Taverna, Biocatalogue, and myExperiment: a three-legged foundation for effective collaboration in E-science

A collaborative talk by Paolo Missier

Information Management Group
School of Computer Science, University of Manchester, UK

with additional material kindly shared by:
Prof. Dave DeRoure and David Newman, University of Southampton
Prof. Carole Goble and the e-Labs design group, University of Manchester
• Vision
  – Embracing collaboration in experimental science
  – Sharing is happening, let’s do it right!

• Technology
  – Current architectural components
    • Taverna, myExperiment, Biocatalogue
  – Available vocabularies, formats that we use
    • SIOC, SKOS, ORE,...

• Open issues (Research??)
  – Sharing in the large: social and technical issues
  – Producing value, avoiding pollution
On sharing in science

• Traditionally protective culture
  – no data sharing prior to publication
  – raw data is public, results are private

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

Mike Ashburner and others
Professor of Genetics,
University of Cambridge, UK
On sharing in science

• Traditionally protective culture
  – no data sharing prior to publication
  – raw data is public, results are private

“A biologist would rather share their toothbrush than their gene name”
Mike Ashburner and others
Professor of Genetics,
University of Cambridge, UK

• Scope
  – Collaboration in the small vs in the wild
    • control over sharing
    • trust management
  – Experimental, information-centric, computational science
    • experimental processes encoded as workflows
1- Momentum on sharing and collaboration

• Ongoing debate in many communities
  – Clinical trials [1]
  – Earth Sciences -- ESIP - data preservation / stewardship, 2009

• Science Commons recommendations for Open Science
  – Open Science recommendations from Science Commons (July 2008) [link]

• Publishers’ requirements for data and methods disclosure
  – for archiving, experiment repeatability

Emerging community standards for information exchange (bioinformatics)

MIBBI: Minimum Information for Biological and Biomedical Investigations

MIAME: Minimum Information About a Microarray Experiment

MIAPE: The Minimum Information About a Proteomics Experiment (PSI - HUPO) [1]

Incentives to sharing

Special issue of Nature on Data Sharing (Sept. 2009)


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009
http://www.nature.com/news/specials/datasharing/index.html
Incentives to sharing

Special issue of Nature on Data Sharing (Sept. 2009)

“The draft human genome sequence was the highest profile dataset rapidly released before publication [...] This experience ultimately demonstrated that the **broad and early availability** of sequence data greatly benefited life sciences research by leading to many new insights and discoveries, including new information on 30 disease genes published prior to the draft sequence.”


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009  
http://www.nature.com/news/specials/datasharing/index.html
Incentives to sharing

Special issue of Nature on Data Sharing (Sept. 2009)

- timeliness requires rapid sharing
- repurposing
- data quality control

“The draft human genome sequence was the highest profile dataset rapidly released before publication [...] This experience ultimately demonstrated that the broad and early availability of sequence data greatly benefited life sciences research by leading to many new insights and discoveries, including new information on 30 disease genes published prior to the draft sequence.”


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009 http://www.nature.com/news/specials/datasharing/index.html
Incentives to sharing

Special issue of Nature on Data Sharing (Sept. 2009)

- timeliness requires rapid sharing
- repurposing
- data quality control

One of the lessons from the Human Genome Project (HGP) was the recognition that making data broadly available prior to publication can be profoundly valuable to the scientific enterprise and lead to public benefits.

This is particularly the case when there is a community of scientists that can productively use the data quickly — beyond what the data producers could do themselves in a similar time period, and sometimes for scientific purposes outside the original goals of the project.


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009
http://www.nature.com/news/specials/datasharing/index.html
Incentives to sharing

**Special issue of Nature on Data Sharing (Sept. 2009)**

- timeliness requires rapid sharing
- repurposing
- data quality control

The draft human genome sequence was the highest profile dataset rapidly released before publication. This experience ultimately demonstrated that the broad and early availability of sequence data greatly benefited life sciences research by leading to many new insights and discoveries, including new information on 30 disease genes published prior to the draft sequence.

One of the lessons from the Human Genome Project (HGP) was the recognition that making data broadly available prior to publication can be profoundly valuable to the scientific enterprise and lead to public benefits. This is particularly the case when there is a community of scientists that can productively use the data quickly — beyond what the data producers could themselves do in a similar time period, and sometimes for scientific purposes outside the original goals of the project.

To date, many papers have been published by third parties reporting research findings enabled by data sets released before publication. The experiences shared in Toronto suggest that these have rarely affected subsequent publications authored by the data producer.


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009 | http://www.nature.com/news/specials/datasharing/index.html
Incentives to sharing

Special issue of Nature on Data Sharing (Sept. 2009)

- timeliness requires rapid sharing
- repurposing
- data quality control

The data users should carefully read the source information, including any marker papers, associated with a released data set. Data analysts should pay particular attention to any caveats about data quality, because rapidly released data are often unstable, in that they may not yet have been subjected to full quality control and so may change.


Prepublication data sharing:
Nature 461, 168-170 (10 September 2009) | doi:10.1038/461168a; Published online 9 September 2009 http://www.nature.com/news/specials/datasharing/index.html
The social process of Science

- Digital Libraries
- Peer-Reviewed Journal & Conference Papers
- Preprints & Metadata
- Repositories
- Local Web
- Technical Reports
- Virtual Learning Environment

scientists

experimentation

- Certified Experimental Results & Analyses
- Data, Metadata
- Provenance
- Workflows
- Ontologies

Undergraduate Students
Graduate Students
The social process of Science 2.0
The social process of Science 2.0

Digital Libraries

Reprints

Peer-Reviewed Journal & Conference Papers

Preprints & Metadata

Local Web

Repositories

scientists

Virtual Learning Environment

Undergraduate Students

Graduate Students

experimentation

Data, Metadata

Certified Experimental Results & Analyses

Provenance

Workflows

Ontologies
Keep up to date
Get the latest news about what your online community is doing and what’s happening with your Research Objects.

Form Friends & Groups
Explore and manage the social network. You have fine control over the privacy and sharing of your Research Objects.

Find Workflows
See the latest and most popular workflows: discover, view, download, run, tag and rate. Upload your workflows.

Build Packs
Share collections of items as individual packs — like all the digital items in an experiment. Include external items too.

Content types
We provide special support for workflow systems including Taverna and Trident, as well as experiment plans, providing a foundation for the e-Laboratory.

Curating process
Workflows capture pieces of research process which are curated by their authors, experts and the community. These curation models are also used in the Biocatalogue service registry.

All about the Research Object
See and manage all the essential extrinsic information and ‘social metadata’ — licence, tags, sharing, ratings.

Credits and attributions are an essential feature to support flow of rights and reputation.

All about me
Easy navigation using a dashboard of all the things relating to me and my social network.
Title: Escherichia coli: From cDNA Microarray Raw Data to Pathways and Published Abstracts

Type: Taverna 1

Download Scalable Diagram (SVG)
Workflow-based experimentation lifecycle

1. **Run**
2. **Analyse Results**
3. **Develop**
4. **Publish**

**Taverna 2**

**BioCatalogue**

Collect and query provenance metadata

myexperiment
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)
We are not alone...
Scaling up along the social dimension

Develop
Run
Analyze
Publish

What?
Where?
Why?
Who?
How?
Scaling up along the social dimension
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.
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - http://rdf.myexperiment.org/ontologies/
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - [http://rdf.myexperiment.org/ontologies/](http://rdf.myexperiment.org/ontologies/)

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384">
    <dcterms:title>What’s On Next</dcterms:title>
    <dcterms:description>Queries the BBC to find out what’s on next...
    </dcterms:description>
    <sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438" />
    <dcterms:created>2008-08-28T15:23:58Z</dcterms:created>
    <mebase:filename>what_s_on_next_18085.xml</mebase:filename>
    <ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384" />
    <mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
```
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - [http://rdf.myexperiment.org/ontologies/](http://rdf.myexperiment.org/ontologies/)

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384"

  <dcterms:title>What’s On Next</dcterms:title>
  <dcterms:description>Queries the BBC to find out what’s on next...</dcterms:description>

  <sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438"/>
  <dcterms:created>2008-08-28T15:23:58Z</dcterms:created>
  <mebase:filename>what_s_on_next_18085.xml</mebase:filename>
  <ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384"/>

  <mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
```

myExperiment / contributions module
MyExperiment’s open information model

• Entries described using an RDF data model
• Entries available through a REST API
• Using a combination of home-grown and standard ontologies
  • [http://rdf.myexperiment.org/ontologies/](http://rdf.myexperiment.org/ontologies/)

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384">
  <dcterms:title>What’s On Next</dcterms:title>
  <dcterms:description>Queries the BBC to find out what’s on next...<dcterms:description>
  <sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438"/>
  <dcterms:created>2008-08-28T15:23:58Z</dcterms:created>
  <mebase:filename>what_s_on_next_18085.xml</mebase:filename>
  <ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384"/>
  <mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
</mecontrib:Workflow>
```

myExperiment / contributions module

Dublin core
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - [http://rdf.myexperiment.org/ontologies/](http://rdf.myexperiment.org/ontologies/)

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384">
  <dcterms:title>What’s On Next</dcterms:title>
  <dcterms:description>Queries the BBC to find out what’s on next...</dcterms:description>
  <sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438"/>
  <dcterms:created>2008-08-28T15:23:58Z</dcterms:created>
  <mebase:filename>what_s_on_next_18085.xml</mebase:filename>
  <ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384"/>
  <mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
</mecontrib:Workflow>
```

- myExperiment / contributions module
- Dublin core
- SIOC ontology
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - http://rdf.myexperiment.org/ontologies/

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384">
  <dcterms:title>What’s On Next</dcterms:title>
  <dcterms:description>Queries the BBC to find out what’s on next...</dcterms:description>
  <sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438" />
  <dcterms:created>2008-08-28T15:23:58Z</dcterms:created>
  <mebase:filename>what_s_on_next_18085.xml</mebase:filename>
  <ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384" />
  <mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
</mecontrib:Workflow>
```

- myExperiment / contributions module
- Dublin core
- SIOC ontology
- myExperiment / base module
MyExperiment’s open information model

- Entries described using an RDF data model
- Entries available through a REST API
- Using a combination of home-grown and standard ontologies
  - [http://rdf.myexperiment.org/ontologies/](http://rdf.myexperiment.org/ontologies/)

```xml
<mecontrib:Workflow rdf:about="http://rdf.myexperiment.org/Workflow/384"
    dcterms:title="What’s On Next"
    dcterms:description="Queries the BBC to find out what’s on next..."
    sioc:has_owner rdf:resource="http://rdf.myexperiment.org/User/438"
    dcterms:created="2008-08-28T15:23:58Z"
    mebase:filename="what_s_on_next_18085.xml"
    ore:isDescribedBy rdf:resource="http://rdf.myexperiment.org/ResourceMap/Workflow/384"
    mecontrib:preview rdf:datatype="&xsd;anyURI">http://www.myexperiment.org/workflow/image/384/what_s_on_next_18085_1.png</mecontrib:preview>
```

- [myExperiment / contributions module](http://myexperiment.org)
- [Dublin core](http://dublincore.org)
- [myExperiment / base module](http://myexperiment.org)
- [SIOC ontology](http://viaf.org/vocabulary/sioc)
- [ORE ontology](http://openarchives.org/ore/)
- [myExperiment / modules](http://myexperiment.org)
• Keys to controlled sharing

<mebase:has-policy>
  <snarm:Policy rdf:about="http://rdf.myexperiment.org/Policy/733">
    <snarm:has-access>
      <snarm:RestrictedAccess rdf:about="http://rdf.myexperiment.org/Policy/733/Access/ViewUser438">
        <snarm:has-accesser rdf:resource="http://rdf.myexperiment.org/User/438"/>
        <snarm:has-access-type rdf:resource="&mespec;View"/>
      </snarm:RestrictedAccess>
    </snarm:has-access>
  </snarm:Policy>
</mebase:has-policy>

Simple Network Access Rights Management (SNARM)
Attribution: **David R. Newman**, myExperiment team, University of Southampton, UK
• Keys to controlled sharing

  <mebase:has-policy>
    <snarm:Policy rdf:about="http://rdf.myexperiment.org/Policy/733">
      <snarm:has-access>
        <snarm:RestrictedAccess rdf:about="http://rdf.myexperiment.org/Policy/733/Access/ViewUser438">
          <snarm:has-accesser rdf:resource="http://rdf.myexperiment.org/User/438"/>
          <snarm:has-access-type rdf:resource="&mespec;View"/>
        </snarm:RestrictedAccess>
        <snarm:has-access>
          <snarm:RestrictedAccess rdf:about="http://rdf.myexperiment.org/Policy/733/Access/ViewUser438">
            <snarm:has-accesser rdf:resource="http://rdf.myexperiment.org/User/438"/>
            <snarm:has-access-type rdf:resource="&mespec;View"/>
          </snarm:RestrictedAccess>
        </snarm:has-access>
      </snarm:has-access>
    </snarm:Policy>
  </mebase:has-policy>

  Simple Network Access Rights Management (SNARM)
  Attribution: David R. Newman, myExperiment team, University of Southampton, UK
http://sioc-project.org/

Good tool support
Commonly used in conjunction with the FOAF vocabulary
Aside from its well-established role in promoting the stabilization of microtubules (MTs), tau may have additional functions as a result of its interactions with other structures and enzymes.
The e-Labs collaboration model

An e-Lab consists of:

1. a community
2. research objects = data + methods + annotations
3. management infrastructure for research objects

- their abstract properties
  - life cycle
  - state
  - compound structure
  - provenance

- the rules and mechanisms for sharing
  - attribution, derivation, access control
  - aggregation / de-aggregation into components
  - introspection
  - exchange protocols
Results

Workflow 16

Analysis Protocol for Candidate Genes and Pathways

This protocol is aimed at providing a guide to the interpretation of the results obtained from both the 3D and 2D analysis of gene expression analysis. The protocol is divided into the following sections:

- Genes and Pathways
- Methods
- Results
- Discussion

The results are generated using the following analysis:

- Genes and Pathways
- Pathway Analysis
- Hierarchical Clustering

The results are presented in a tabular format and include:

- Gene ID
- Gene Name
- Gene Expression

Metadata

Workflow Inputs:
- microarray_pathways
- qtl_pathways
- common_pathways
- kegg_pathways

Workflow Outputs:
- intersecting_pathways

Common pathways

Workflow 13

Paul’s Pack

Logs

Slides

Paper

QTL
Paul's Pack

Workflow 16

Results

Logs

Metadata

Slides

Paper

QTL

Workflow Inputs
- microarray_pathways
- cell_pathways
- common_pathways
- kegg_pathways

Workflow Outputs
- intersecting_pathways

Common pathways

Workflow 13

Analysis Protocol for Candidate Genes and Pathways

This protocol is aimed at providing a guide to the interpretation of the results obtained from both microarray and pathway datasets. The protocol includes a detailed analysis of the results obtained from the datasets, along with a discussion of the implications of the findings. The protocol also includes a list of key genes and pathways that are of interest, along with a summary of the results obtained from each dataset.
Current support in myExperiment: packs

Title: Taverna 2 workflow example and associated provenance

Description

test workflow along with a raw provenance dump which can be imported into mySQL

[edit]

Items (2)

Workflow: Workflow for provenance testing -- example 1 (Paolo)

[Add a comment here]
Added by Paolo ... 3 days ago (10/09/09 @ 14:22:33)

File: Provenance DB dump to be used as a reference (Paolo)

[Add a comment here]
Added by Paolo ... 3 days ago (10/09/09 @ 14:22:28)

Paolo uses myExperiment to interact with users whilst testing his Taverna provenance component.
1. An e-Lab is built from a collection of services, consuming and producing Research Objects.
Anatomy of a research object

1. Composite
2. Distributed
3. Annotated
4. Repeatable
5. Interoperable
ORE: representing generic aggregations

http://www.openarchives.org/ore/1.0/primer.html section 4
Research Objects: a concrete example

Source: e-Labs and Research Objects White Paper, School of Computer Science, University of Manchester, upcoming
Moving forward

- Research objects as the next currency in the reputation market?
- From quantity to quality to value
Summary

- Collaborative / open science gaining momentum

- How do we get the collaboration model right?

- Technology:
  - many initiatives in the vocabulary / ontology space
  - it’s mostly about adoption, integration
  - myExperiment moving quickly (and adventurously)
    - second round of funding secured from UK funding body

- The (potential) pitfalls:
  
  quantity without quality = trouble...