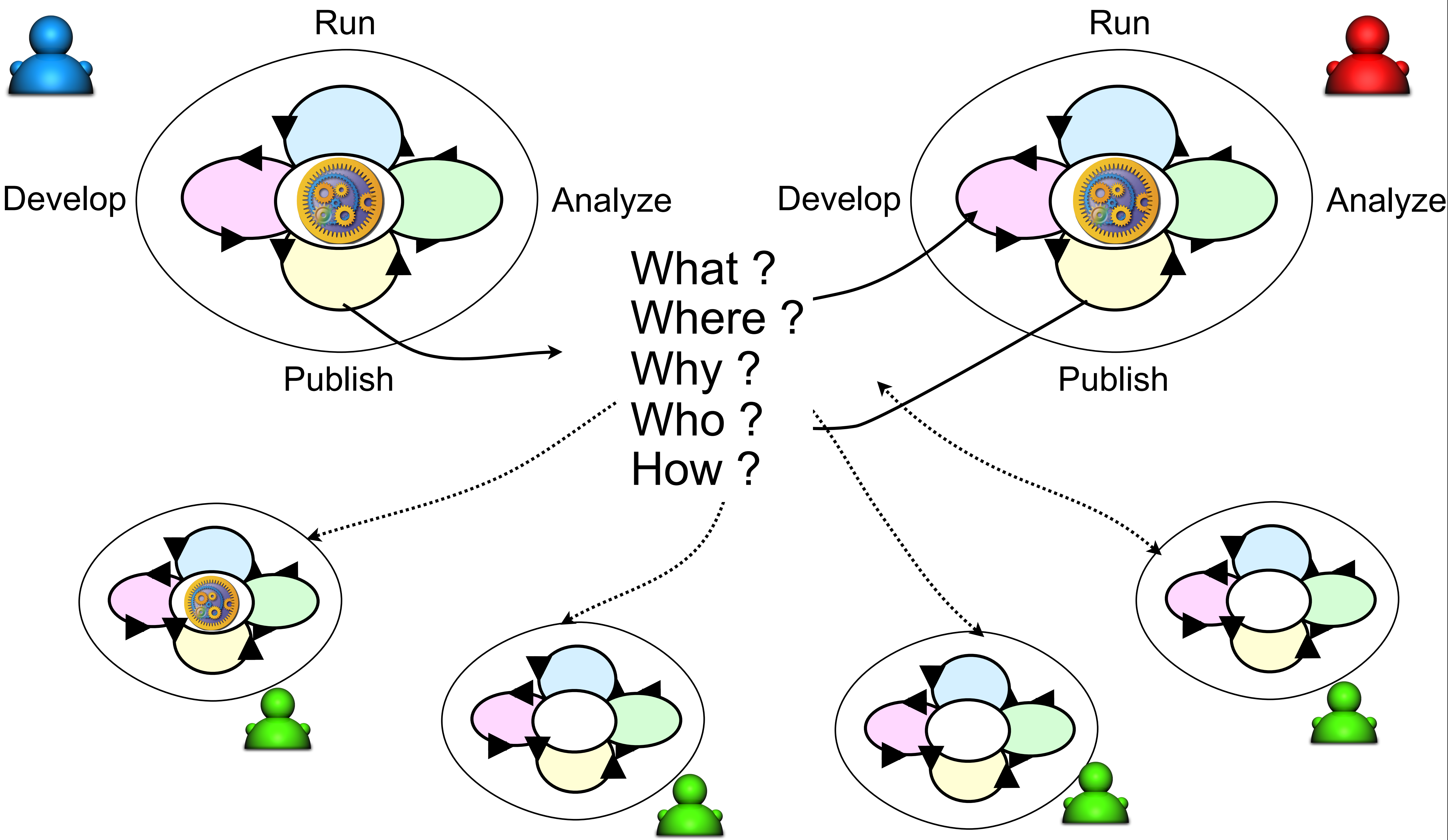
Comments: 0 | **Viewed internally:** 60 times | **Downloaded internally:** 4 times myexperiment





15,902 Absolute Unique Visitors

Collaboration in the workflow space



Run

Develop

Analyze

Publish

Run

Develop

Analyze

Publish

What ?
Where ?
Why ?
Who ?
How ?

Technical implications of open science

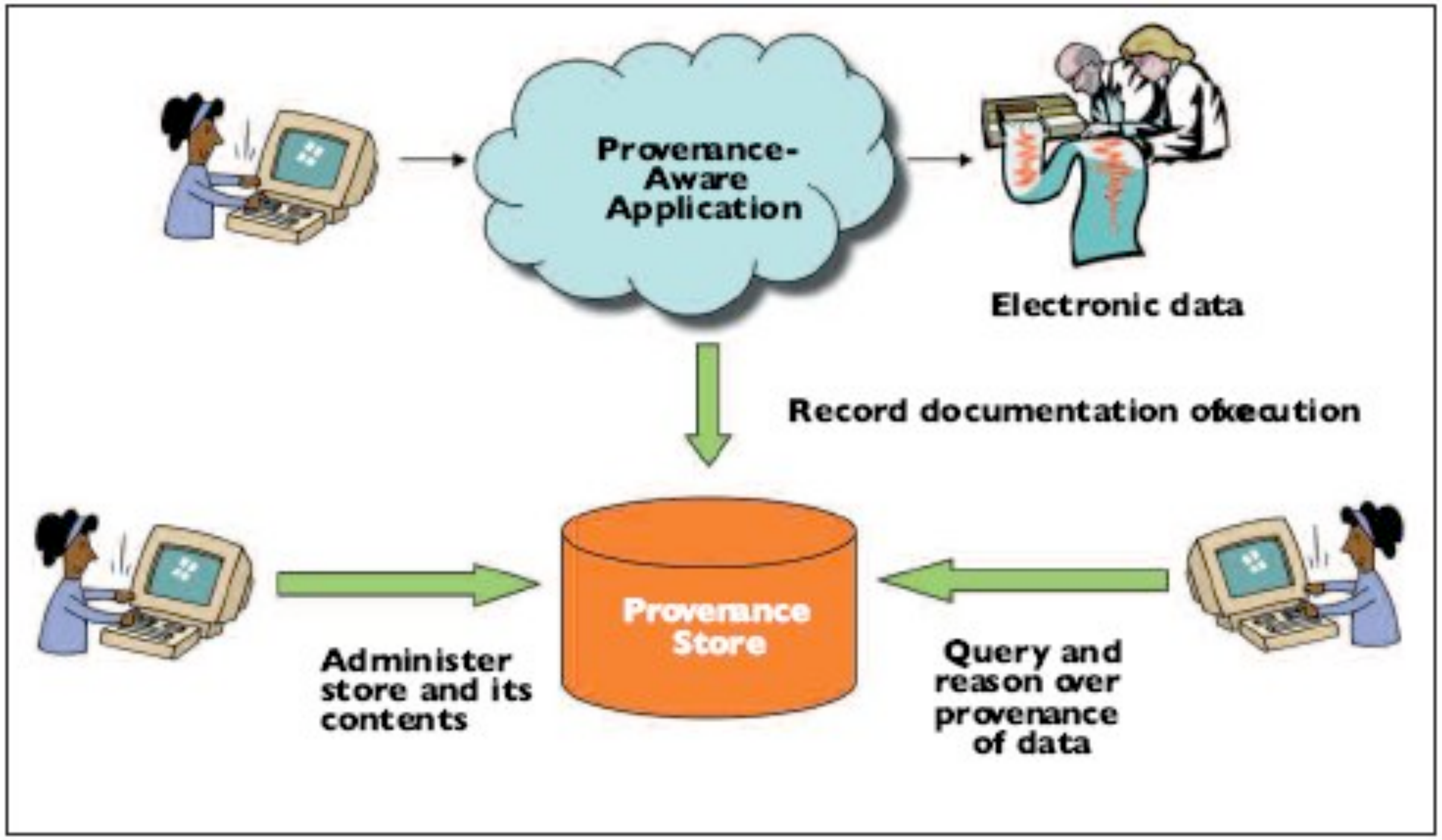
- Process interoperability
  - SOA principles: runtime interoperability
  - but, still no common workflow model after all!
- Data interoperability
  - Traditional heterogeneity / integration issues
  - Dataspaces
  - LinkedData
  - ...
- Aggregation: creating logical units
  - process + inputs + outputs + provenance traces + ...
  - Research Objects
- Provenance interoperability

Technical implications of open science

- Process interoperability
  - SOA principles: runtime interoperability
  - but, still no common workflow model after all!
- Data interoperability
  - Traditional heterogeneity / integration issues
  - Dataspaces
  - LinkedData
  - ...
- Aggregation: creating logical units
  - process + inputs + outputs + provenance traces + ...
  - Research Objects
- Provenance interoperability

Provenance of data

- “It would include details of the processes that produced electronic data as far back as the beginning of time or at least the epoch of provenance awareness.”

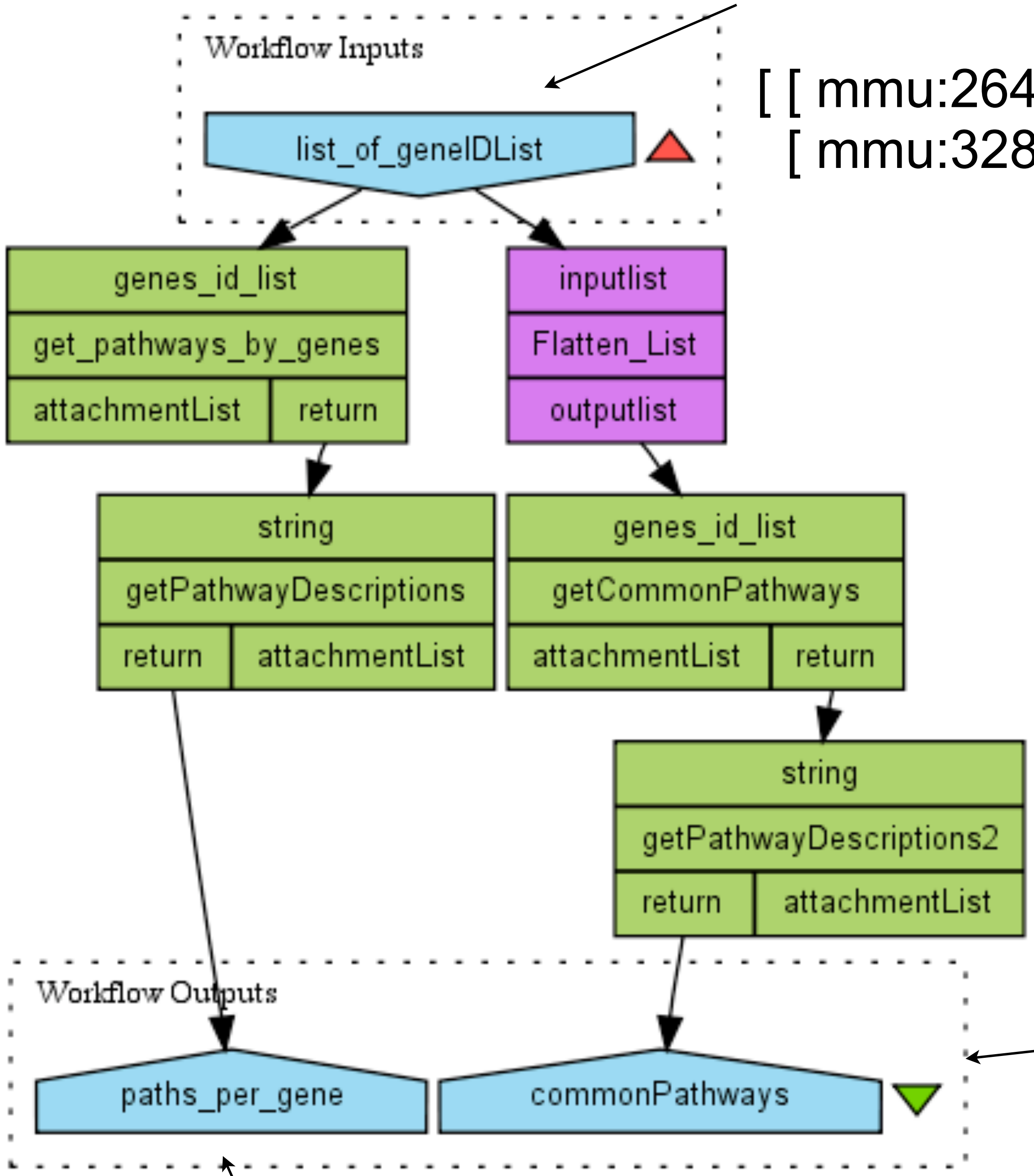

Provenance-Aware Application
Electronic data
Record documentation ofexecution
Provenance Store
Administer store and its contents
Query and reason over provenance of data

Luc Moreau, Paul Groth, Simon Miles, Javier Vazquez-Salceda, John Ibbotson, Sheng Jiang, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, Laszlo Varga, *The provenance of electronic data,* Communications of the ACM, Vol. 51 No. 4, Pages 52-58

Analysis of process results

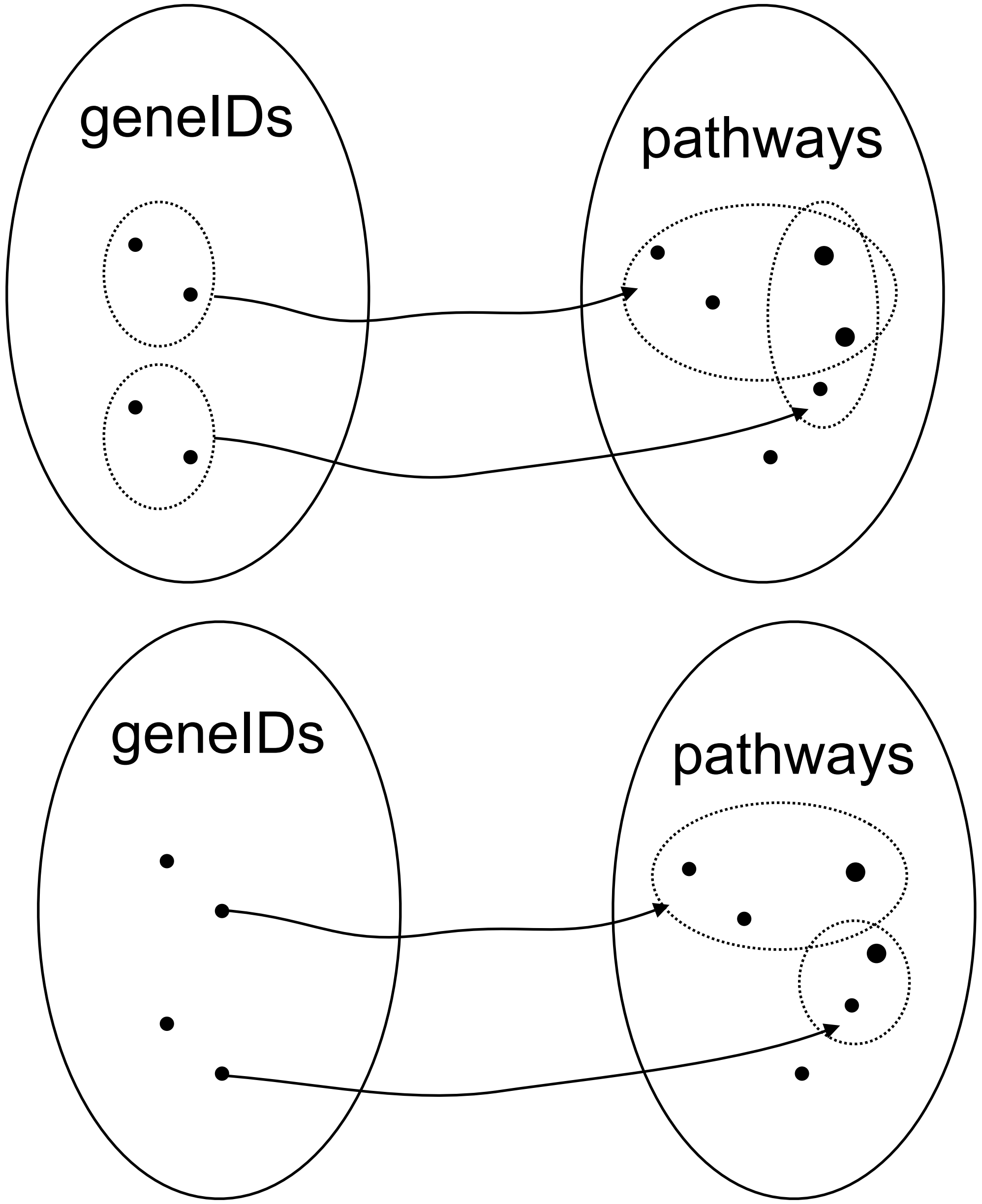
- Causal relations:
  - ☑ which pathway sets come from which gene sets?
  - which processes contributed to producing this image?
  - which process(es) caused this data to be incorrect?
  - which data caused this process to fail?
- Process and data analytics:
  - show me the variations in output in relation to an input parameter sweep (multiple process runs)
  - how often has my favourite service been executed?
    - on what inputs?
  - who produced this data?
  - how often does this pathway turn up when the input genes range over a certain set S?

Example: inverse data associations



List-structured KEGG gene ids:

[[mmu:26416],
  [mmu:328788]]



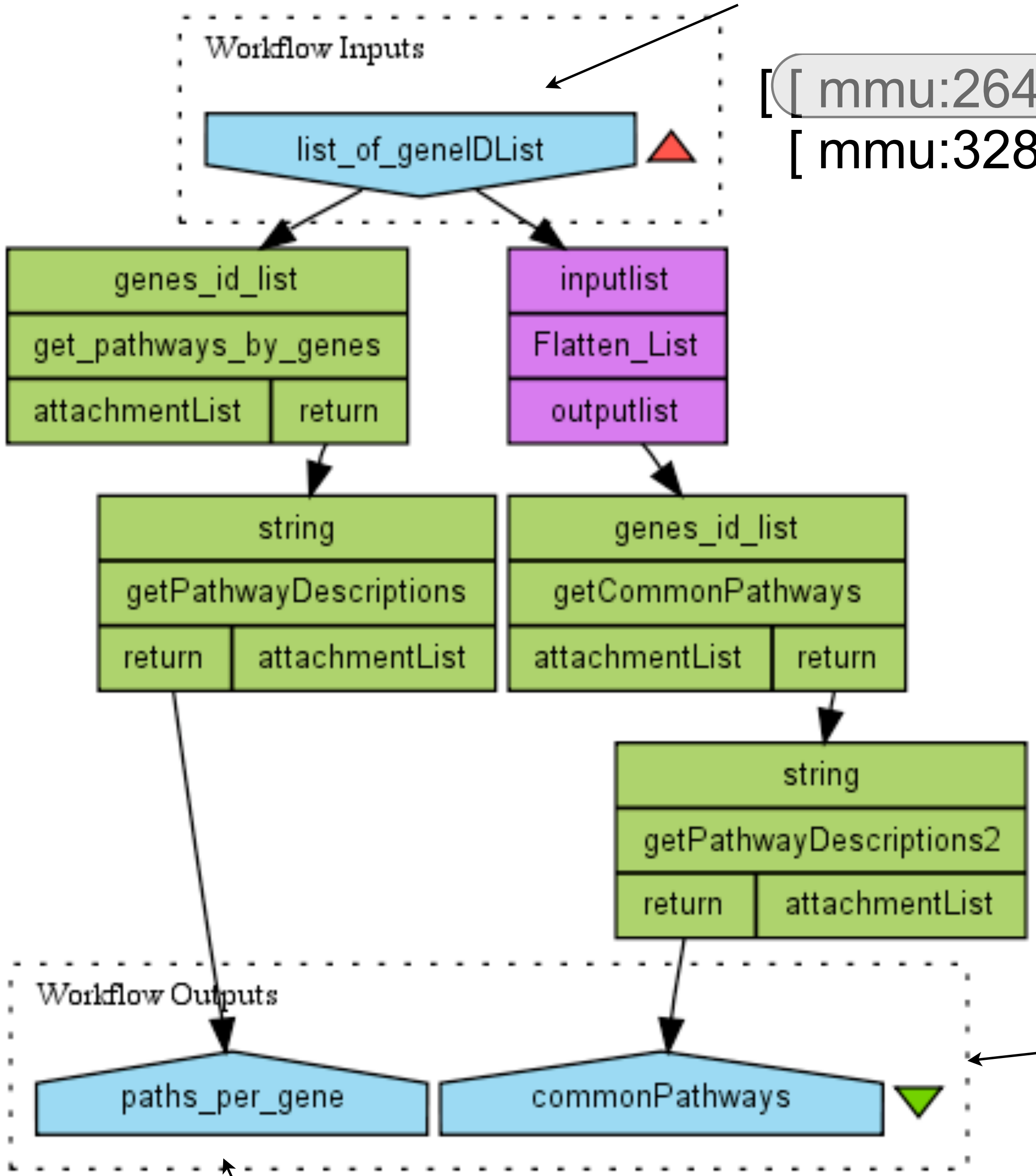
[path:mmu04010 MAPK signaling,
  path:mmu04370 VEGF signaling]

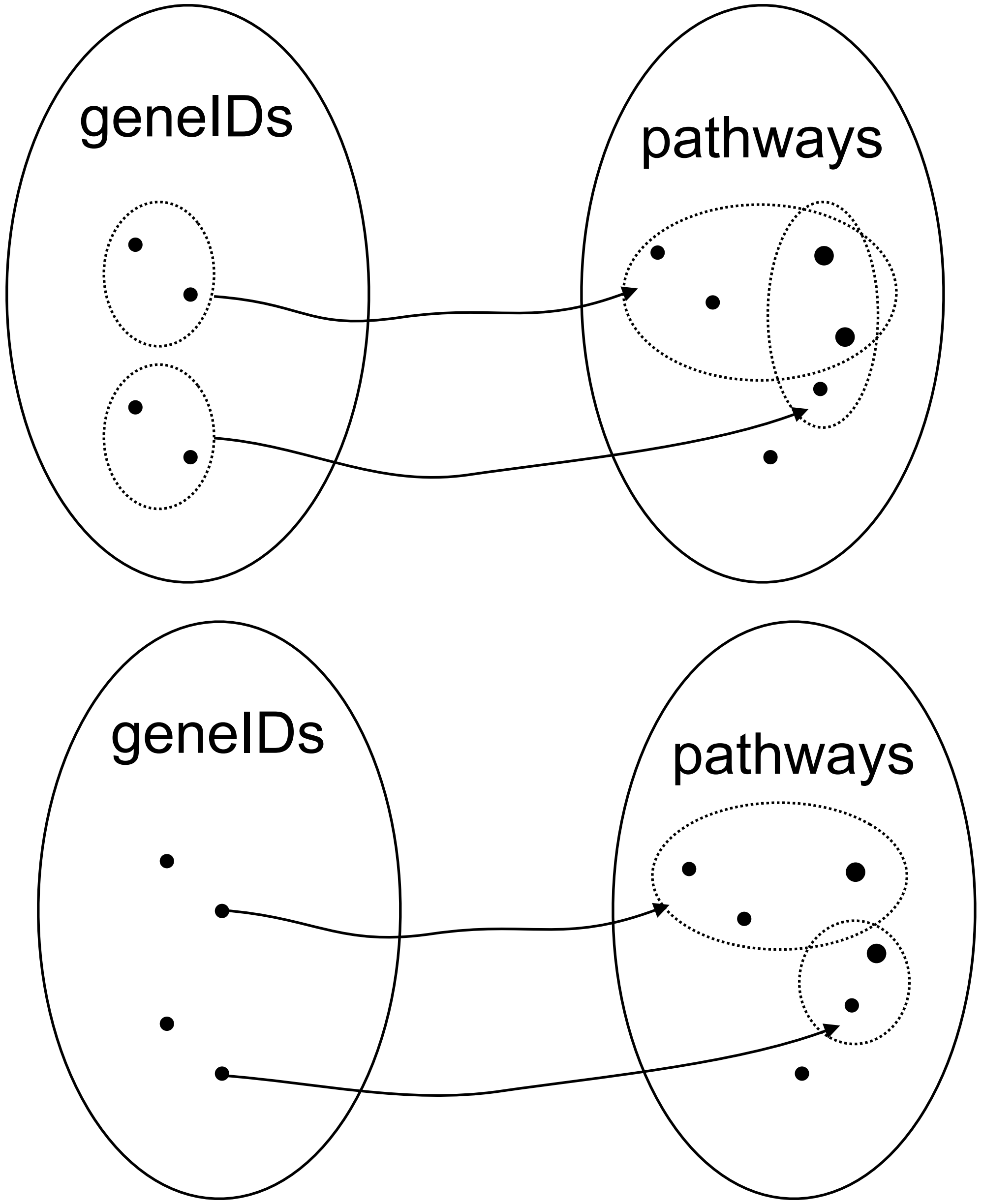
[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
  [path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Example: inverse data associations

List-structured KEGG gene ids:

[[mmu:26416],
  [mmu:328788]]

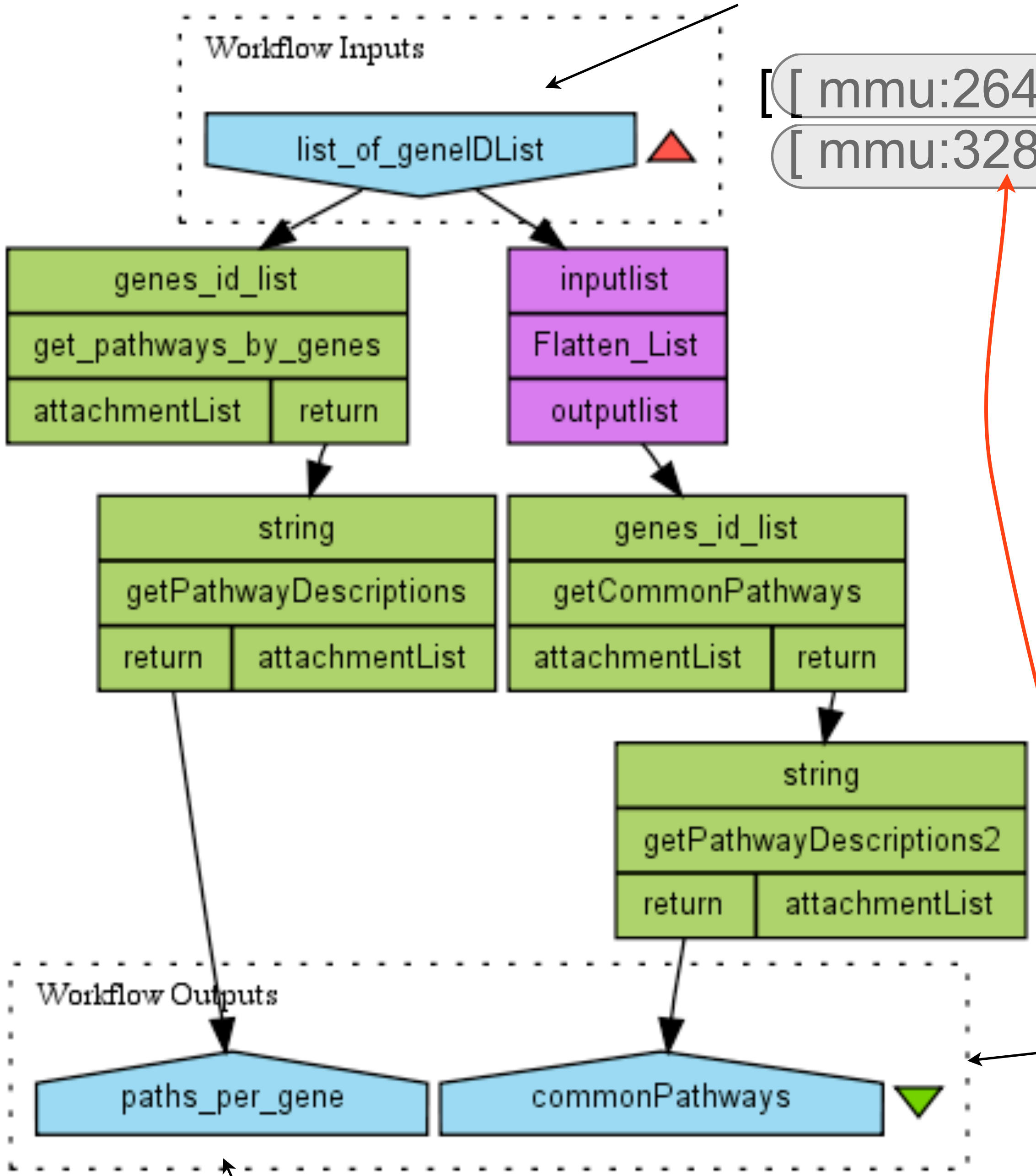




[path:mmu04010 MAPK signaling,
path:mmu04370 VEGF signaling]

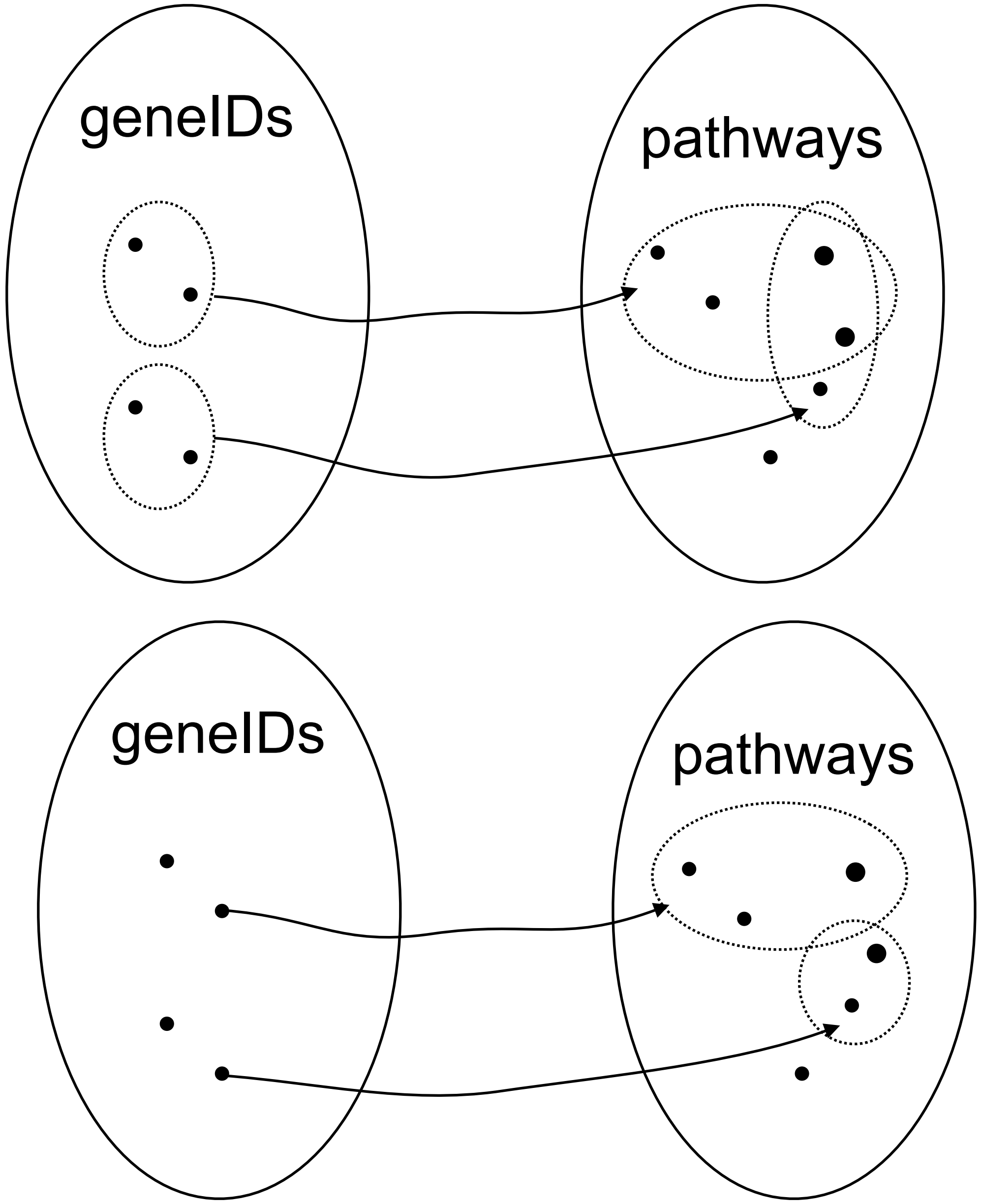
[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
  [path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Example: inverse data associations



List-structured
KEGG gene ids:

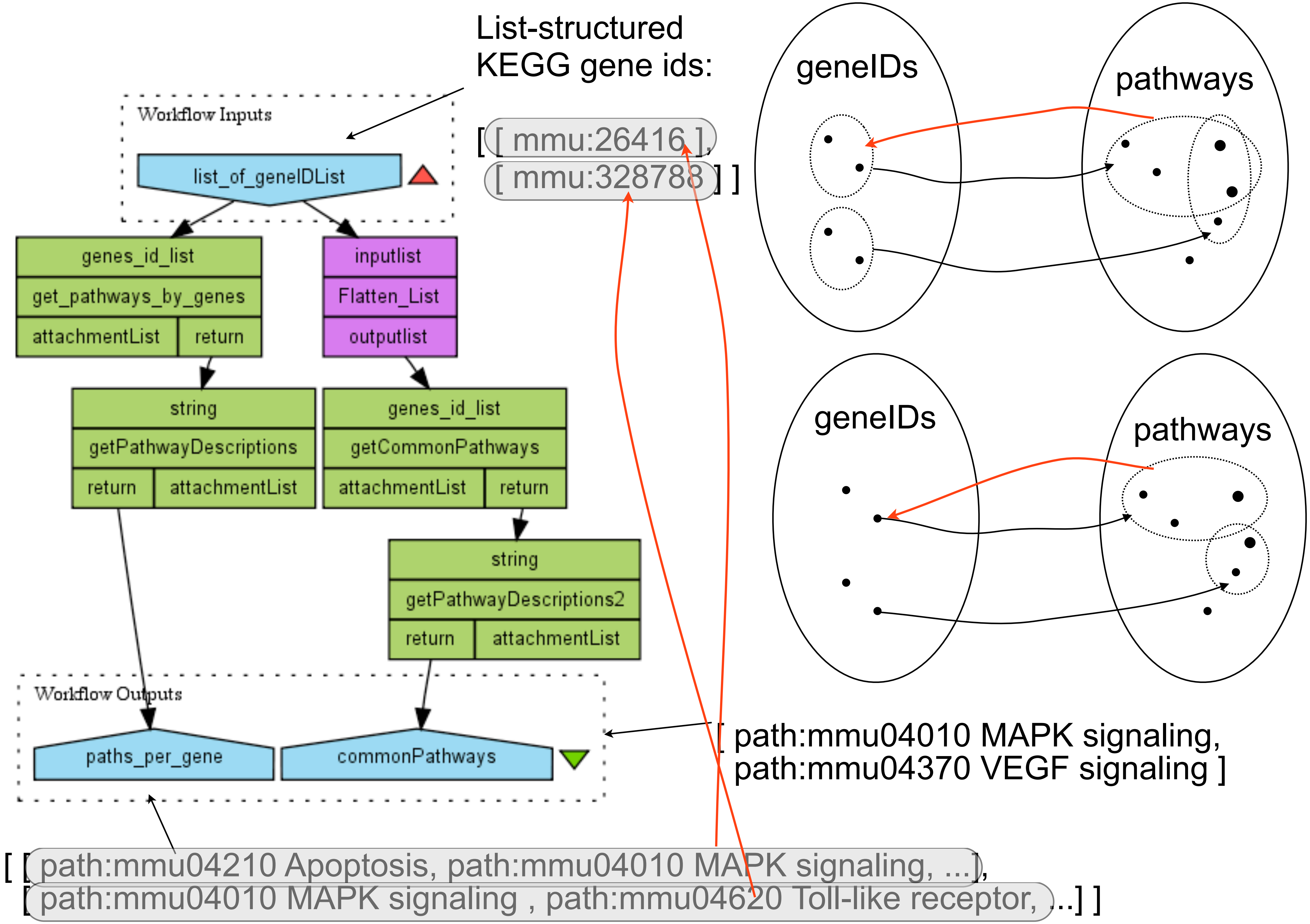
[[mmu:26416],
[mmu:328788]]



[path:mmu04010 MAPK signaling,
path:mmu04370 VEGF signaling]

[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
[path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Example: inverse data associations



List-structured KEGG gene ids:

[[mmu:26416],
[mmu:328788]]

[path:mmu04010 MAPK signaling,
path:mmu04370 VEGF signaling]

[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
[path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, ...]]

Taverna + provenance

- Taverna type system: strings + nested lists
  - “cat”, [“cat”, “dog”], [[“cat”, “dog”], [“large”, “small”]]

- Taverna dataflow model: data-driven execution
  - services activate when input is ready

- Workflow provenance: a detailed trace of workflow execution
  - which services were executed
  - when
  - inputs used, outputs produced

Taverna + provenance

- Taverna type system: strings + nested lists
  - “cat”, [“cat”, “dog”], [[“cat”, “dog”], [“large”, “small”]]
- Taverna dataflow model: data-driven execution
  - services activate when input is ready
- Workflow provenance: a detailed trace of workflow execution
  - which services were executed
  - when
  - inputs used, outputs produced

Taverna dataflow model + provenance traces can be a powerful combination

Forms of provenance ...

Focus is on the data: the observable outcomes of a process

	raw provenance metadata	provenance metadata + interpretation framework
design	• process structure (workflow graph) • history of process composition - reuse • process versions	• service annotations: • ex. get_pathways_by_genes • who created /edited: attribution • why: purpose, intent
execution	process events: - service invocation - data production / consumption - causal dependency graphs ex.: - list_of_geneIDList = [a, b, c] - paths_per_gene = [[d,e,f], [g,h,j]] - ... in run #32	- data annotations, results interpretation in terms of conceptual data model: set of pathways → gene sets

...and their uses and associated challenges

	raw provenance metadata	**provenance metadata + interpretation framework**
design	• exploiting semantic properties of the process structure to improve provenance exploitation • exploring process space across versions and structural similarities • graph matching	• semantic-based search of process space
execution	- enabling partial re-runs of resource-intensive workflows - storing very large provenance traces that accumulate over time - efficient query over large traces - presentation of query answers	- semantic-based query answering over annotated traces

...and their uses and associated challenges

	raw provenance metadata	provenance metadata + interpretation framework
design	• exploiting semantic properties of the process structure to improve provenance exploitation • exploring process space across versions and structural similarities • graph matching	• semantic-based search of process space
execution	- enabling partial re-runs of resource-intensive workflows - storing very large provenance traces that accumulate over time - efficient query over large traces - presentation of query answers	- semantic-based query answering over annotated traces

fully implemented in Taverna 2

to be released in Sept. 2009

...and their uses and associated challenges

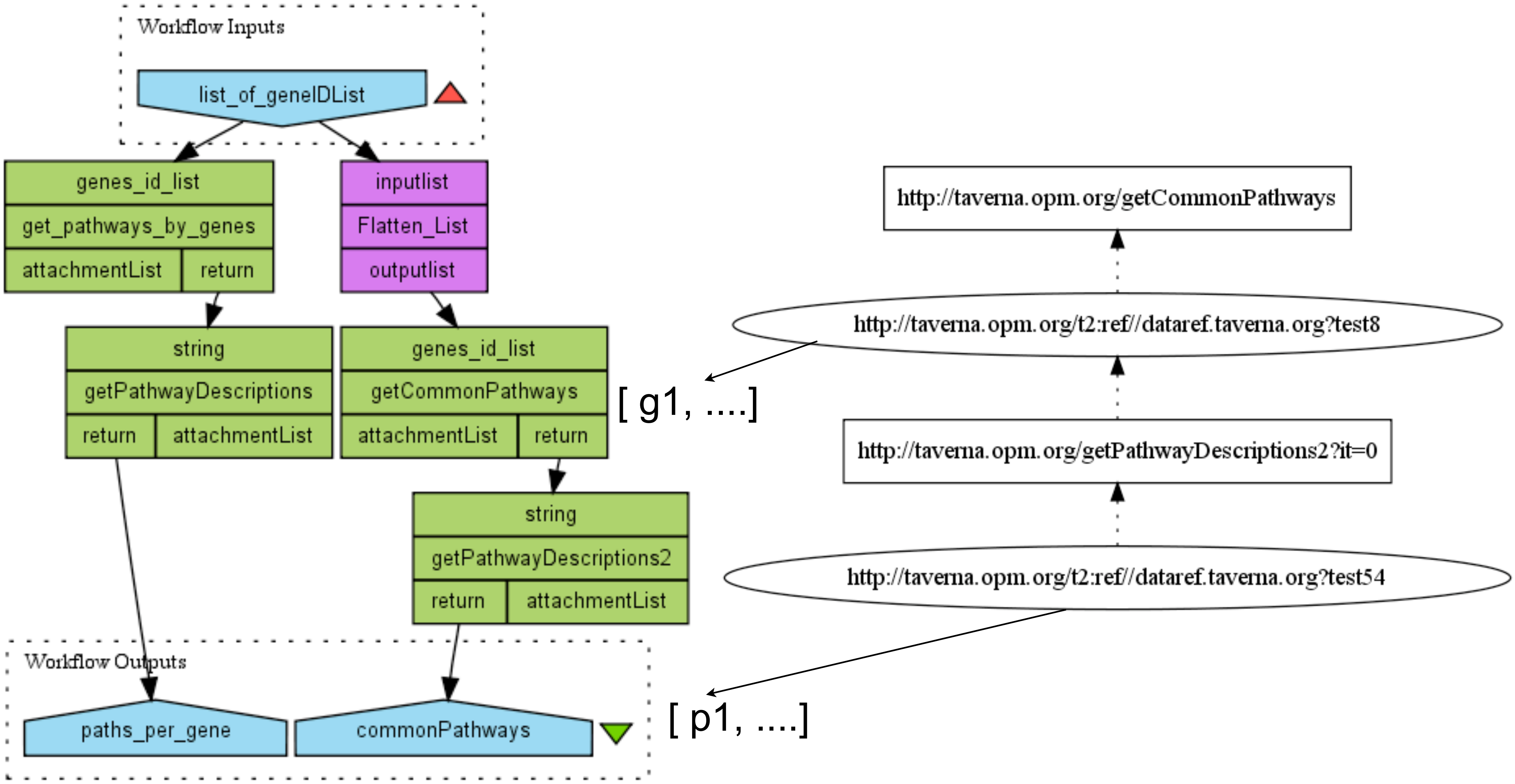
	raw provenance metadata	provenance metadata + interpretation framework
design	• exploiting semantic properties of the process structure to improve provenance exploitation • exploring process space across versions and structural similarities • graph matching	• semantic-based search of process space
execution	- enabling partial re-runs of resource-intensive workflows - storing very large provenance traces that accumulate over time - efficient query over large traces - presentation of query answers	- semantic-based query answering over annotated traces

fully implemented in Taverna 2

to be released in Sept. 2009

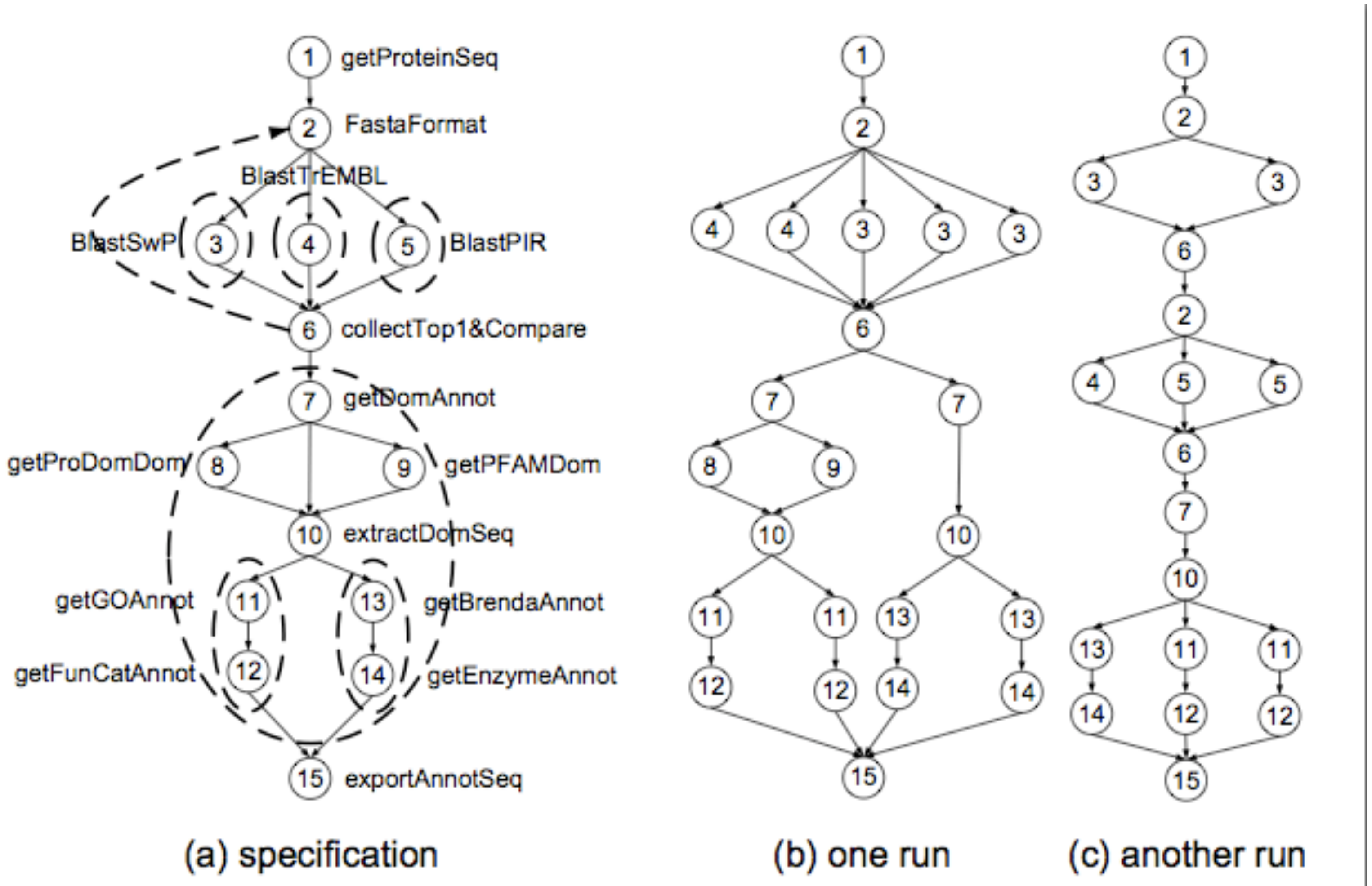
Querying provenance traces

- Lineage queries involve traversing a *provenance graph* from bottom to top



Naive provenance trace queries

- In most approaches, the originating process are not used for querying
- consequence: query requires provenance graph traversal
  - large traces → computationally complex
  - view materialization used in practice to get around the computational complexity



(a) specification

(b) one run

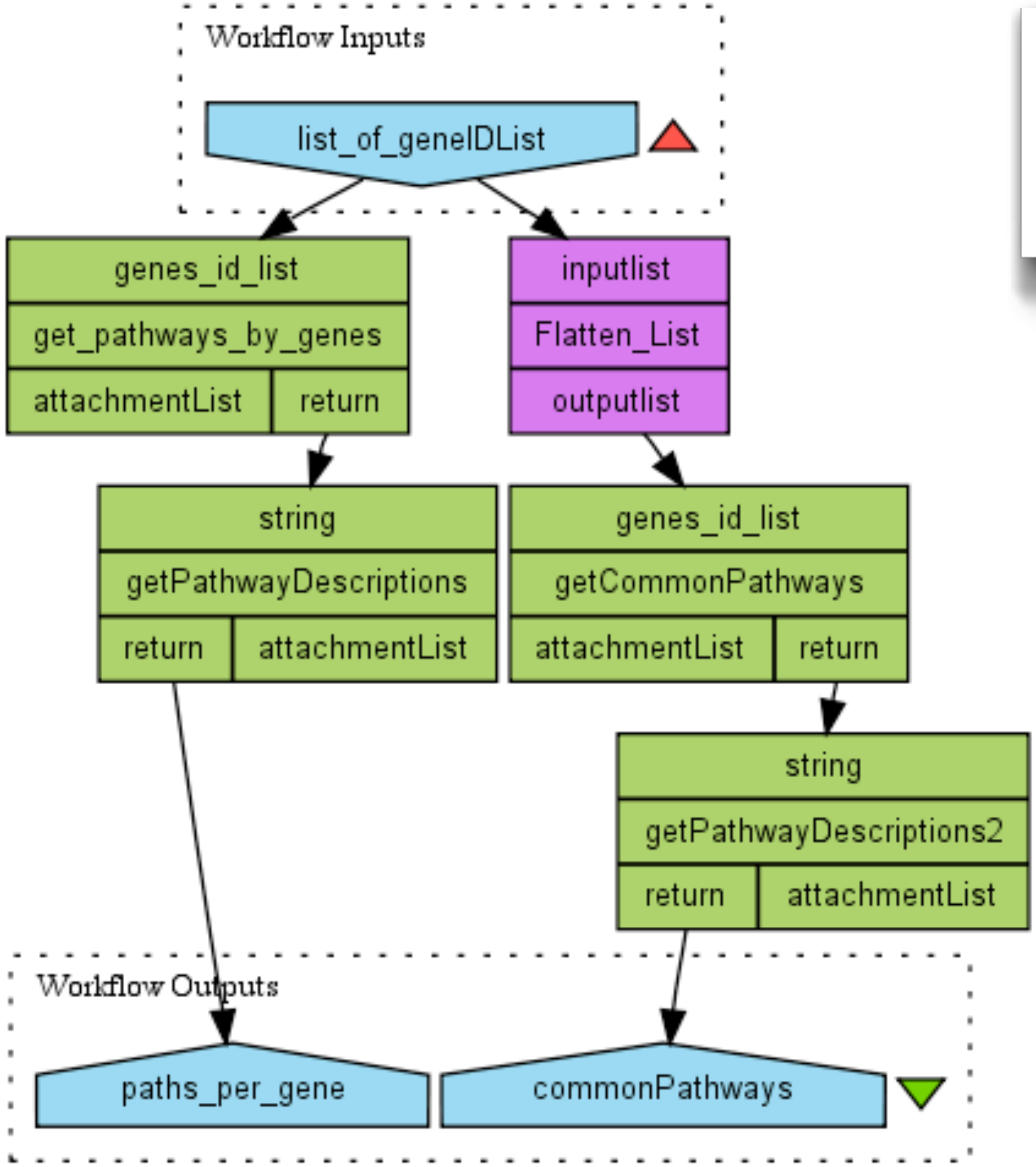
(c) another run

Fig. 1. Pr

Z. Bao and S. Cohen-Boulakia and S. Davidson and A. Eyal and S. Khanna, *Differencing Provenance in Scientific Workflows*, Procs. ICDE, 2009

Querying provenance graphs in Taverna

- Users are rarely interested in the complete provenance graph
  - noisy, possibly large, difficult to navigate

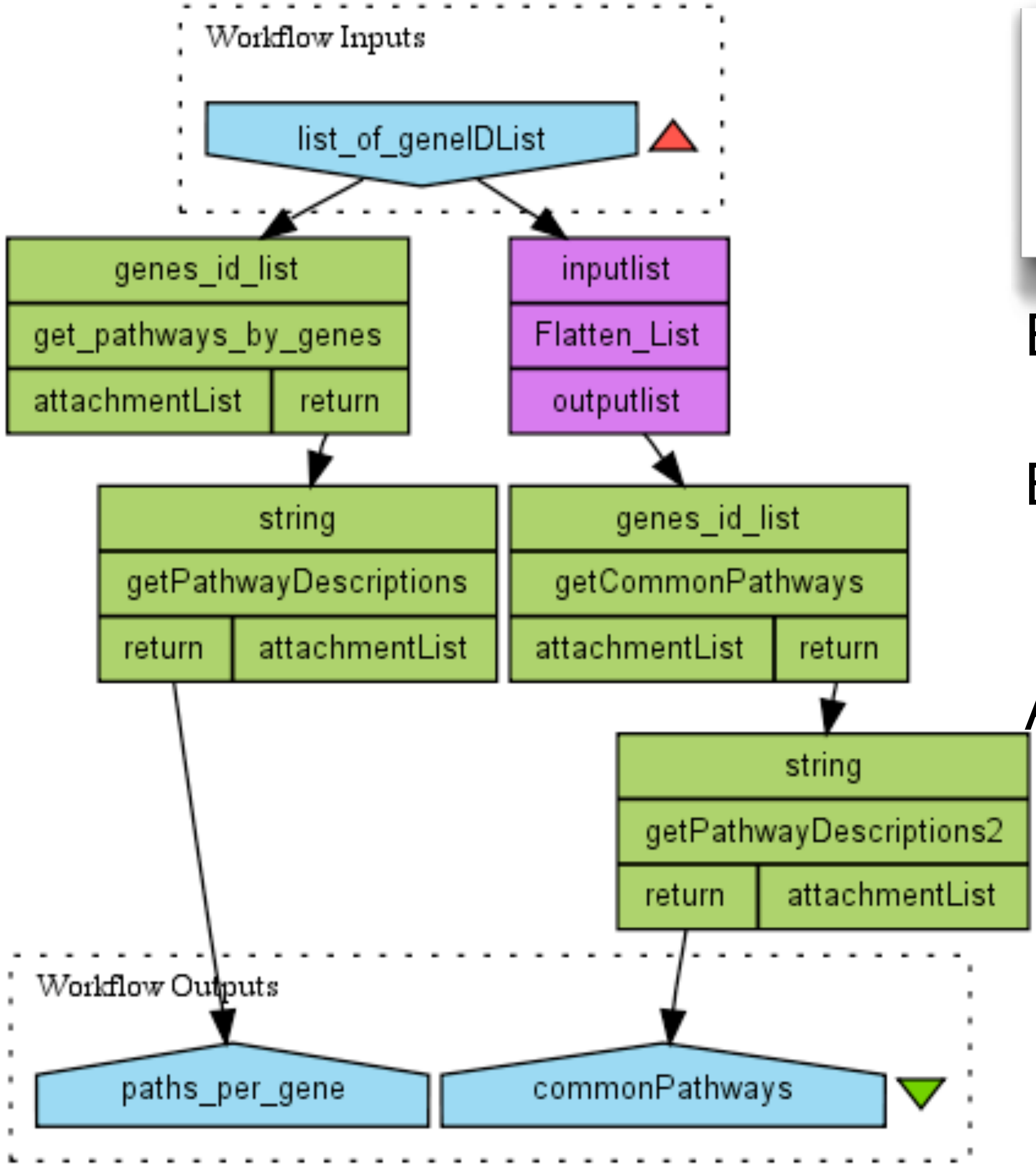


select interesting outputs
select interesting processors

This results in a more efficient lineage query algorithm that scales to large provenance graphs

Querying provenance graphs in Taverna

- Users are rarely interested in the complete provenance graph
  - noisy, possibly large, difficult to navigate



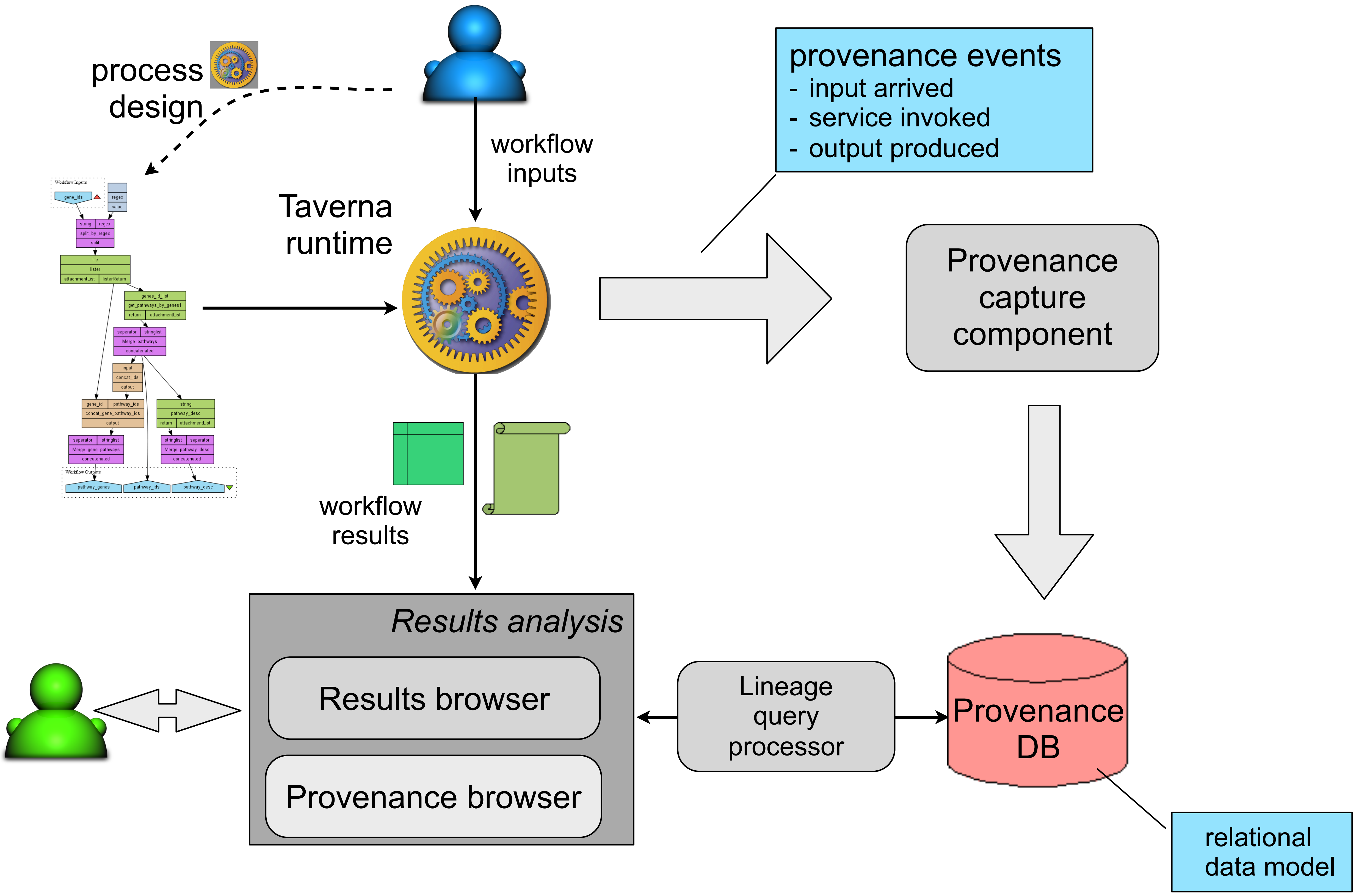
select interesting outputs
select interesting processors

Example:

```
BACKTRACE
    (paths_per_gene[3,4],  paths_per_gene[1,2])
        AT get_pathway_by_genes
AND
    commonPathways[1]
       AT TOP
```

This results in a more efficient lineage query algorithm that scales to large provenance graphs

Provenance management architecture



process design

workflow inputs

Taverna runtime

workflow results

provenance events
- input arrived
- service invoked
- output produced

Provenance capture component

Results analysis

Results browser

Provenance browser

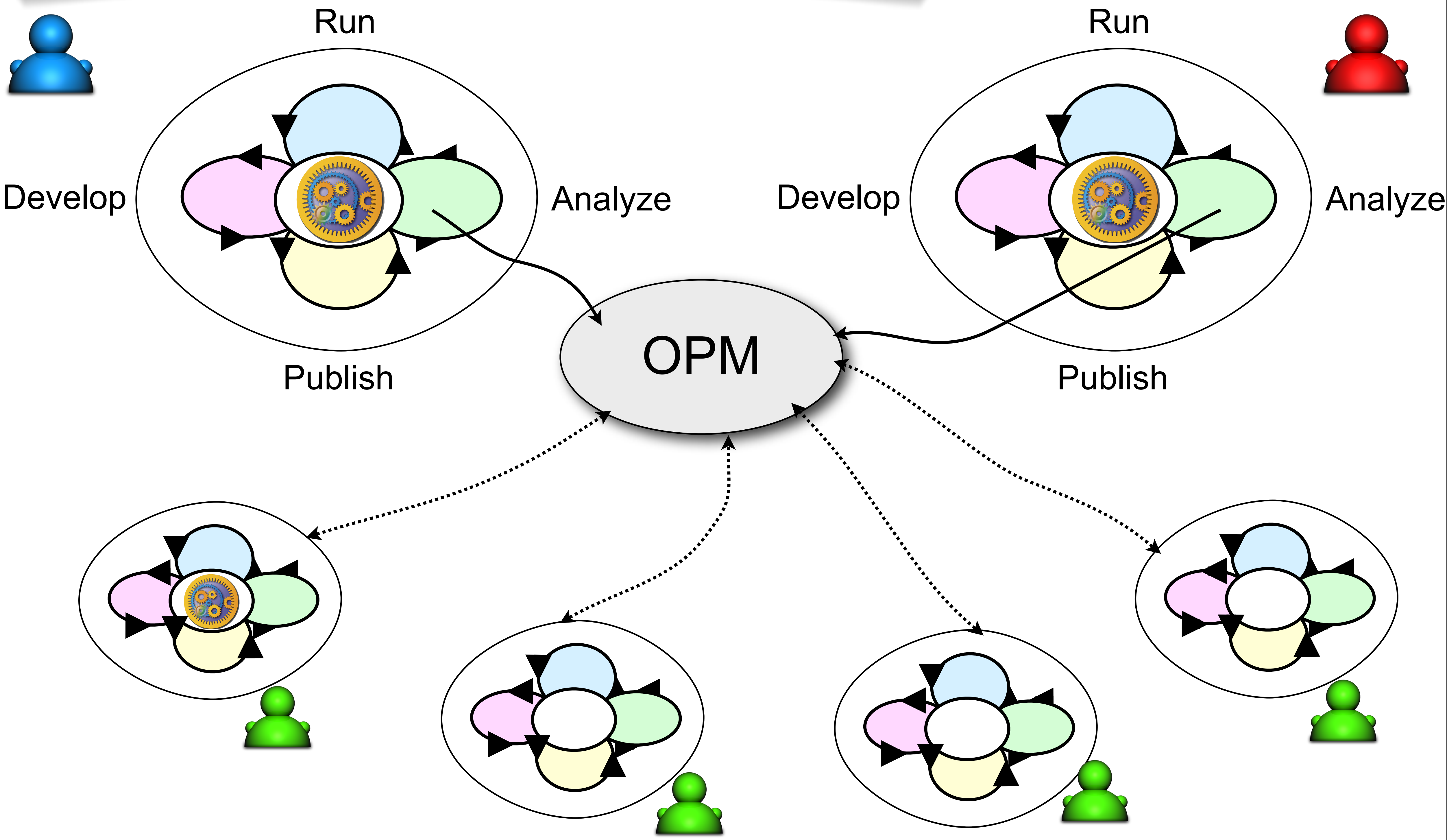
Lineage query processor

Provenance DB

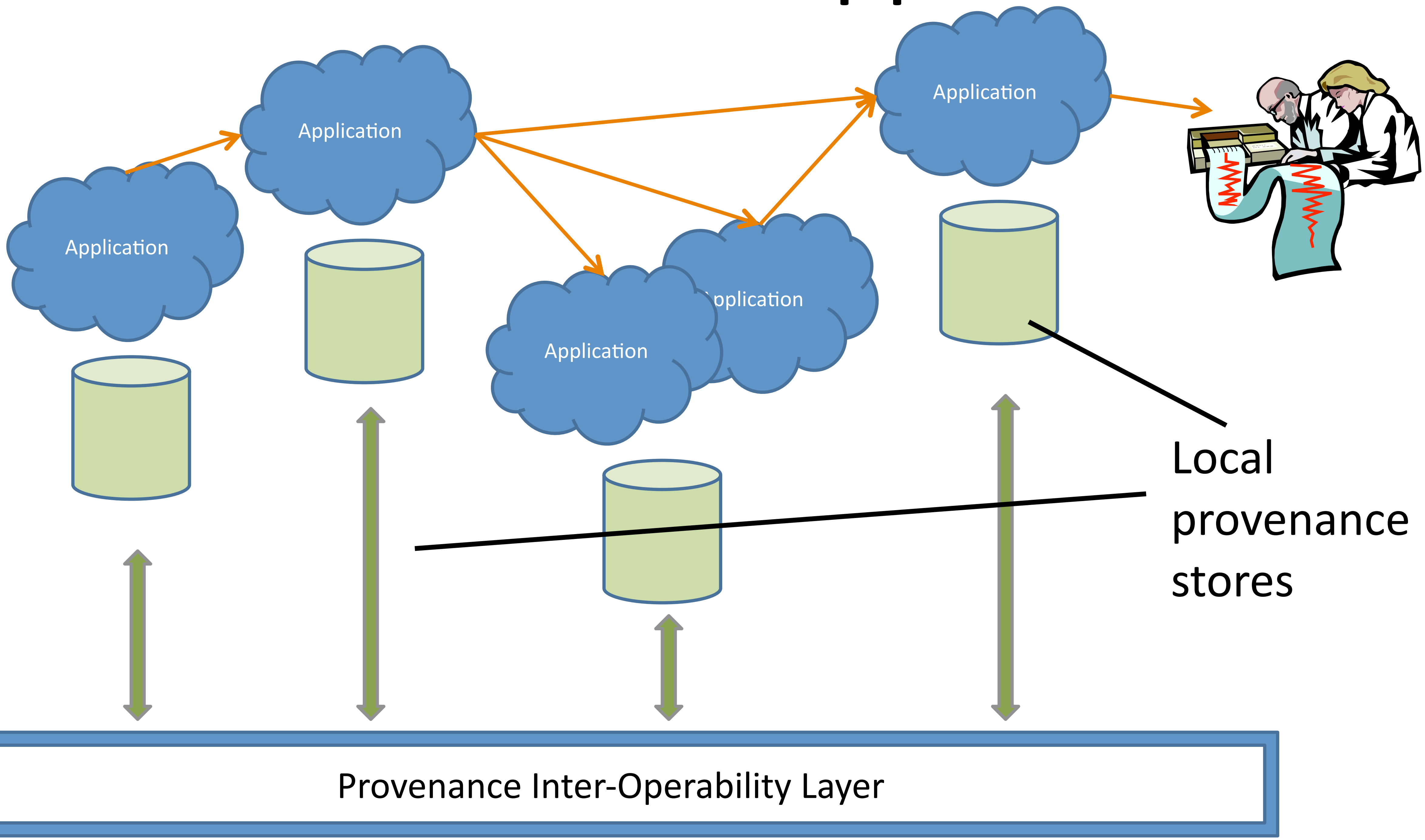
relational data model

Provenance interoperability for open science

OPM: the Open Provenance Model



Provenance Across Applications



Application

Application

Application

Application

pplication

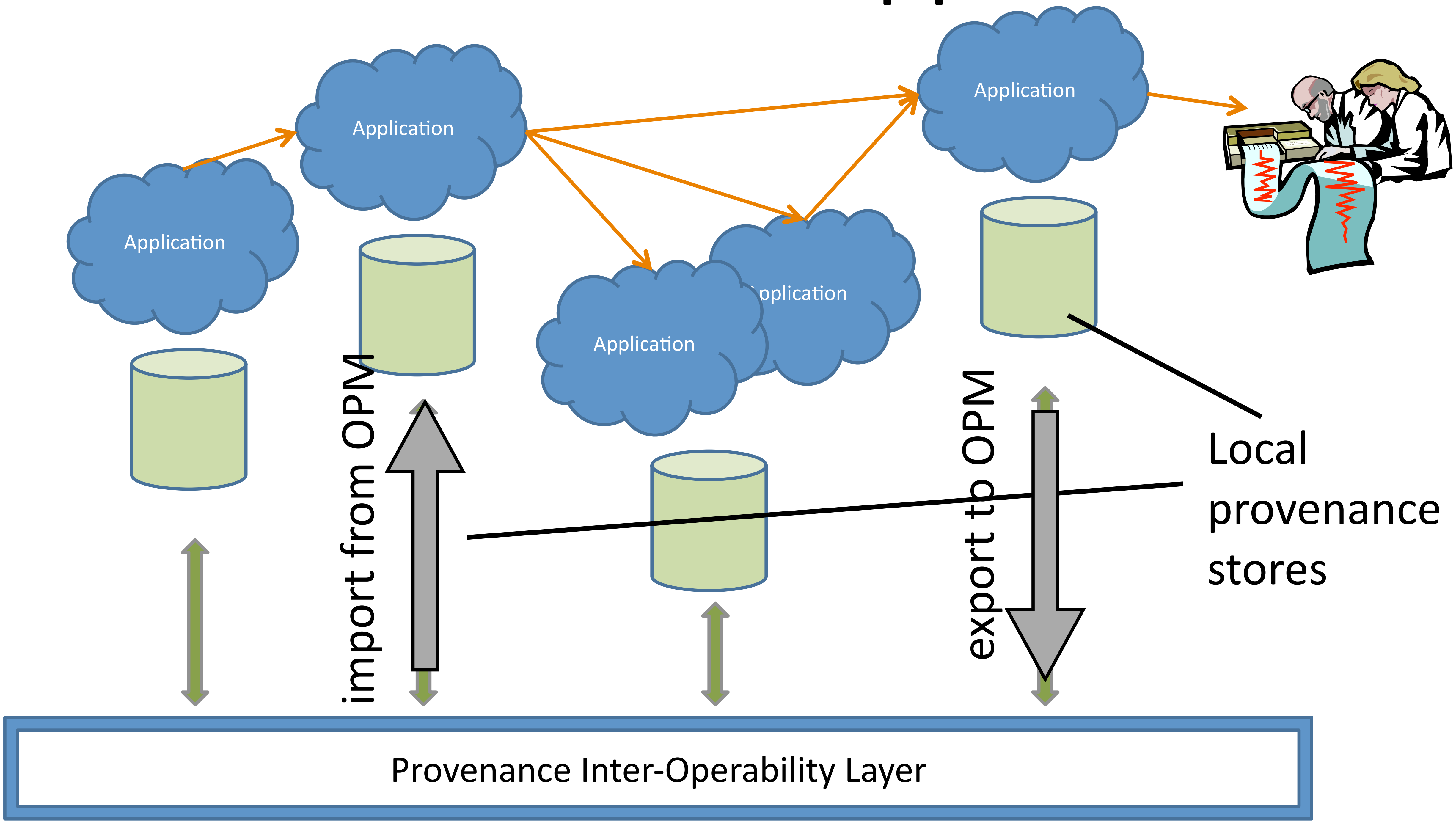
Application

Local provenance stores

Provenance Inter-Operability Layer

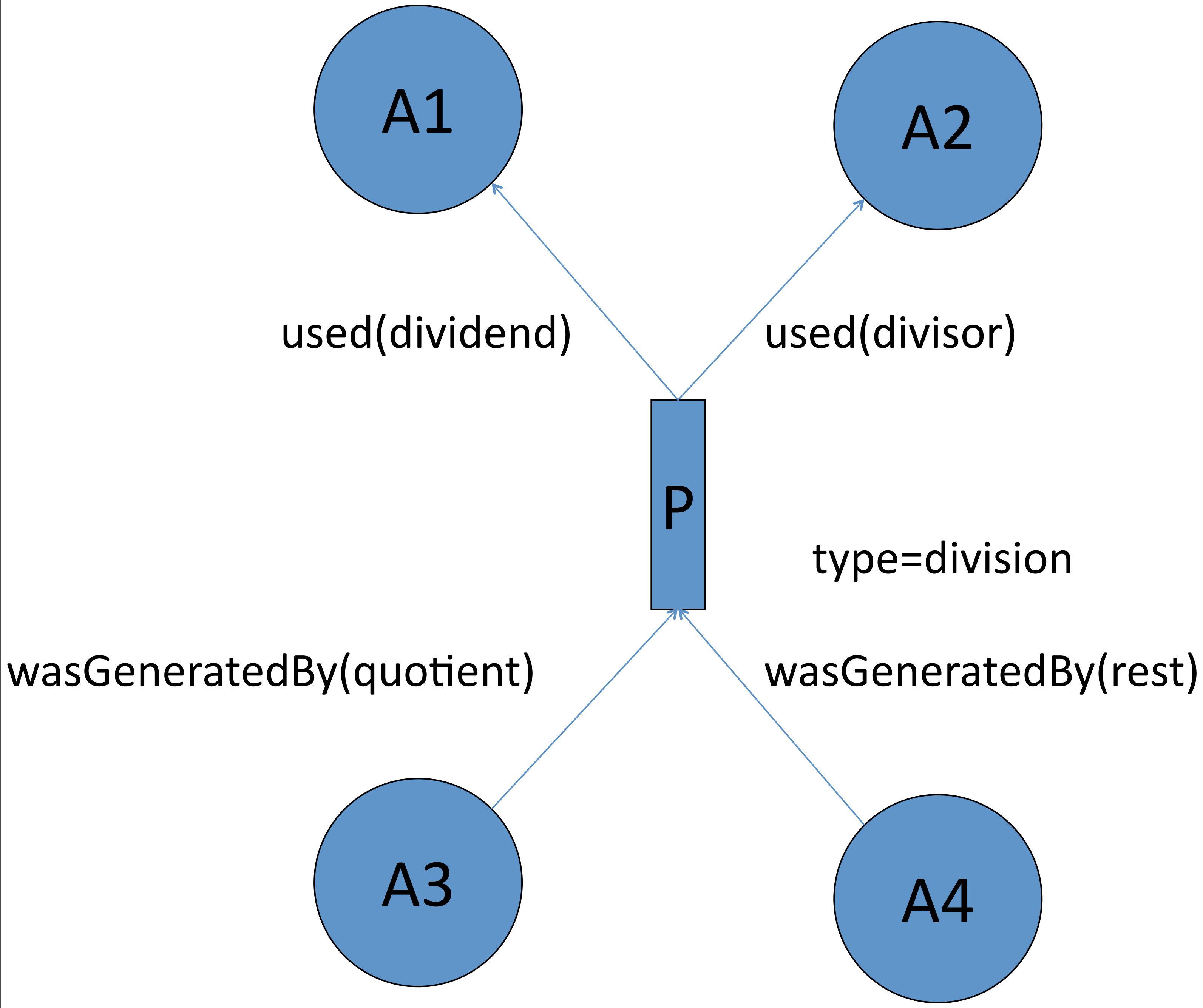
Adapted from Luc Moreau’s slides: “The Open Provenance Model” (Univ. of Southampton,UK), 2009

Provenance Across Applications



Adapted from Luc Moreau's slides: "The Open Provenance Model" (Univ. of Southampton,UK), 2009

Illustration



- Process “used” artifacts and “generated” artifact
- Edge “roles” indicate the function of the artifact with respect to the process (akin to function parameters)
- Edges and nodes can be typed

Causation chain:

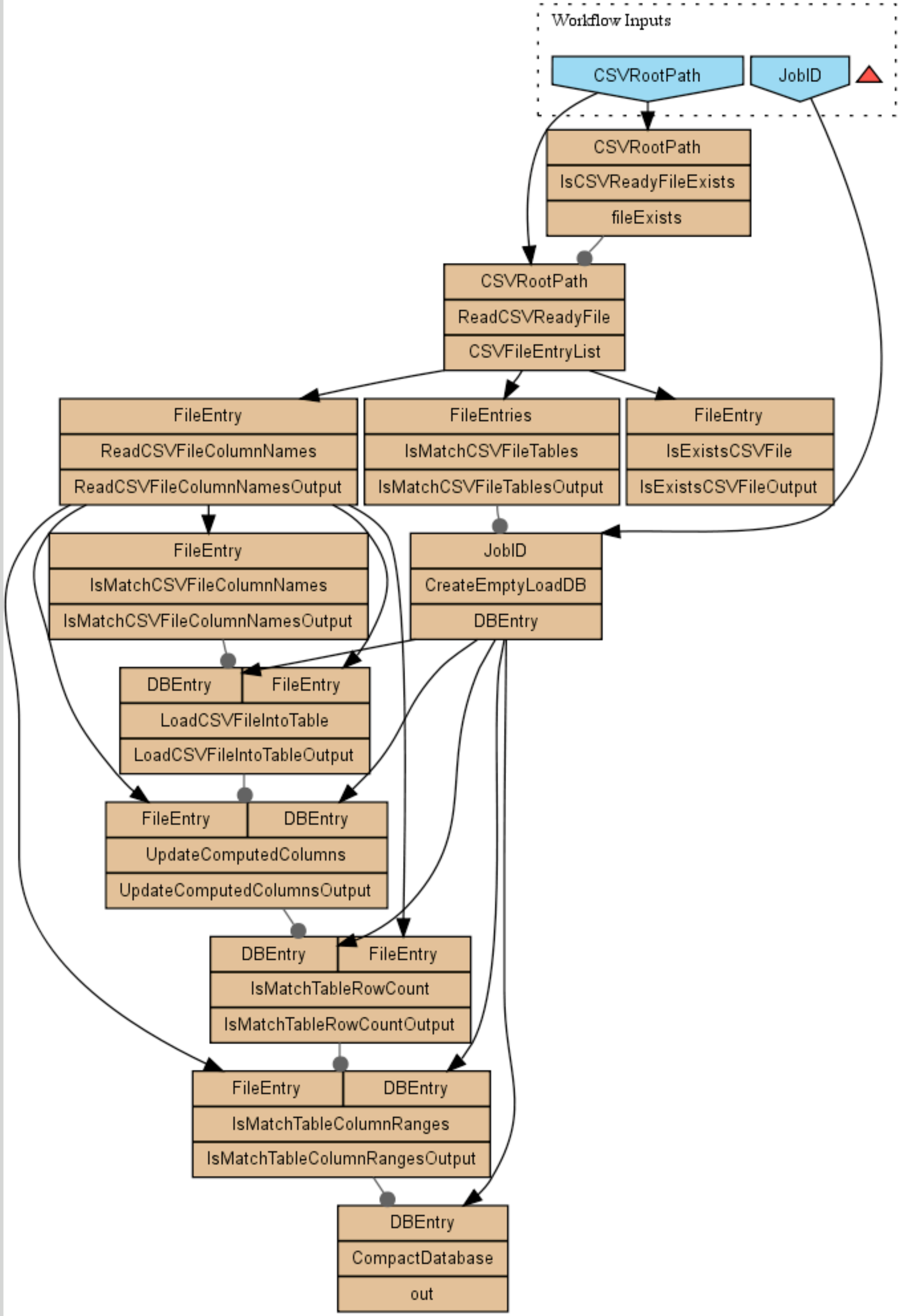
- P was caused by A1 and A2
- A3 and A4 were caused by P
- Does it mean that A3 and A4 were caused by A1 and A2?

From Luc Moreau’s slides set: “The Open Provenance Model” (Univ. of Southampton,UK), 2009

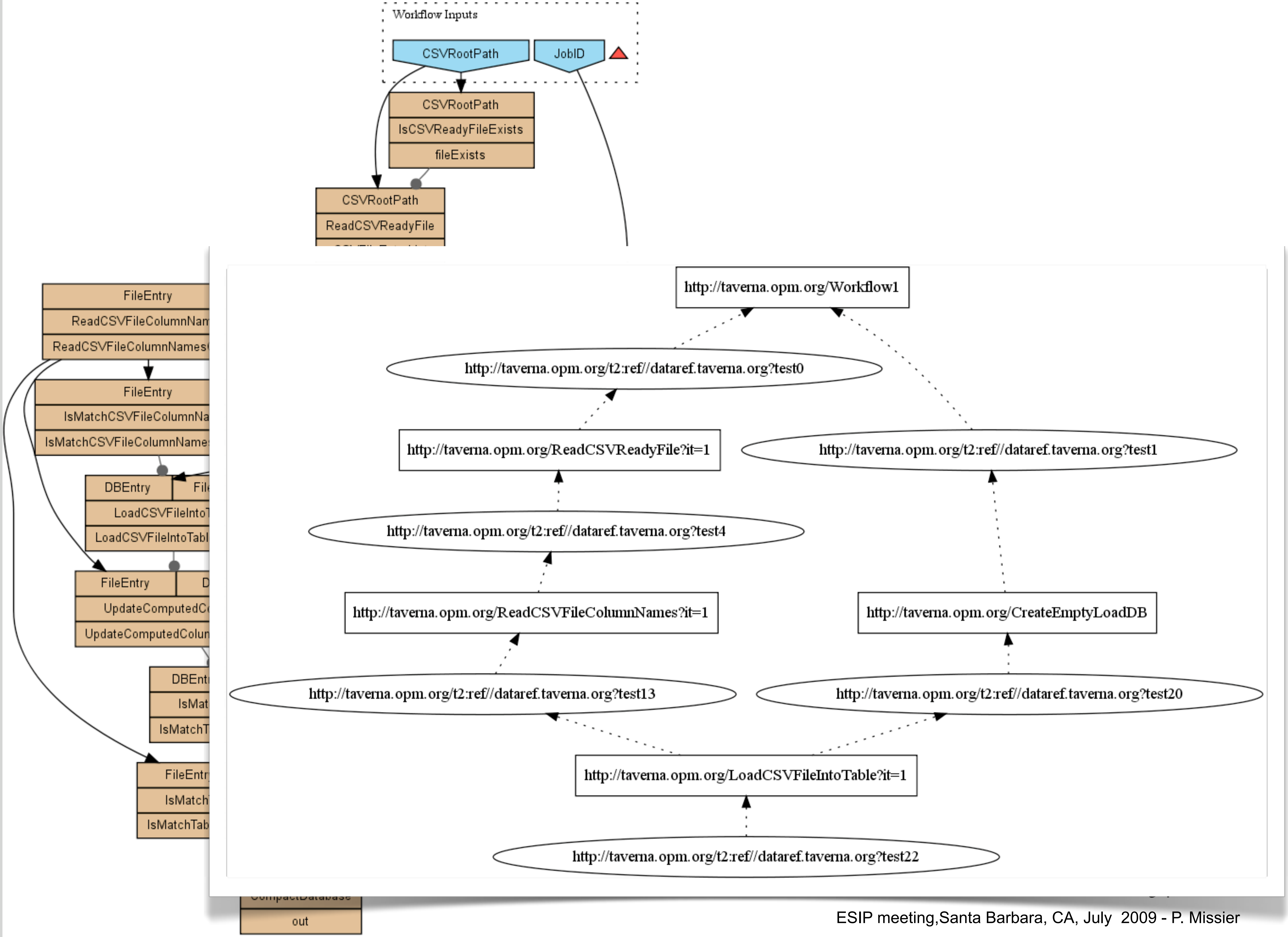
MANCHESTER 1824

The University of Manchester

Integrated OPM generation in Taverna

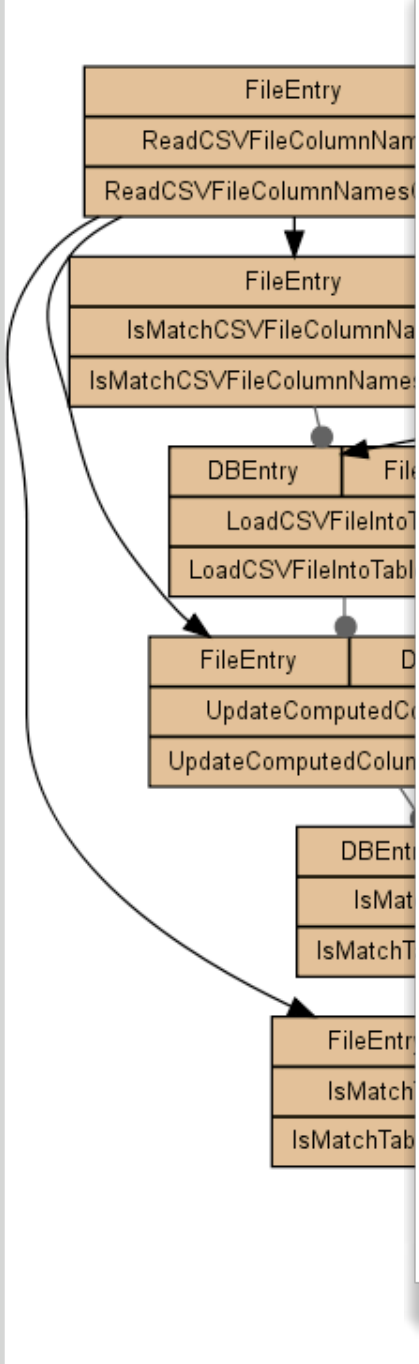


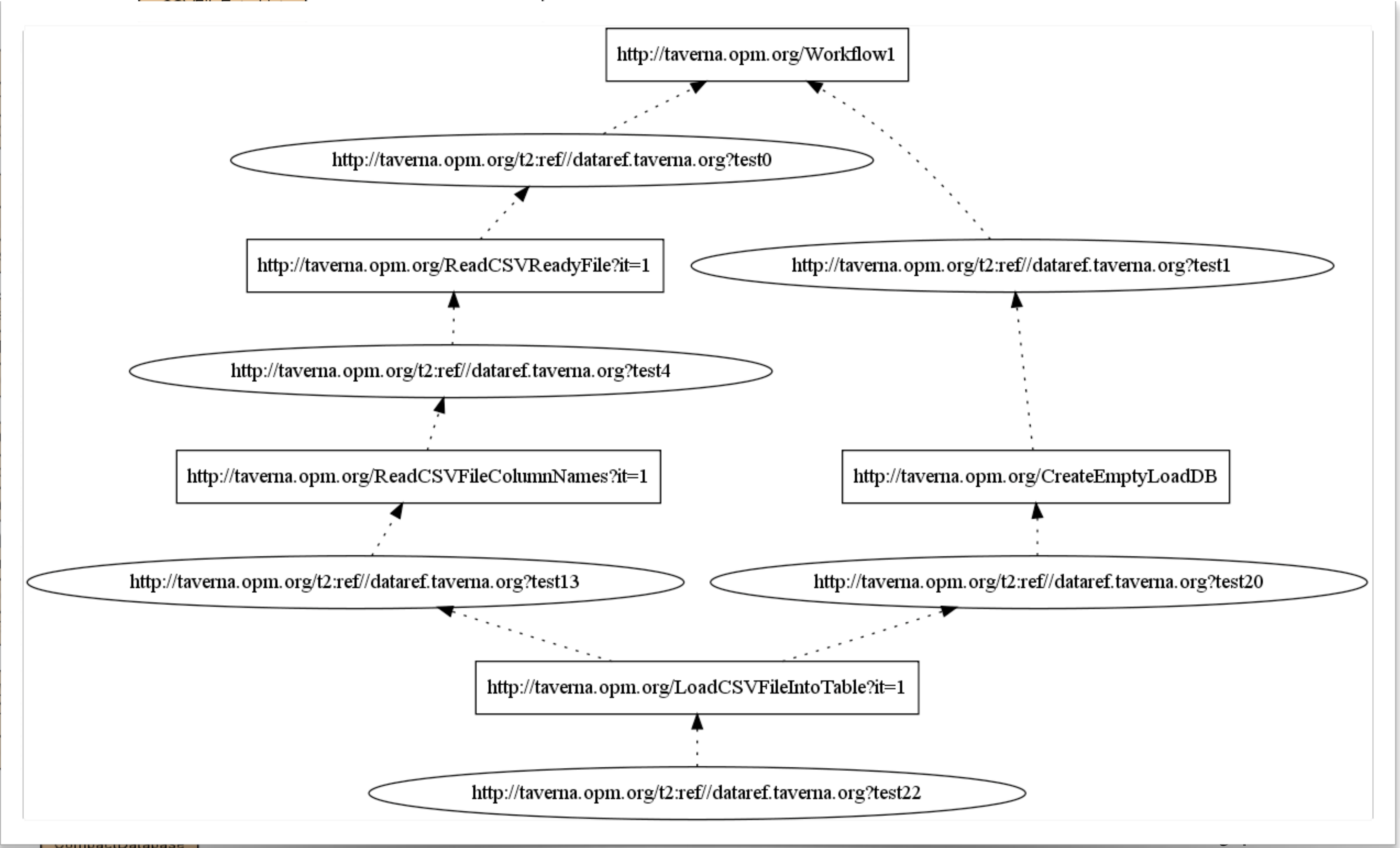
Integrated OPM generation in Taverna


Workflow Inputs
CSVRootPath
JobID
CSVRootPath
IsCSVReadyFileExists
fileExists
CSVRootPath
ReadCSVReadyFile
FileEntry
ReadCSVFileColumnNam
ReadCSVFileColumnNames(
FileEntry
IsMatchCSVFileColumnNa
IsMatchCSVFileColumnName
DBEntry
File
LoadCSVFileIntoT
LoadCSVFileIntoTabl
FileEntry
D
UpdateComputedC
UpdateComputedColun
DBEntr
IsMat
IsMatchT
FileEntr
IsMatch
IsMatchTab
out
http://taverna.opm.org/Workflow1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test0
http://taverna.opm.org/ReadCSVReadyFile?it=1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test4
http://taverna.opm.org/ReadCSVFileColumnNames?it=1
http://taverna.opm.org/CreateEmptyLoadDB
http://taverna.opm.org/t2:ref//dataref.taverna.org?test13
http://taverna.opm.org/t2:ref//dataref.taverna.org?test20
http://taverna.opm.org/LoadCSVFileIntoTable?it=1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test22

Integrated OPM generation in Taverna

- ➡ the answer to any TP query can be viewed as an OPM graph
- ➡ encoded as RDF/XML using the Tupelo provenance API (NCSA)




http://taverna.opm.org/Workflow1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test0
http://taverna.opm.org/ReadCSVReadyFile?it=1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test4
http://taverna.opm.org/ReadCSVFileColumnNames?it=1
http://taverna.opm.org/CreateEmptyLoadDB
http://taverna.opm.org/t2:ref//dataref.taverna.org?test13
http://taverna.opm.org/t2:ref//dataref.taverna.org?test20
http://taverna.opm.org/LoadCSVFileIntoTable?it=1
http://taverna.opm.org/t2:ref//dataref.taverna.org?test22

More information on OPM

- The OPM wiki:
  - http://twiki.ipaw.info/bin/view/OPM/
    - open to discussions and contributions
    - please read the governance doc
- The 3rd provenance challenge:
  - produce and export OPM graphs
    - interoperable XML and RDF serializations
  - import and query third party graphs
  - http://twiki.ipaw.info/bin/view/Challenge/ThirdProvenanceChallenge
  - The University of Manchester's contribution to the challenge:
    - http://twiki.ipaw.info/bin/view/Challenge/UoM
  - latest meeting held in June, 2009 (Amsterdam)

From knowledge capture to exploitation

- answering user questions effectively

  (using provenance + semantics infrastructure)
  - has a similar investigation been undertaken before? when, by whom? what was the outcome?
  - have alternative services being used? to what effect?
  - what have been the users' decisions, and why?

- Enabling collaborative science:
  - provide users with recommendations on the next steps in their session, based on analysis of their past actions;
  - cluster users within a group based on their common interests, observed through the choices they make during the sessions
  - promote socal/scientific networking
    - a "blog the lab" flavour

From knowledge capture to exploitation

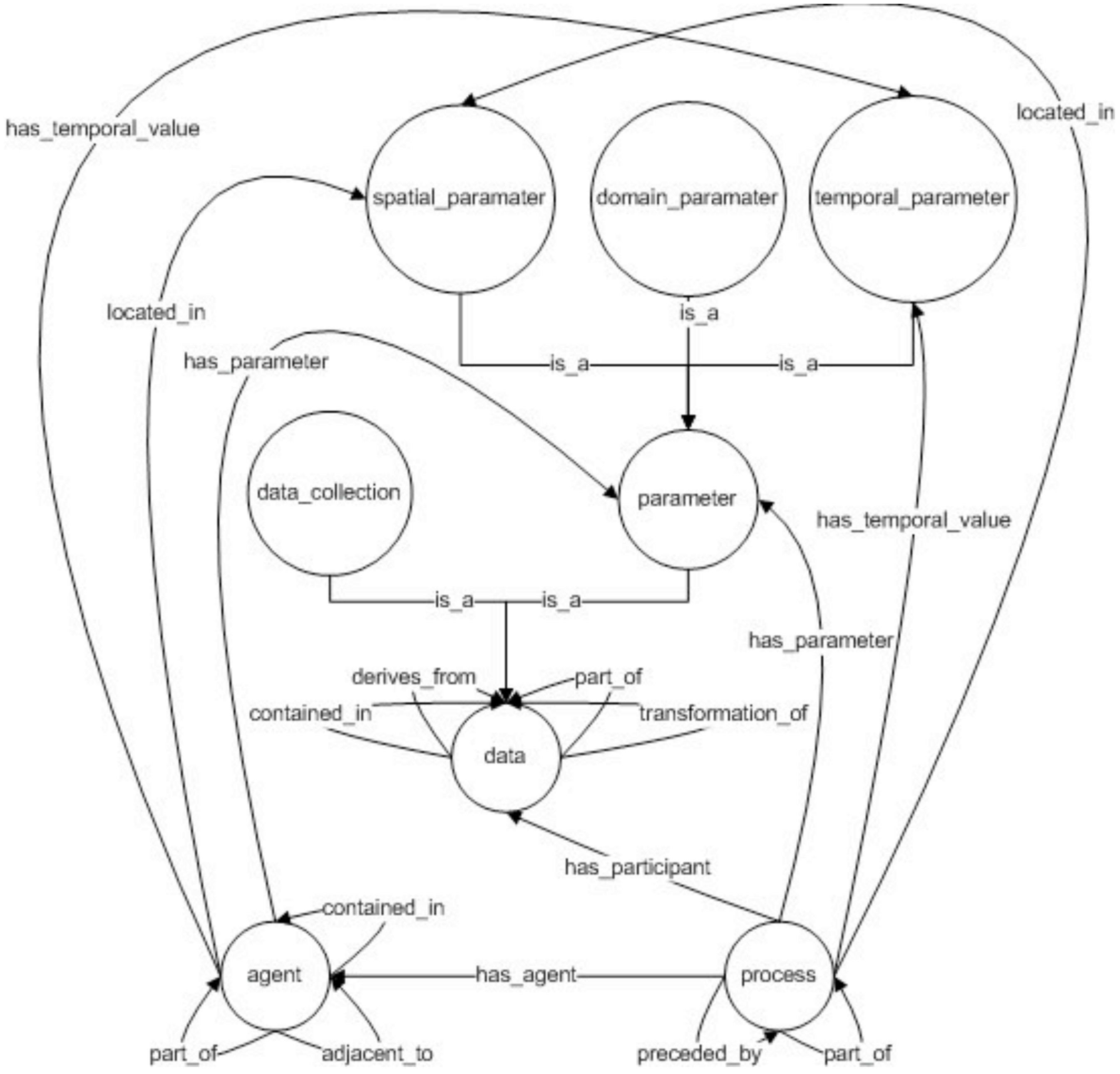
- answering user questions effectively (using provenance + semantics infrastructure)
  - has a similar investigation been undertaken before? when, by whom? what was the outcome?
  - have alternative services being used? to what effec
  - what have been the users' decisions, and wh

- Enabling collaborative scie
  - provide users with recon n the next steps in their session, based on analysis o tions;
  - cluster users based on their common interests, observed through ey make during the sessions
  - promo cientific networking
    - a "bl the lab" flavour

(Semantic) provenance analytics

The Provenir ontology

- Upper ontology with for domain-specific extensions
- OWL, designed for reasoning and RDF queries



Satya S. Sahoo, Roger S. Barga, Jonathan Goldstein, Amit P. Sheth, *Where did you come from...Where did you go?" An Algebra and RDF Query Engine for Provenance*, TR-2009-03, Kno.e.sis Center, CSE Dept., Wright State University, Dayton, OH, March, 2009

Upcoming events

SWPM 2009:
The First International Workshop on
the Role of Semantic Web in Provenance Management

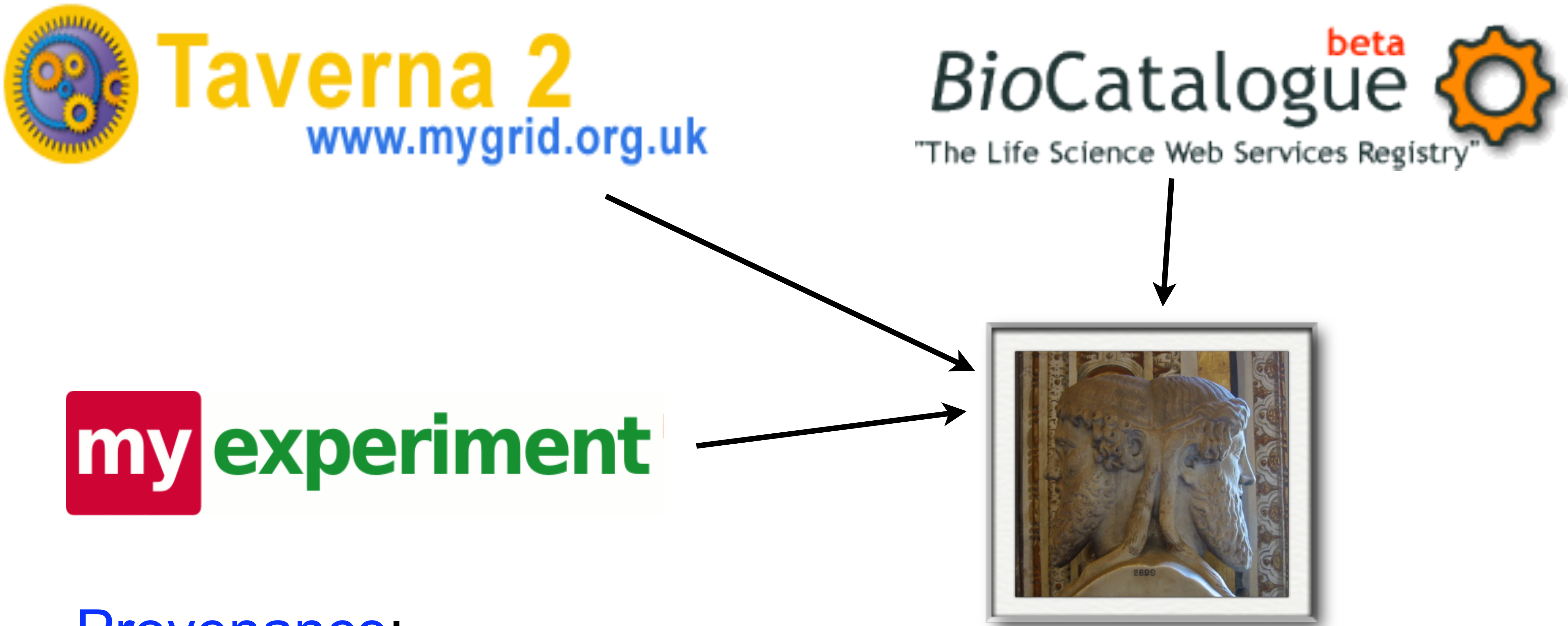
http://wiki.knoesis.org/index.php/SWPM-2009

Co-located with ISWC'09, October 25/26 2009, Washington D.C., USA
Submission Deadline: Friday, July 31, 2009

Special issue of **Future Generation Computer Systems Journal** (FGCS) on the third provenance challenge
(to be announced)

expected deadline: **Dec., 2009**

Summary: Support for collaborative science



Provenance:

– automated metadata collection and processing

1. data management angle: “logs with a proper data model”

– storage and query issues

– interoperability

2. **social angle**: attribution chain for experimental artifacts

– processes, data, annotations

– myExperiment packs → Research Objects

Selected literature on provenance

- P. Buneman, S. Khanna, W. Chiew Tan, Why and Where: *A Characterization of Data Provenance*, Procs. **ICDT, 2001**
- Susan B. Davidson and Juliana Freire, *Provenance and scientific workflows: challenges and opportunities*, Procs. **SIGMOD, 2008**
- Z. Bao and S. Cohen-Boulakia and S. Davidson and A. Eyal and S. Khanna, *Differencing Provenance in Scientific Workflows*, Procs. **ICDE, 2009**
- Luc Moreau, Paul Groth, Simon Miles, Javier Vazquez-Salceda, John Ibbotson, Sheng Jiang, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, Laszlo Varga, *The provenance of electronic data*, **Communications of the ACM, Vol. 51 No. 4, Pages 52-58, 2008**
- P. Missier, K. Belhajjame, J. Zhao, C. Goble, *Data lineage model for Taverna workflows with lightweight annotation requirements*, Procs. **IPAW 200**8
- M. Anand, S. Bowers, T. McPhillips, B. Ludaescher, *Efficient Provenance Storage over Nested Data Collections*, Procs. **EDBT, 2009**
- J. Zhao, C. Goble, R. Stevens, D. Turi, *Mining Taverna's semantic web of provenance*, **Concurrency and Computation: Practice and Experience, Vol. 20 no. 5, 2008**.
- R. S. Barga and L. A. Digiampietri, *Automatic capture and efficient storage of e-Science experiment provenance*, **Concurrency and Computation: Practice and Experience, Vol. 20 no. 8, 2008**