

The Reality of Reproducibility in Computational Science

reproduce? repeat? rerun? does it matter?

Prof Carole Goble FREng FBCS

The University of Manchester, UK

carole.goble@manchester.ac.uk

Based on:

e-Science 2012 Chicago, October 2012

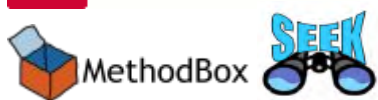
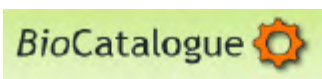
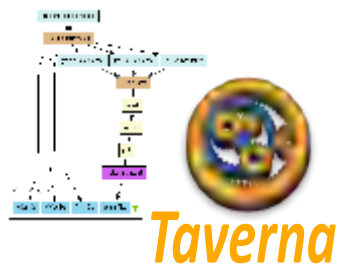
<https://dl.dropbox.com/u/617206/eScience-2012-GOBLE-release-nonotes.ppt>

JCDL 2012 Washington DC, June 2012

<https://dl.dropbox.com/u/617206/JCDL%20Goble%20Final%20Clean-nobigbird.ppt>

Scholarly Communication Workshop, 14-15 January 2013, Pittsburgh, USA

Products



Methods

Computational Methods

Scientific workflows. In the wild.
Distributed web/grid/cloud services
Cyber-Infrastructure

Social Methods: Sharing and Exchange

e-Laboratories for scientific artefacts. Libraries, Repositories and Catalogues for data, models, web services, workflows, scripts, SOPs...

Knowledge Management

Semantic technology, semantic applications, Linked Open Data, research objects, executable papers, publishing

Software Engineering

Software Sustainability Institute
Open Middleware Infrastructure Institute, S/W and Data Policy
Institutional Repository

Applications

Astronomy

**Library
Digital
Preservation**

Biodiversity

Biology

Systems Biology

Chemistry

Public Health

Astro-Physics

Social Science

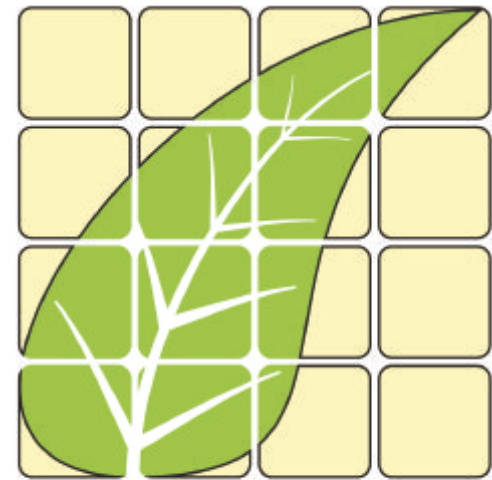


Service and Workflows



Systems Biology of Microorganisms

Systems Biology data, models and SOPs



BioVeL

Data, Service and Workflows



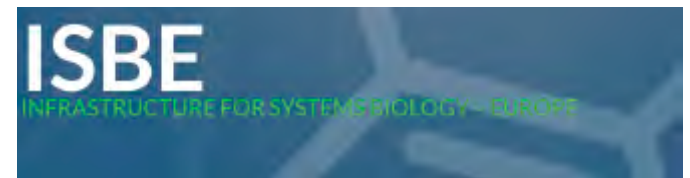
Research Objects

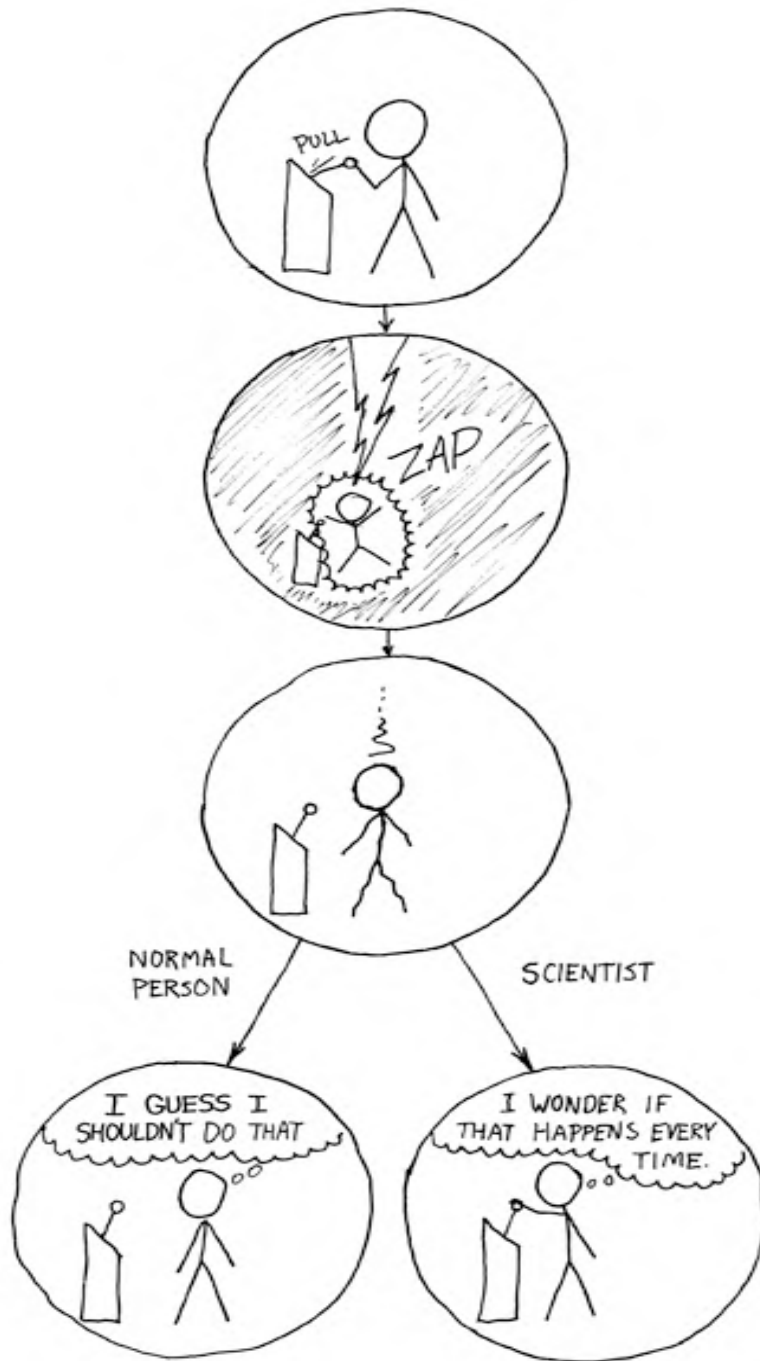


Open PHACTS

Open Pharmacological Space

Nanopublications





Reproducibility

a principle of the scientific method

Evidence to test and justify claims

Comparison of results and methods

Peer review

“An experiment is reproducible until another laboratory tries to repeat it.”

Alexander Kohn

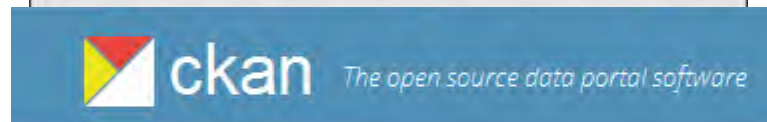
The Reproducibility Initiative

Reproducibility as a Service

PLoS, FigShare

<http://reproducibilityinitiative.org>

Data Journals / Repositories



In silico (Computational) Science

Simulations, data exploration, data processing, analytics, database based, text mining, auto recommendation, visual analytics...(Digital Science = Science)

Datasets
Data collections
Algorithms
Configurations
Tools and Apps
Codes
Workflows
Scripts
Code Libraries
Services,
Infrastructure,
Compilers
Hardware

RESEARCH PRIORITIES

Shining Light into Black Boxes

A. Morin,¹ J. Urban,² P. D. Adams,³ I. Foster,⁴ A. Sali,⁵ D. Baker,⁶ P. Sliz^{1*}

Funders, publishers, and research institutions must act to ensure that research computer code is made widely available.

Science 13 April 2012: 336(6078) 159-160

DOI: 10.1126/science.1218263



**executable paper
grand challenge**

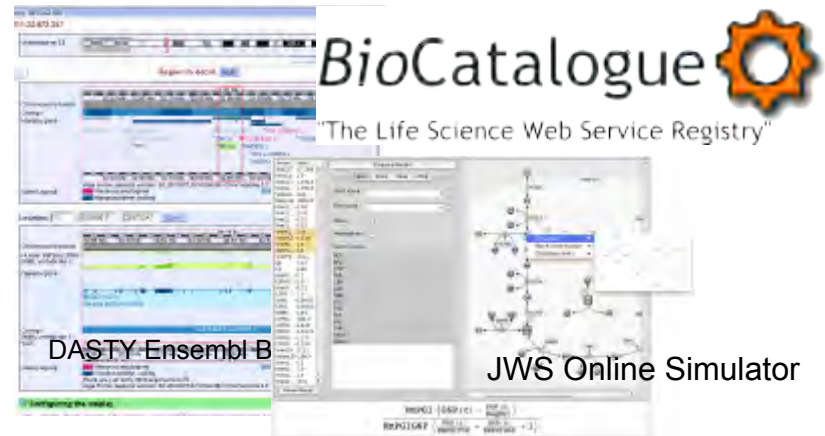
knowledge enhancement in the computational sciences



Specialist Codes Libraries, Platforms, Tools



Service based Science



Data Collections Catalogues



My Data **My Process**
My Codes **My Libraries**
My Special Tweaks

github
 SOCIAL CODING
 Software
 Repositories

Commodity Platforms



(Cloud) Hosted Services



Compound Assemblies: Workflows

See *Tom Moritz talk*

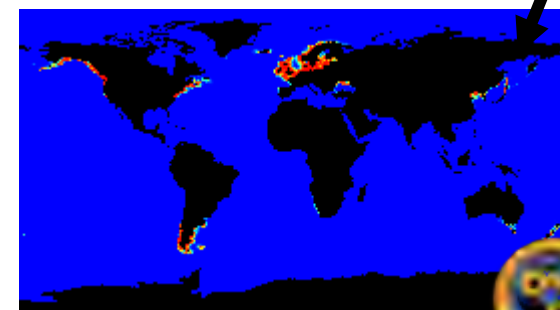
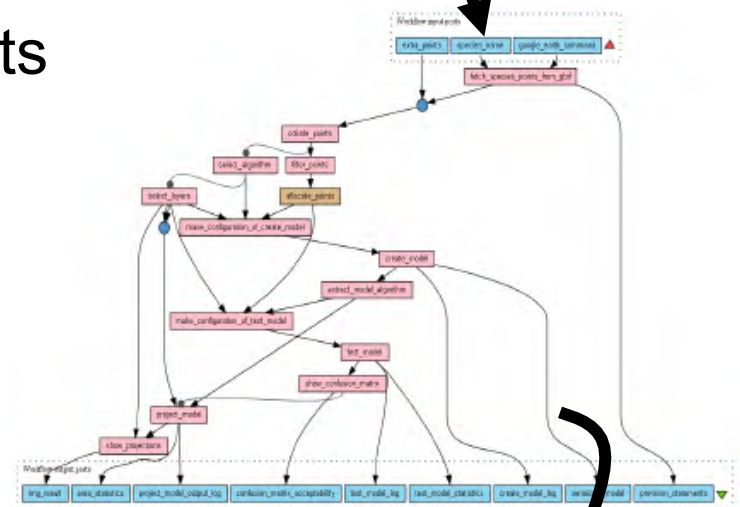
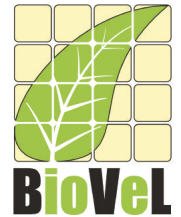
Execution

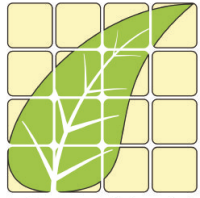
- Multi-step coordinated execution of (distributed) computational components
- Repeatable and comparative
- Explicated computation

Virtual Witnessing / Minute-Taking

- Transparent, precise, citable documentation
- Accurate logs
- Reusable protocols, know-how, best practice

nameComplete
Ameira divagans
Boccardia redeki
Bougainvillia rugosa
Branchiura sowerbyi
Cercopagis pengoi
Chelicorophium curvispinum
Chionoecetes opilio

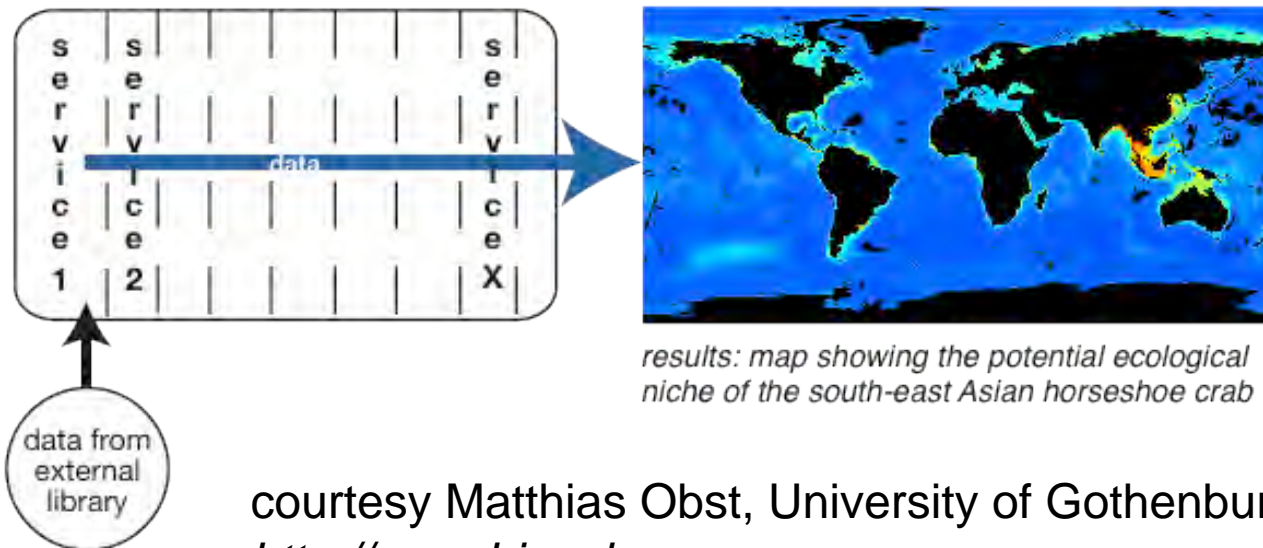




BioVeL

Study on the ecological niche of the south east Asian horseshoe crab

- Generate input files: Import south east Asian data from public archives + Clean data + Merge with own data
- Run large number of niche model analyses
- Visualise ecological niche maps to interpret and compare



results: map showing the potential ecological niche of the south-east Asian horseshoe crab

courtesy Matthias Obst, University of Gothenburg, Sweden
<http://www.biovel.eu>

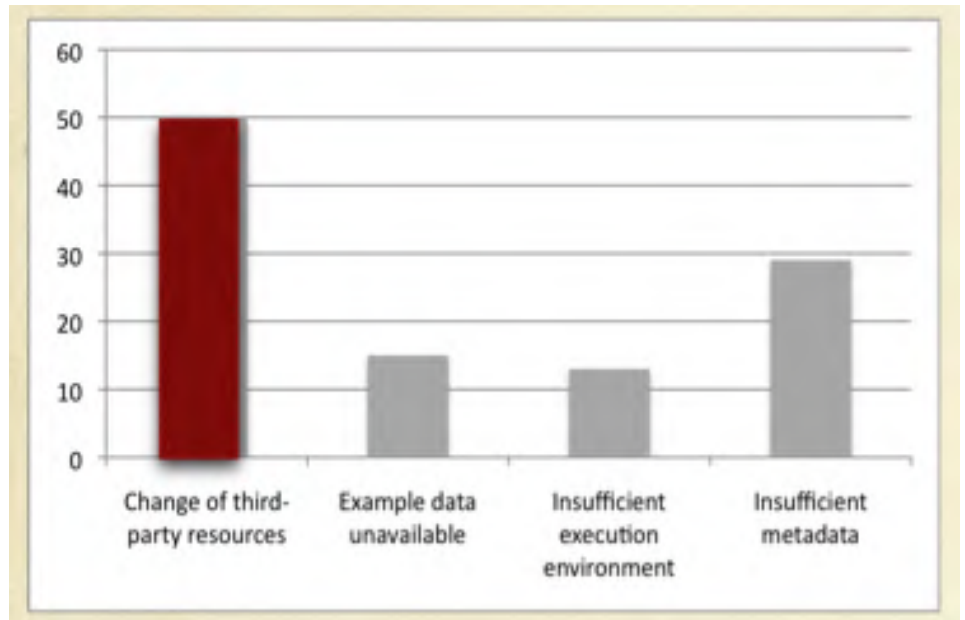
Reproducibility Issues

Read It: Description

- Obfuscated: too vague / detailed
- Black Box data/processes
- Tweaking
- Scattering
- Logging

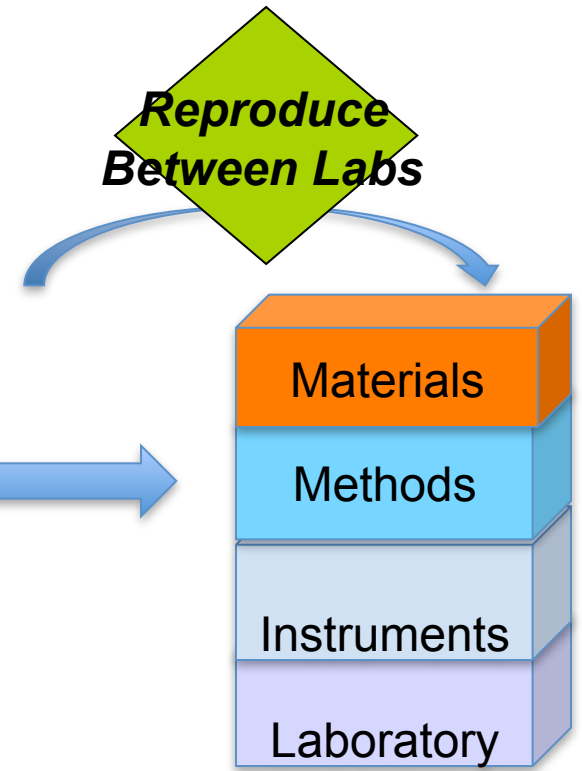
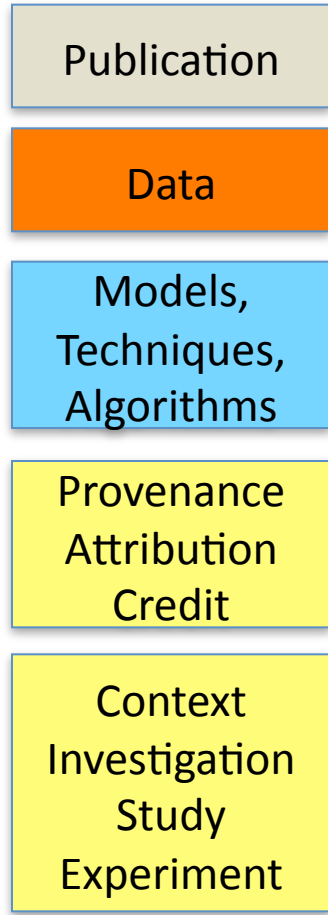
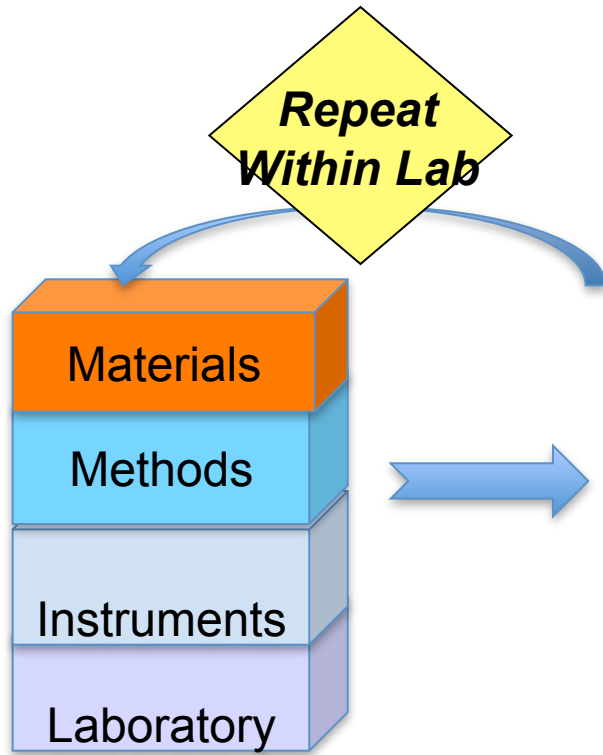
Run it: Environment

- Dependencies/Stewardship
- Stability/Reliability
- Availability: one-off processes
- Black box platforms
- Scattering
- Tweaking
- State: Snapshot or Live



Do It: Governance

- Capability
- Cost / Burden
- Credit / Reward



Replicate / Repeat
 Exactly replicate the original experiment and experimental conditions. Eliminate change. Observe.

Reproduce
 Run experiment with differences in experimental conditions.. Compare to test for same result. Observe.

Capture Curate Discover Use Reuse Preserve

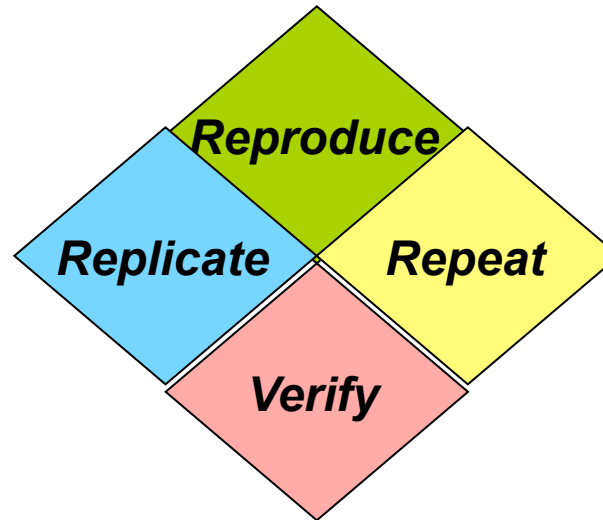
Re* <verb> Bingo

Vary and compare

Recreate results without existing code or data, independently.

Fix and Compare

Regenerate results from existing code, data.



Re-run to determine the sensitivity of results when underlying measurements are retaken

Review the Record

(Re)examine accuracy, wrt underlying model (Verify), or data (model error, measurement error) (Validate)

Detect and Repair

Prevent

Virtual Machines, Deployed Codes

Community Workflows

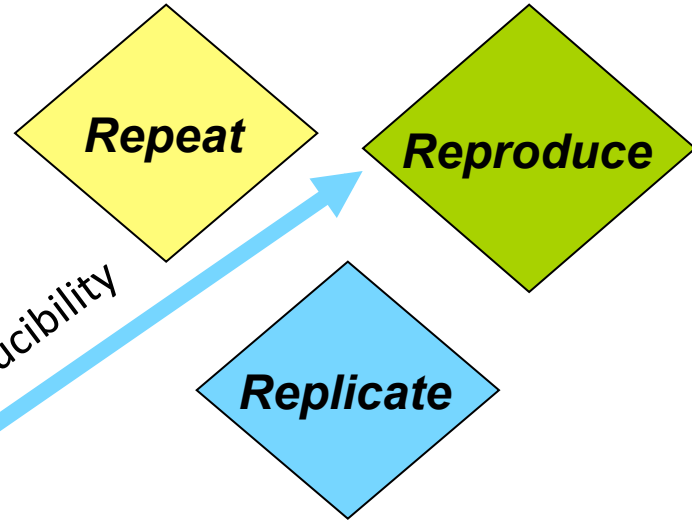
Version Control
(data, services, workflows)

Dependencies
Snapshots

No Version Control

Workflows in the Wild

External Dependencies
Mixed Environments
(open service set)



[adapted from Watson and Missier]

Reproducible Research Systems

There are many emerging (time for “standards”?)

- **ID it to Cite It:** ORCID (people), DOI (data, models, tools ...)
- **Tracking:** local helper systems to instrument and track provenance
- **Science as a Service:** Virtual Machines, Cloud Appliances, Hosted platforms deploys on your behalf, no installations, common platforms (e.g. Galaxy)
- **Libraries and Repositories:** with rich documentation
- **Publish:** executable papers, companion web sites, embedded notebooks/publishing, active publications
- **Explication of experimental mechanics:** pipelines, workflows, script systems with documentation, common tools/languages (e.g. MatLab)

Home

Users

Groups

Workflows

Files

Packs

Services

Topics

Services

All

Home > Packs > Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Pack: Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Created at: 24/11/11 @ 19:12:05 Last updated: 01/12/11 @ 22:13:34

Tags (6) | Featured in Packs (0) | Favourited By (0) | Comments (0)

New/Upload

Workflow

Log in / Register

Title: Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Description

The scientific experiment represented by this research object pertains to the wavelength study for a sample of the most isolated galaxies in the local universe. This study characterizes each galaxy of this sample through both the measurement of basic astrophysical properties:

- The equatorial coordinates in J2000 epoch
- The velocities in km/s (v)
- The dust extinction coefficient (ag)
- The axis ratio of the isophote 25 mag/arcsec² (logr25)
- The apparent total B magnitude (BT)
- The morphological type (t)

and the calculation of the more complex properties:

- The distance in Mega parsecs (D)
- The corrected apparent B magnitude (btc)
- The optical luminosity in B-band (LB)

Specifically, this research object is focused on the calculation of the intrinsic luminosity in the Johnson B-band, in order to achieve it the measurement and calculation of all those astrophysical properties is needed.



- Workflow: Propagation of physical quantities in the calculation of luminosities of galaxies** (Susana)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:38:06)

[more](#)
- Workflow: Calculation of distances, magnitudes and luminosities using HyperLEDA** (Susana)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:37:54)

[more](#)
- File: Content description of RO Propagation of quantities** (Jose Enrique Ruiz)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:33:19)

[more](#)
- File: How to use RO Propagation of quantities** (Jose Enrique Ruiz)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:33:10)

[more](#)
- File: Purpose of RO Propagation of quantities** (Jose Enrique Ruiz)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:30:33)

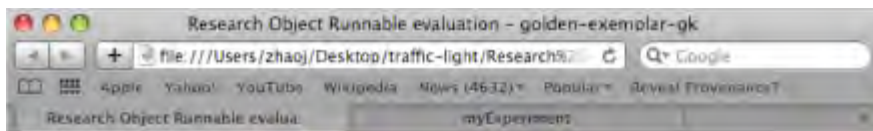
[more](#)
- File: Data involved in RO Propagation of quantities** (Jose Enrique Ruiz)

Added by [Jose Enrique Ruiz](#) ... 192 days ago (01/12/11 @ 13:30:20)

[more](#)



Resource



Research object **golden-exemplar-gk**

Target **golden-exemplar-gk** *minimally satisfies* checklist for **Runnable**.

All declared workflow inputs are accessible

Workflow description http://ns.taverna.org.uk/2010/workflowBundle/c4a53ec6-d774-4b4e-b60f-2899dea1dee7/workflow/Filter_concepts_with/is_not_accessible

Workflow description found

All declared workflow inputs are aggregated

All declared workflow descriptions are aggregated

Workflow input(s) found

[Wf4Ever project](#)

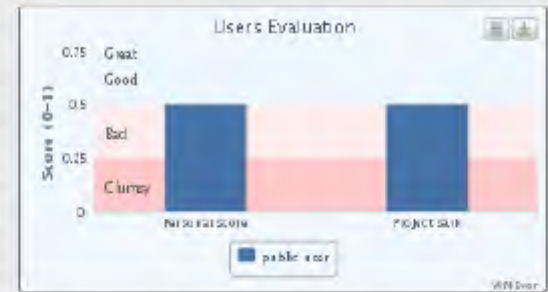
Display a menu



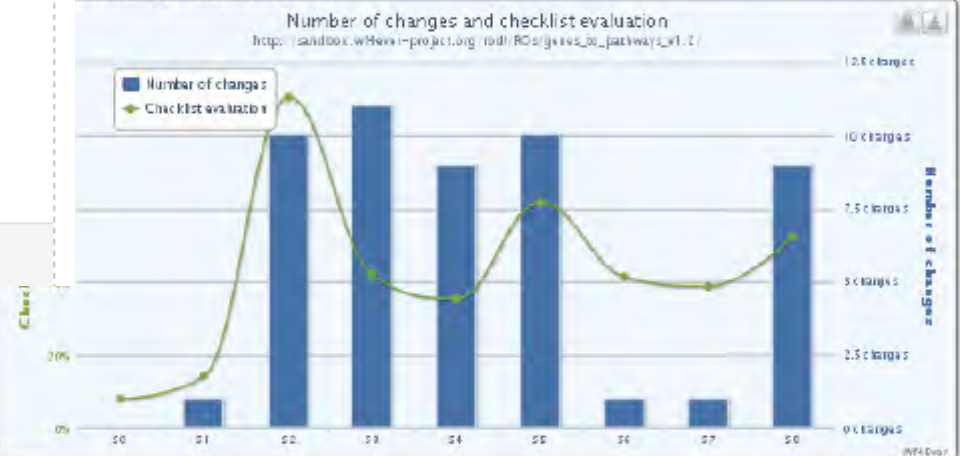
URI: http://sandbox.wf4ever-project.org/rodi/ROs/genes_to_pathways_v1.1/

... you are able to analyze information ... evolution and evaluation of a RO. ... right of the screen you have user ... based on the quality of their ... One of the evaluations represents the ... of the user and the other represents the ... of the user in the RO. ... stem chart compares the number of ... against quality for each snapshot, which ... clicked in order to get additional ...

... the Stability Value represents how stable ... in the quality of the RO over time.



Stability Value: 53.70%



Snapshot checklist evaluation

http://sandbox.wf4ever-project.org/rodi/ROs/genes_to_pathways_v1.2_snapshot_0/

✗ MUST:

- ✓ Must check number 1
- ✗ Must check number 2
- ✓ Must check number 3

✓ SHOULD:

- ✓ Should check number 1
- ✓ Should check number 2
- ✓ Should check number 3

✗ MAY:

- ✓ May check number 1
- ✓ May check number 2
- ✗ May check number 3
- ✗ May check number 4
- ✓ May check number 5



Changes info:

Additions:

- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)
- [Pathway.DOC.txt](#)

1. Reproducibility is a means to an end, not an end in itself

- all science becomes less reproducible / repeatable over time...

and some can never be

... stochastic experiments or large scale data collections.

- when does it matter?



Results may vary

Defend results are correct and method convincing and repeatable.

Review & Learn Verify the results empirically. Trust. Understand. Convince, comfort, credibility.

Reuse Use the explained and trusted results (data, method) for new / my science on demand. Compare. Extend.

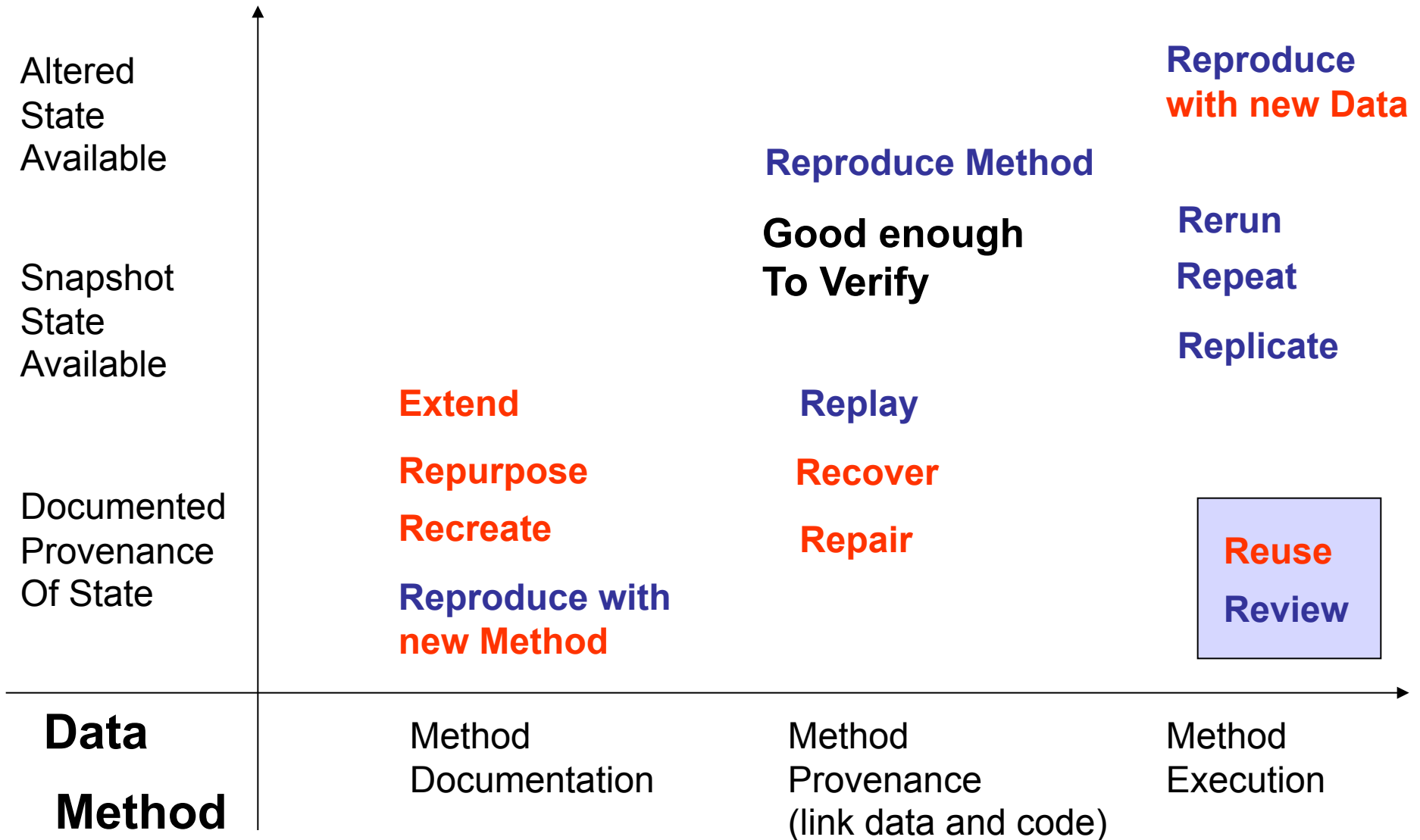
Is it “true”?

Can I repeat it?

Can I use it?

Can I reproduce it?

2. Reproducibility is a Spectrum



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online

Peng RD, Reproducible Research in Computational Science *Science* 2 Dec 2011: 1226-1227.

De Roure <http://www.scilogos.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/>

Reproducibility is a Spectrum

Partial reproducibility – over proprietary steps or difficult-to-reproduce subparts, or just through examining the log

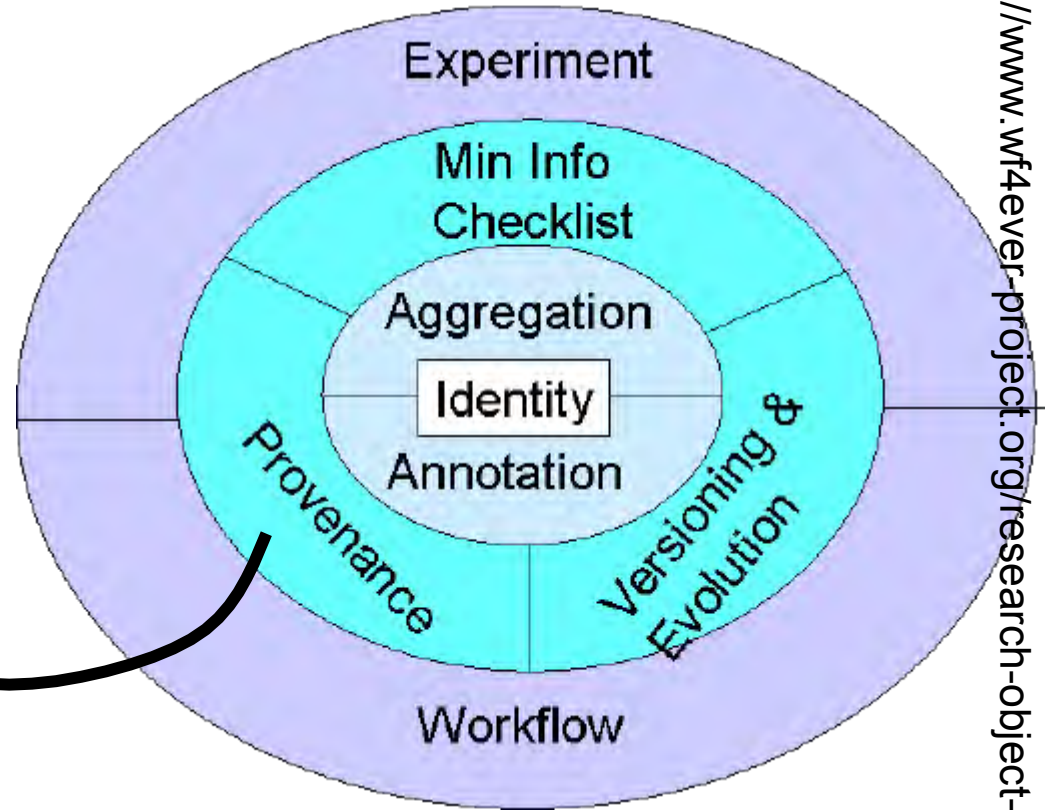
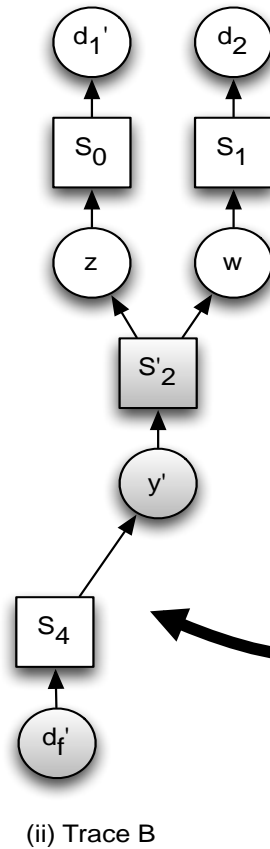
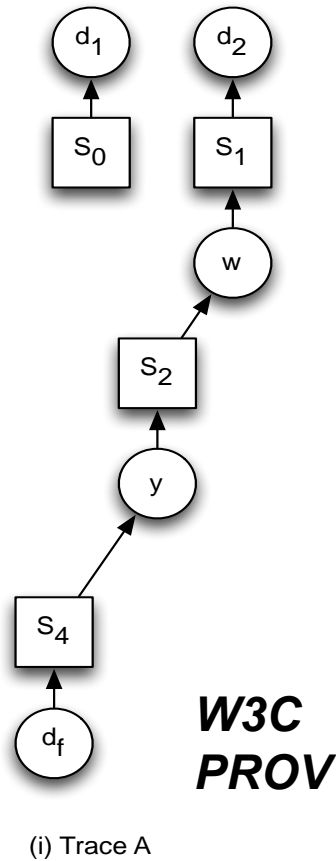
“perfect is the enemy of the good”
Voltaire

3. Reproducibility through Inspection

Archived Record to Manage



[Woodman, et al, 2011]



http://www.wf4ever-project.org/research-object-model

Log, Fix, Replay, Analyse -> Instrument Systems and Apps

4. Reproducibility by Invocation

Active Instrument to Maintain

- **Active Preservation:**
 - Preservation vs Just in Time Just Enough restoration/reconstruction: The natural state is broken.
- **Stop Publishing, Start Releasing**
 - Software release practices for workflows and scripts, services, data and articles [Schopf, JC DL 2012]
- **Librarianship, Stewardship and Best Practices of Everything**
 - “Better Science through Superior Software” – C Titus Brown
 - Zeeya Merali , Nature 467, 775-777 (2010) | doi: 10.1038/467775a



~~Software~~

Data Stewardship

“Better Science through Superior Software” – C Titus Brown
Open does not mean understandable.

Software sustainability
Software practices
Software deposition
Long term access to software
Credit for software
Software Journals
Licensing
Open Source Software



www.software.ac.uk



Best Practices for Scientific Computing <http://arxiv.org/abs/1210.0530>

Stodden, Reproducible Research Standard, *Intl J Comm Law & Policy*, 13 2009

Prlić A, Procter JB (2012) Ten Simple Rules for the Open Development of Scientific Software. *PLoS Comput Biol* 8(12): e1002802. doi:10.1371/journal.pcbi.1002802

5. Governance, Economics and Burden

Why?

Make it Matter. Trade, Asset and Curation economics

What?

Numerous standards: formats, terminologies and checklists

When?

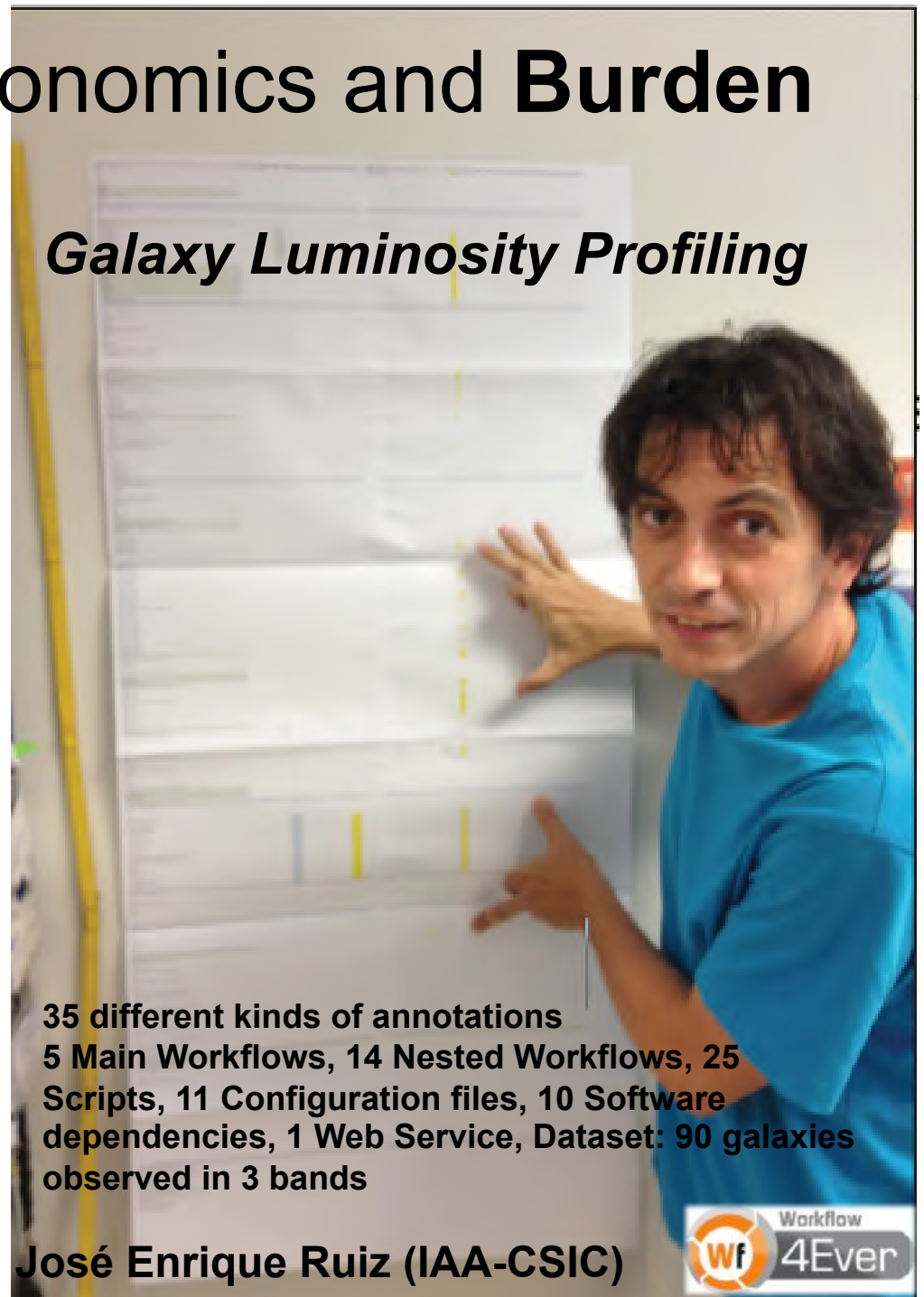
Incremental, Eager and Lazy, UpStream, Downstream

How?

Ramps: Automation & Integrated Tools

Who?


Copy editing Method, Curation Service, Authors? Reviewers? Editors? Readers? Curators?



Galaxy Luminosity Profiling

35 different kinds of annotations
5 Main Workflows, 14 Nested Workflows, 25 Scripts, 11 Configuration files, 10 Software dependencies, 1 Web Service, Dataset: 90 galaxies observed in 3 bands

José Enrique Ruiz (IAA-CSIC)

 Workflow 4Ever

Accessible

Reusable

Capable

Publication



Data



INSTRUMENTS
Samples, Specimens
Strains

Models,
Techniques,
Algorithms



Context
Investigation
Study
Assay



74% / 26%

31% / 8%

Provenance
Attribution
Credit



See Fran Berman Talk

ISB International Society
for Biocuration

<http://biocurator.org/>

NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY |

My data are your data

Vivien Marx

Nature Biotechnology 30, 509–511 (2012) | doi:10.1038/nl
Published online 07 June 2012

Encouraging more broad and inclusive data sharing in
community efforts to overcome technical barriers and t

- Introduction

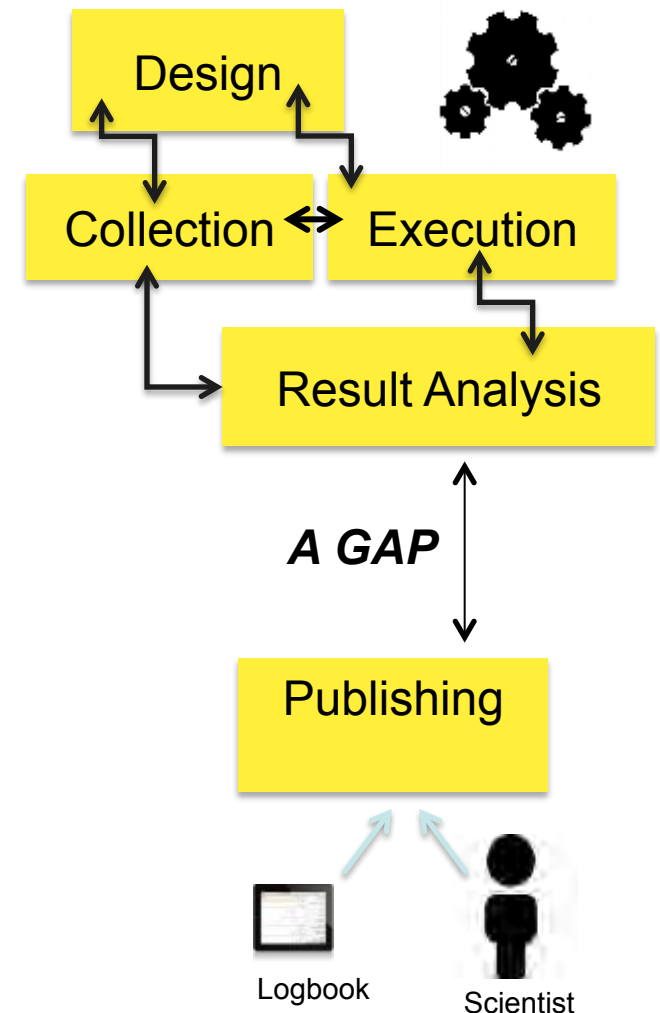
[Introduction](#) • [References](#) • [Supplementary Information](#)

Hugging
Flirting
Voyerism
Creeping
Comprehending

Trading
Credit
Economics

Integrated Reproducible Research Systems that the 95% use

- **Active Reproducible Research Environment**
 - Instrumented infrastructure and services for producing and working with reproducible research.
- **Active Reproducible Research Publication Environment**
 - Instrumented infrastructure and services for distributing and reviewing; academic credit; legal licensing, watching and preserving etc.
- **Safe Havens, Rescue Teams, Scholarship Services**
- **Top Down and Bottom Up**



* Adapted from Mesirov, J. Accessible Reproducible Research *Science* 327(5964), 415-416 (2010)

Utopia Documents – Two-step and one-step secretion mechanisms in Gram-negative bacteria: contrasting the type IV secretion system and the chaperone-usher pathway of pilus biogenesis

Documents Information: **BJ**

Two-step and one-step secretion mechanisms in Gram-negative bacteria: contrasting the type IV secretion system and the chaperone-usher pathway of pilus biogenesis

Régo, Ana Toste, Chandran, Vidya, Waksman, Gabriel

Biochem. J. 425 (Pt 3)
Pages: 475-488
Published: 2010-02-01

Keywords
chaperone-usher; protein secretion; structural biology; type IV secretion; virulence factor

Abbreviations used in the Document
EM: electron microscopy; AAD: all- α -helical domain; O-layer: outer layer; T1SS: T2SS etc., type I secretion system, type II secretion system etc.; I-layer: inner layer; TM: transmembrane; DSE: donor-strand exchange; Nte: N-terminal extension; DSC: donor-strand complementation; CU: chaperone-usher; NBD: nucleotide-binding domain.

UTopia Documents
www.getutopia.com

Ramps for Authoring & Reading

<http://www.rightfield.org.uk>

IDFExcelExample2_lem.xls

Element	Example	Format	Constraint	Types in dropdown	Automatic Key Type	Key Type	Explanation
1 Investigation Title	Transcription profiling of wild-type and ATFS-	text	NOT NULL				
5 Experimental Design Type *	Transcription profiling of wild-type and ATFS-	text	NOT NULL			PK	Explanation
6 Experimental Factor Category *	compound_treatment_design	text	NOT NULL			CV	PK - primary key
7 Person Last Name	OSMPOUND	text	NOT NULL			CV	FK - foreign key
8 Person First Name	Maxwell	text	NOT NULL				
9 Person Mid Initial	Diaz	text	NOT NULL				
10 Person Affiliation	Institute for Systems Biology	text	NOT NULL				
11 Role *	submitter	text	NOT NULL				
12 Quality Control Description Type *		text	NOT NULL				
13 Replicate Description Type *		text	NOT NULL		X	CV	
14 Normalization Description type *		text	NOT NULL		X	CV	
15		text	NOT NULL		X	CV	
16 Experiment Description	transcriptome-wide analysis of wild-type and ATFS-1 in mouse bone marrow macrophages were carried out to investigate the transcriptional network in Toll-like receptor	text		X			
17 Protocol Name		text					
19 Protocol Type *		text				CV	
20 Protocol Description	osmpon L1	text					
21 Protocol Parameters		integer	NOT NULL			X	
22		text	NOT NULL		X		
23 SRRF File		text	NOT NULL			X	
24 data file		text	NOT NULL				
25		text	NOT NULL	X			
26		integer	NOT NULL		X	X	
27 ADP file		text	NOT NULL				
28 Term Source Name		integer	NOT NULL		X	X	
29 Term Source File		text	NOT NULL		X	X	
30 Term Source Version		integer	NOT NULL		X	X	
31		integer	NOT NULL		X	X	
32		text					
33		integer			X	X	
34		integer			X	X	
35		text			X	X	
36		text			X	X	
37		integer			X	X	
38		integer			X	X	
39		integer			X	X	
40		integer			X	X	
41		text			X	X	
42		text			X	X	
43		text			X	X	
44		integer			X	X	
45		integer			X	X	

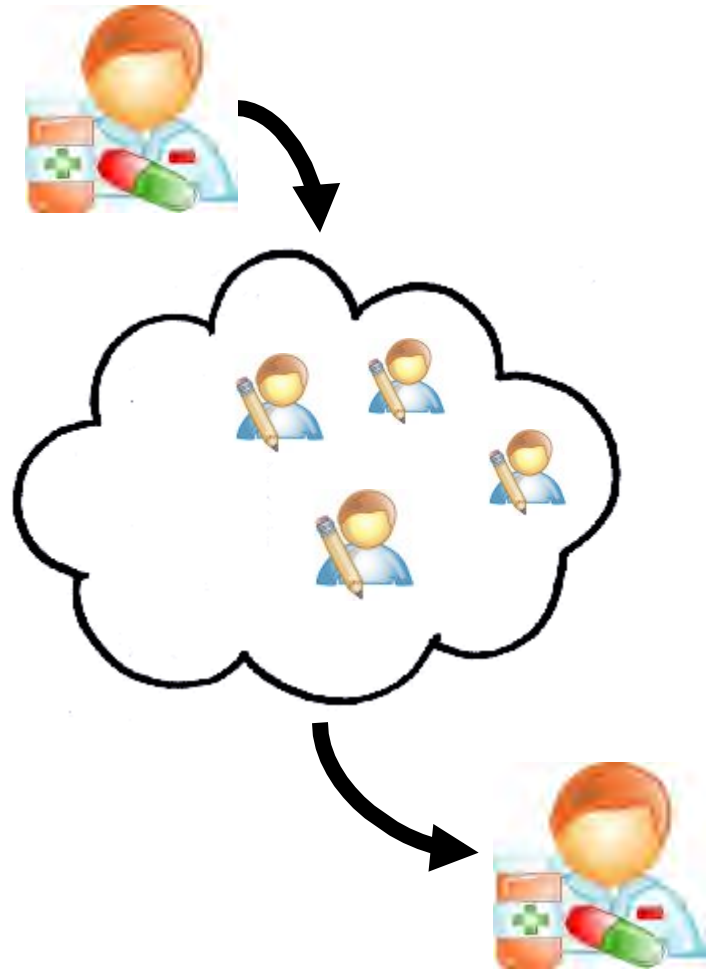
RightField

Semantics, Thursday, 10:30–12:00

Wolstencroft, Owen, Goble, Nguyen, Krebs and Müller. RightField: Semantic Enrichment of Systems Biology Data using Spreadsheets

Governance, Ecosystems and the Scholarly Process

- Local or Central responsibility
- Responsibility/Role/Reward of:
 - Institution? Funders?
 - Library? Publishers?
 - Reviewers? Trainees?
 - Authors? Readers?
 - Communities? Curators?
 - Information Brokers?
 - Third party vendors?
 - Research Management service providers?
- Cost/Capacity/Reward for review
- Sustainability, Silos, Packaging
- The 95%



“An experiment is reproducible until another laboratory tries to repeat it.”

Alexander Kohn

**Its harder than you might think.
And less common than it could be.**

Its about capturing, preserving,
reusing and curating.

Bottom Up Perspective

Summary

- Couple together Library, Infrastructure, Publishing, Culture, Social, Policy
- Reproducibility for the 95%
- Bottom up not just top down
- “Weak” reproducibility is better than none at all and could be enough.

Archived Record



Documentation, enough information to make a judgement call...

Inspection

Active Instrument



... and reproduce the workflow if needed

Invocation



[Events](#) >

Beyond the PDF 2

Date: Tuesday, March 19, 2013 to Wednesday, March 20, 2013

Location: Amsterdam, NL

Go directly to: [Registration](#) | [Location](#) | [Transportation](#) | [Hotels](#) | [Amsterdam Guide](#) | [Committee](#) | [Sponsors](#) | [Preliminary Program](#)

Conference Registration*

On-line registration is available at this website: <http://www.regonline.com/beyondthepdf2>

Regular Registration Fee € 150

Student Registration Fee € 70

Scholarly communication across all disciplines is changing profoundly under the influence of new technologies. New models, tools and standards are being developed that aim to enhance, enable or entirely replace formerly ingrained forms of scholarly communication, including publication courses, conferences and policy. The **Beyond the PDF** conference brings together scholars, librarians, archivists, publishers and research funders in a lively forum, not just to broaden awareness of current efforts across disciplines, but to define the future through discussions, challenge projects, demonstrations and seeding new partnerships and collaborations. Individually and collectively, we aim to bring about a change in modern scholarly communications through the effective use of information technology. Beyond the PDF is organized by [FORCE11](#), a group of stakeholders that arose organically from the first [Beyond the PDF](#) workshop, held at the University of California, San Diego, in 2011, and a follow on workshop held at [Dagstuhl](#) that same year. We will actively engage the membership of FORCE11 to shape and evolve this and future workshops, as the conference itself provides a platform for those interested in creating new modes of conference organization and participation. FORCE11 and the Beyond the PDF conference are supported by a grant from the Alfred P. Sloan Foundation. [Sponsorships](#) are available. If you are interested, please contact Maryann Martone: [mmartone \(you know what goes here!\)@ucsd.edu](mailto:mmartone@ucsd.edu)

- **Dates:** Main conference: March 19-20th.
 - March 19, 2012 - 09:30 - 17:00 and an evening social event
 - March 20, 2012 - 09:30 - 17:00

Acknowledgements and Inspirations

- David De Roure
 - Tim Clark
 - Sean Bechhofer
 - Robert Stevens
 - Christine Borgman
 - Victoria Stodden
 - Marco Roos
 - Jose Enrique Ruiz del Mazo
 - Oscar Corcho
 - Anton Güntsch
 - Cherian Mathew
 - Ian Cottam
 - Steve Pettifer
 - Robin Williams
 - Pinar Alper
 - C. Titus Brown
 - Greg Wilson
 - Juliana Freire
 - Jill Mesirov
 - Simon Cockell
 - Paolo Missier
 - Paul Watson
 - Gerhard Klimeck
 - Matthias Obst
 - Jun Zhao
 - Pinar Alper
 - Daniel Garijo
 - Yolanda Gil
-
- Wf4ever, SysMO, BioVel, UTOPIA and myGrid teams

Further Information

- myGrid
 - <http://www.mygrid.org.uk>
- Taverna
 - <http://www.taverna.org.uk>
- myExperiment
 - <http://www.myexperiment.org>
- BioCatalogue
 - <http://www.biocatalogue.org>
- SysMO-SEEK
 - <http://www.sysmo-db.org>
- MethodBox
 - <http://www.methodbox.org.uk>
- Rightfield
 - <http://www.rightfield.org.uk>
- UTOPIA Documents
 - <http://www.getutopia.com>
- Wf4ever
 - <http://www.wf4ever-project.org>
- Software Sustainability Institute
 - <http://www.software.ac.uk>
- BioVeL
 - <http://www.biovel.eu>
- Force11
 - <http://www.force11.org>
- <http://reproducibilityinitiative.org>
- <http://reproducibleresearch.net>

