Provenance as a key material to improve FAIR-ness in data-driven (life) sciences

Alban Gaignard, PhD, CNRS

### Colloque "Réplicabilité et Reproductibilité de la Recherche" MITI, CNRS Paris, 08 septembre 2023







Predict the evolution ? Understand the pathology ?

**Prevention strategies ?** 

### ICAN cohort: 34 univ. hospitals / 3000 subjects

3.000 ToF MRIs600 whole genomes (analysis in progress)

Predict the evolution ? Understand the pathology ?

**Prevention strategies ?** 

Bourcier R, Chatel S, Bourcereau E, Jouan S, Marec HL, Daumas-Duport B, Sevin-Allouet M, Guillon B, Roualdes V, Riem T, Isidor B, Lebranchu P, Connault J, Tourneau TL, Gaignard A, Loirand G, Redon R, Desal H; ICAN Investigators. Understanding the Pathophysiology of Intracranial Aneurysm: The ICAN Project. Neurosurgery. 2017 Apr 1;80(4):621-626. <u>https://doi.org/10.1093/neuros/nyw135</u>



**Workflows** to enhance trust in scientific results :

- → automation (scalability)
- → abstraction (methods sharing)
- → traceability (provenance)

# nextflow

Galaxv

COMMUNITY HUE









	ŷ1 branch	© 0 releases		42 contributors	
h: master + New pull request		Create new file	Upload files	Find file	Clone or download
ditommaso Update README.md				Latest comm	iit 07ee96a 25 days ag
NTRIBUTING.md	Added	contributing			4 years ag
ADME.md	Update	README.md			25 days ag
ADME.md					
Awesome Pipeline	•				
A curated list of awesome pipeline to	polkits inspired by Aweso	me Sysadmin			
Dinalina francusales 9 li	hunding				
Pipeline frameworks & li	braries				
ActionChain - A workflow system	m for simple linear succe	ss/failure workflows.			
Adage - Small package to description	ribe workflows that are no	ot completely known at de	finition time.		
<ul> <li>Airflow - Python-based workflo</li> </ul>	w system created by AirB	3nb.			
Anduril - Component-based wo	rkflow framework for scie	entific data analysis.			
Antha - High-level language for	biology.				
• Bds - Scripting language for dat	ta pipelines.				
BioMake - GNU-Make-like utility	v for managing builds and	d complex workflows.			
BioQueue - Explicit framework v	with web monitoring and	resource estimation.			
BioQueue - Explicit framework v     Bistro - Library to build and exe	with web monitoring and cute typed scientific work	resource estimation. kflows			
BioQueue - Explicit framework v     Bistro - Library to build and exe     Boine - Tool for running and ma	with web monitoring and cute typed scientific work	resource estimation. kflows.			
BioQueue - Explicit framework v     Bistro - Library to build and exe     Bpipe - Tool for running and ma     Dictory to build and exe	with web monitoring and cute typed scientific work naging bioinformatics pip	resource estimation. kflows. pelines.			
BioQueue - Explicit framework v     Bistro - Library to build and exe     Bpipe - Tool for running and ma     Briefly - Python Meta-program	with web monitoring and cute typed scientific worl naging bioinformatics pip ning Library for Job Flow	resource estimation. kflows. pelines. Control.			
BioQueue - Explicit framework v     Bistro - Library to build and exe     Bpipe - Tool for running and ma     Briefly - Python Meta-programm     Cluster Flow - Command-line te	with web monitoring and cute typed scientific worl naging bioinformatics pip ming Library for Job Flow wol which uses common c	resource estimation. kflows. belines. Control. cluster managers to run bio	informatics	pipelines.	
BioQueue - Explicit framework of Bistro - Library to build and exe Bpipe - Tool for running and ma Briefly - Python Meta-program Cluster Flow - Command-line to Clusterjob - Automated reprodu	with web monitoring and cute typed scientific work naging bioinformatics pip ming Library for Job Flow sol which uses common o cibility, and hassle-free s	resource estimation. kflows. belines. Control. Juster managers to run bio submission of computation	informatics al jobs to clu	pipelines. Isters.	
BioQueue - Explicit framework vi Bistro - Library to build and exe Bipipe - Tool for running and ma Briefly - Python Meta-program Cluster Flow - Command-line to Cluster Flow - Command-line to Clusterjob - Automated reprodu Compss - Programming model	with web monitoring and cute typed scientific worf naging bioinformatics pip ning Library for Job Flow ool which uses common c icibility, and hassle-free s for distributed infrastruct	resource estimation. kflows. belines. Control. cluster managers to run bio submission of computation rures.	informatics al jobs to clu	pipelines. Isters.	
BioQueue - Explicit framework vi Bistro - Library to build and exe Spipe - Tool for running and ma Briefly - Python Meta-program Cluster Flow - Command-Ine to Clusterjob - Automated reprodu Compas - Programming model <sup>1</sup> Connar2 - Light-weight workflow	with web monitoring and cute typed scientific worf naging bioinformatics pip ning Library for Job Flow ool which uses common c cibility, and hassle-free e for distributed infrastruct v management applicatio	resource estimation. kflows. pelines. Control. cluster managers to run bio submission of computation ures. n.	informatics al jobs to clu	pipelines. Isters.	
BioQueue – Explicit framework vi Bistro – Library to build and exe Bejpe – Tool for running and ma Briefly – Python Meta-program Cluster Flow – Command-line to Cluster Flow – Command-line to Cluster Flow – Command-line to Cluster Flow – Command-line to Cluster Compase – Programming model Conan2 – Light-weight workflow Consecution – A Python pipelin	with web monitoring and cute typed scientific worf naging bioinformatics pip ning Library for Job Flow sol which uses common c cibility, and hassle-free s for distributed infrastruct v management applicatio a abstraction inspired by	resource estimation. kflows. pelines. Control. cluster managers to run bio submission of computation rures. n. Apache Storm topologies.	informatics al jobs to clu	pipelines. Isters.	
BioQueue - Explicit framework u     Bistor - Library to build and exe     Bigter - Tool for running and ma     Briefly - Python Meta-program     Cluster Row - Command-line t     Clusterjob - Automated reprodu     Compas - Programming model     Conanc2 - Light-weight workflow     Consecution - A Python pipelin     Cosenso: Python library for ma	with web monitoring and icute typed scientific word naging bioinformatics pip iming Library for Job Flow bol which uses common c icibility, and hassle-free s for distributed infrastruct v management applicatio abstraction inspired by ssively parallel workflows	resource estimation. kflows. Selines. Control. Juster managers to run bio submission of computation ures. n. Apache Storm topologies. L	informatics al jobs to clu	pipelines. Isters.	
BioQueue - Explicit framework:     Bistro - Library to build and exe     Bipte - Tool for running and ma     Briefly - Python Meta-program     Cluster Flow - Command-line to     Clusters Flow - Command-line to     Clusters (and the comparised of the comparise of the comparise - th	with web monitoring and cute typed scientific worl naging bioinformatics pip ning Library for Job Flow olo which uses common c scibility, and hassle-free s for distributed infrastruct v management applicatio a abstraction inspired by saively parallel workflows ent System geared towar	resource estimation. kflows. belines. Control. Juster managers to run bio aubmission of computatior ures. n. Apache Storm topologies. i.	informatics   al jobs to clu m the Broad	pipelines. ısters. Institute.	
BioQueue - Explicit framework:     Biatro - Library to build and exe     Bippe - Tool for running and me     Briefly - Python Meta-program     Cluster/bb - Automated reprod.     Compas - Programming model     Connard - Light-weight workflov     Connea: - Light-weight workflov     Connea: - Light-weight workflow     Connea: - Python Iberary for ma     Cromwell - Workflow Managem     Cunderform - Advanced function	with web monitoring and icute typed scientific word ming Library for Job Flow ool which uses common c icibility, and hassle-free § for distributed infrastruct remanagement applicatio e abstraction inspired by ssively parallel workflows ent System geared towar al workflow language and al workflow language and spiret	resource estimation. kflows. selines. Control. sluster managers to run bio submission of computation ures. n. Apache Storm topologies. k. ds scientific workflows froo framework, implemented	informatics   al jobs to clu m the Broad in Erlang.	pipelines. Isters. Institute.	
BioCouse – Explicit framework:     Bistro - Library to build and exe     Bistro - Library to build and exe     Single - Tool for running and ma     Briefly – Python Meta-program     Cluster Floor - Command-line to     Cluster Floor - Command-line to     Cluster Sloor - Command-line to     Consos - Python Ikary for am     Consos - Python Ikary for am     Consone - Dython Ik	with web monitoring and cute typed scientific worl naging bioinformatics pip ning Library for Job Flow old which uses common c (cibility, and hassel-ree = for distributed infrastruct management applicatio a sistivety parallel workflows ent System geared towar- all workflow language and ob scheduler in Python.	resource estimation. kflows. Defines. Control. Juster managers to run bis submission of computation ures. n. Apache Storm topologies. k scientific workflows fro d scientific workflows fro	informatics   al jobs to clu m the Broad in Erlang.	pipelines. ısters. Institute.	
BioQueue - Explicit framework:     Biatro - Library to build and exe     Bippe - Tool for running and ma     Briefly - Python Meta-program     Cluster/bb- Automated reprod.     Comps - Porgamming model     Consecution - A Python pipelin     Commod - Workflow Managem     Cuneiform - Advanced function     Dagr - A scala based DSL and + Dagr	with web monitoring and cute typed scientific worf anging bioinformatics pip ming Library for Job Flow ol which uses common c scibility, and hassle-free s for distributed infrastruct a abstraction inspired by ssively parallel workflows sisvely parallel workflows all workflow language and ob scheduler in Python.	resource estimation. kflows. Selines. Control. sluster managers to run bic sluster managers to run bic ubmission of computation ures. Apache Storm topologies. d scientific workflows fro § framework, implemented executing bioinformatics.	informatics ; al jobs to clu m the Broad in Erlang. oipelines as l	pipelines. Isters. Institute. Directed A	cyclic GRaphs.
BicQueya – Explicit framework:     Bistro – Library to build and exe     Bipte – Tool for running and ma     Brefly – Python Meta-program     Cluster Flow – Command-line to     Cluster Flow – Command-line to     Cluster Flow – Command-line to     Conges – Programming model     Conges – Programming model     Congenz – Light-weight workflow     Consecution – A Python Bitrary for ma     Cromwell – Workflow Managem     Cuneiform – Advanced function     Dagobah – Simple DAG-based     Dags – As ia leable partiel	with we horitoring and cust typed scientific wort anging bioinformatics pip ming Library for Job Flow old which uses common or for distributed infrastruct management application a bastraction inspired by ssively parallel workflow a la workflow language and a workflow language and ob scheduler in Python. ramework for writing ad computing library for an	resource estimation. kflows. Control. Juater managers to run bic submission of computation ures. n. Apache Storm topologies. I. di framework, implemented executing bioinformatics alyrics.	informatics   al jobs to clu m the Broad in Erlang, pipelines as	pipelines. Isters. Institute. Directed Av	cyclic GRaphs.
BioQueue - Explicit framework:     Bistro - Library to build and exe     Bippe - Tool for running and ma     Briefly - Python Meta-program     Cluster Flow - Command-line to     Cluster Flow - Command-line to     Cluster Flow - Command-line to     Compas - Programming model     Connac - Liph-weight workflow     Consecution - A Python pipelin     Cosmos - Python library for ma     Commel- Workflow Annagem     Cunefitorm - Advanced function     Dagobah - Simple DAG-based j     Dagr - A scala based DSL and f     Dask- Dask ia fitexible parallel     Dockerflow - Workflow Innagem	with web monitoring and cute typed scientific work anging bioinformatics pip ming Library for Job Flow of which uses common cubility, and hassle-free as for distributed infrastructure a batraction inspired by saively parallel workflow a batraction inspired by saively parallel workflow a batraction inspired by a batra	resource estimation. kifows. Belines. Control. Juster managers to run bio Juster managers to run bio Juster managers to run bio Juster managers Apache Storm topologies. L Apache Storm topologies. S scientific workflows for framework, implemented executing bioinformatics a series of tasks in Docke	informatics   al jobs to clu m the Broad in Erlang. pipelines as	pipelines. Isters. Institute. Directed A	cyclic GRaphs.
BioDusue - Explicit framework :     Bistro - Uirary to build an dree     Bipipe - Tool for running and ma     Briefly - Python Meta-program     Cluster Flow - Command-line to     Cluster/bo - Automated reprod.     Conse: A programming model     Conserved - Programming model     Conserved - Programming model     Conserved - Python Bioray for an     Comwell - Workflow Managem     Cunderfor - A september DAG-based j     Dag-bah - Simple DAG-based j     Dag - A scale based DSL, and f     Dask - Dask ta a facible parallel     Dokt - Task management A avin	with web monitoring and used typed scientific work anging bioinformatics pip ming Library for JoB Towo of which uses common cicklity, and hastel-free or distributed infrastruct- wranagement application abstraction inspired by a bastraction inspired by a b	resource estimation. kt/lows. beliens. Control. Ulater managers to run bi ulumission of computation ures. A pache Storm topologies. A da scientific workflows fro d s scientific workflows fro f framework, implemented executing bioinformatics alydics. a series of tasks in Docket	informatics ; al jobs to clu m the Broad in Erlang. Dipelines as l	pipelines. ısters. Institute. Directed A	cyclic GRaphs.
BioQueue - Explicit framework:     Bistro - Library to build and exe     Bipte - Tool for running and ma     Bindly - Python Meta-program     Cluster Flow - Command-line to     Cluster Flow - Command-line to     Cluster Flow - Command-line to     Compas - Programming model     Cosmos - Lython bitrary for ma     Commed-line to - A Python pipelin     Cosmos - Python Bitrary for ma     Commed-line - Workflow Managem     Cunefform - Advanced function     Dagobah - Simple DA/O-based)     Dagy - A scala based DSL, and f     Dask - Dask is a flexible parallel     Dockerflow - Workflow runner t     Dock-Task management & auto     Drake - Boux DSL, alsh of the	with web monitoring and use typed scientific work use typed scientific work of which uses common cickling, and hastel-free se for distributed infrastructure abstraction inspired by a bastaction inspired by any science assistively parallel workflows ent. System geared toward all workflow language and is computing library for an at uses DataBow for writing and is computing library for an at uses DataBow for No.	resource estimation. kt/ows. kt/ows. Control. Luter managers to run bic submission of computation ures. n. Apache Storm topologies. is a scientific workflows for framework, implemented executing bioinformatics alytics. a series of tasks in Docker e.	informatics ; al jobs to clu m the Broad in Erlang. bipelines as l	pipelines. Isters. Institute. Directed A	cyclic GRaphs.
ElicQueue - Explicit framework :     Elistro-Uterry to build and exe     Epipe - Tool for running and ma     Eriefty - Python Meta-program     Cluster Flow - Command-line to     Cluster Job - Automated reprod.     Compas - Programming model     Conand - Light-weight workflow     Compas - Programming model     Commel - Workflow Managem     Commel - Workflow Managem     Commel - Workflow Managem     Compas - Angehan Based DSL and f     Dask - Ascala Based DSL and f     Dask - Ascala Is afleable paralle     DockerTow - Workflow runner t     Drak - Robust DSL alin to Mal     Drake - Robust DSL alin to Mal     Drake - Robust DSL alin to Mal	with web monitoring and used typed scientific work anging bioinformatics pip ming Library for JoB Towo of which uses common cobility, and hastel-free or distributed infrastruct- ur management application is abstraction inspired by solvey parallel workflows ent System gareet toward a workflow language and workflow language ob scheduler in Python. ranework for writing and computing library for an hat uses Dataflow to run mation tool. w, implemented in Colgung.	resource estimation. kt/lows. kt/lows. Control. Uater managers to run bic lubarission of computation re. Apache Storm topologies. A da scientific workflows fro framework, implemented executing bioinformatics aytics. a series of tasks in Docker e.	informatics al jobs to clu m the Broad in Erlang. Dipelines as l	pipelines. Isters. Institute. Directed Av	cyclic GRaphs.

# Interpreting bioinformatics analysis results ?





 $\ll$  Which alignment algorithm was used when predicting this pathogenic score ?  $\gg$ 



« A new version of a reference genome is available, which genome was used when predicting these phenotypes ? »

Need for an overall tracking of provenance over multiple workflows !

# Provenance in Computer Science

« Provenance information describes the **origins** and the **history** of data in its **life cycle**. »

« Today, data is often made **available on the Internet** with **no centralized control** over its integrity: data is constantly being **created**, **copied**, **moved** around, and **combined** indiscriminately. Because information sources (or different parts of a single large source) may vary widely in terms of quality, it is essential to provide **provenance** and other **context** information which can **help end users judge** whether query **results are trustworthy**. »

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in Databases: Why, How, and Where. Found. Trends databases 1, 4 (April 2009), 379-474. DOI=<u>http://dx.doi.org/10.1561/190000006</u>

# 20 years of provenance ...

Foundations and Trends<sup>®</sup> in Databases Vol. 1, No. 4 (2007) 379–474 © 2009 J. Cheney, L. Chiticariu and W.-C. Tan DOI: 10.1561/1900000006



### Provenance in Databases: Why, How, and Where

### James Cheney<sup>1</sup>, Laura Chiticariu<sup>2</sup> and Wang-Chiew Tan<sup>3</sup>

<sup>1</sup> University of Edinburgh, UK, jcheney@inf.ed.ac.uk

<sup>2</sup> IBM Almaden Research Center, San Jose, CA, USA,

chiti@almaden.ibm.com

<sup>3</sup> University of California, Santa Cruz, CA, USA, wctan@cs.ucsc.edu

### Abstract

Different notions of provenance for database queries have been proposed and studied in the past few years. In this article, we detail three main notions of database provenance, some of their applications, and compare and contrast amongst them. Specifically, we review why, how, and where provenance, describe the relationships among these notions of provenance, and describe some of their applications in confidence computation, view maintenance and update, debugging, and annotation propagation.

Which DB tuples contribute to a query result ? Which operators ?

### PROV-O: The PROV Ontology W3C Recommendation 30 April 2013

### This version:

W3C Recommendation

http://www.w3.org/TR/2013/REC-prov-o-20130430/ Latest published version: http://www.w3.org/TR/prov-o/ Implementation report: http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/ **Previous version:** http://www.w3.org/TR/2013/PR-prov-o-20130312/ Editors: Timothy Lebo, Rensselaer Polytechnic Institute, USA Satya Sahoo, Case Western Reserve University, USA Deborah McGuinness, Rensselaer Polytechnic Institute, USA **Contributors:** (In alphabetical order) Khalid Belhajjame, University of Manchester, UK James Cheney, University of Edinburgh, UK David Corsar, University of Aberdeen, UK Daniel Garijo, Ontology Engineering Group, Universidad Politécnica de Stian Soiland-Reyes, University of Manchester, UK Stephan Zednik, Rensselaer Polytechnic Institute, USA Jun Zhao, University of Oxford, UK

A standardized lightweight, extensible model to represent Provenance on the Web.













# **PROV** extensions

nature > scientific data > articles > article

### Open Access Published: 06 December 2016

### Sharing brain mapping statistical results with the neuroimaging data model

Camille Maumet ⊠, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Krzysztof J. Gorgolewski, Karl G. Helmer, Mark Jenkinson, David B. Keator, B. Nolan Nichols, Jean-Baptiste Poline, Richard Reynolds, Vanessa Sochat, Jessica Turner & Thomas E. Nichols

Scientific Data 3, Article number: 160102 (2016) Cite this article

5277 Accesses 30 Citations 42 Altmetric Metrics

https://doi.org/10.1038/sdata.2016.102





### Research | Open Access | Published: 06 January 2022

End-to-End provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach

Sheeba Samuel 🖂 & Birgitta König-Ries

Journal of Biomedical Semantics 13, Article number: 1 (2022) Cite this article 3475 Accesses | 3 Citations | 8 Altmetric | Metrics

https://doi.org/10.1186/s13326-021-00253-1



### ACCEPTED PAPERS



### Astronomical Pipeline Provenance: A Use Case Evaluation

Michael A. C. Johnson, Institute of Data Science (DLR) and Max Planck Institute for Radio Astronomy; Marcus Paradies and Marta Dembska, Institute of Data Science (DLR); Kristen Lackeos, Hans-Rainer Klöckner, and David J. Champion, Max Planck Institute for Radio Astronomy, Sirko Schindler, Institute of Data Science (DLR)

https://doi.org/10.48550/arXiv.2109.10759

Home > Provenance and Annotation of Data and Processes > Conference paper

### Towards a Provenance Management System for Astronomical Observatories

Mathieu Servillat C, François Bonnarel, Catherine Boisson, Mireille Louys, Jose Enrique Ruiz & Michèle Sanguillon

Conference paper | First Online: 09 July 2021

515 Accesses 1 Citations 8 Altmetric

Part of the Lecture Notes in Computer Science book series (LNISA, volume 12839)

https://doi.org/10.1007/978-3-030-80960-7 20

Research | Open Access | Published: 31 January 2022

linimini Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation

Max Schröder 🖂, Susanne Staehlke, Paul Groth, J. Barbara Nebe, Sascha Spors & Frank Krüger

Journal of Biomedical Semantics 13, Article number: 4 (2022) Cite this article

4265 Accesses | 6 Citations | 9 Altmetric | Metrics

https://doi.org/10.1186/s13326-021-00257-x

# Many expectations ...

Comparability, transparency, confidence +

- Citing researchers and organisations
- Identifying critical data / software resources associated to scientific results
- Identifying possible bias when reusing / sharing pre-trained models

# **Reuse** instead of re-execution ?





# Is PROV enough for reuse ?

11 a prov:Bundle, prov:Entity;

- 12 prov:wasAttributedTo <#galaxy2prov>;
- 13 prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
- 14 .
- 15 16 <#72486b583fe152f0>
- 17 a prov:Activity ;
- 18 prov:wasAssociatedWith <#catl> ;
- 19 prov:startedAtTime "2015-12-15T12:54:50.749845"^^xsd:dateTime;
- 20 prov.endedatmime\_"2015\_12\_15m12.55.57\_016799"^^yed.datemime

### Visualise

GVGW 61/CHIPATED CHIPATEL AND A CHIP setat 059.dat LOCARS LOCARS and Locars and 1050.dat BOOM TO TO T L002172 004 fasto gz 0044 sea 001/dataset\_1062.dat rod-galax, alaxy/galaxy

CA L004 R1 004.fastg.gz es/galaxy-datasets/001/dataset\_1176.dat deilasoloalaxy-files/galaxy-datasets/001/dataset\_1177.dat Heren and Polician and as(Pfiles/galles

phased and WT2: ascended distaset 1173.dat Ac\_LOOdal Stophar 9 a WT1 accepted dataset\_1167.dat P1(002) atasets/001/dataset\_1094.dau axy/prod-gala

TopHat on 6WT3 : accepted\_hits 1093.dat

C Dalla LOOR R 03\_R2\_0046843340529/p set 1089.dat NGWING DELASUA LOOK HENORALES ALLONS TO 1/02 438 9 0784 dat axam\_es/galaxy-datasets/001/dataset\_1088.dat Ba.dat ASI - 97092.dat

LOCHES/Galage-battasgis/001/dataset 1120.dat

Logal Set 505 tasaset 1104 dat 1128.dat

1056.dat abstets 000 fastatasset 1077.dat

1076.dat

and by 4 and 3 . The axy (galaxy - s) galaxy datasets /001/dataset, 1469 dat maximum datasets /001/dataset, 1469 datasets /001/dataset, 1469 datasets /001/dataset, 1469 datasets /001/datasets /001/datas TopHat on 45 WT1 states w 3 and states and s TopHat on 45WT3 : insertions

Logalaxy/galaxy-datasets/001/dataset\_11/4.dat 4 T2\_fbpfaston 45WT2: accepted\_hits //galaxy/prod-galaxy/galaxy/galaxy-galaxy-datasets/001/dataset\_11/4.dat 4 T2\_fbpfaston 45WT2: accepted\_hits //galaxy/prod-galaxy/galaxy/galaxy-galaxy-datasets/001/dataset\_11/4.dat 4 T2\_fbpfaston 45WT2: accepted\_hits //galaxy/prod-galaxy/galaxy/galaxy-datasets/001/dataset\_11/4.dat 4 T2\_fbpfaston 45WT2: accepted\_hits

/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/00 /galaxy/prod-galaxy/galaxy/galaxy/galaxy-files/galaxy-datas

Cufflinks or OLENG TAGENERS SAVE AND THE PROPERTIES IN THE PROPERTIES OF THE PROPERT 

Cufflinks on 45 Wuff: as the product of a tand the solution of ufflinks on 6WT2 : transcript

Cuffmerge on data 43, data 66, and other Cuffquant on data 42 and data 43: Abunda

/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset\_1450.dat http://fr.symetric#toolshed.g2.bx.psu.ed /galaxy/prod-galaxy/galaxy/galaxy/galaxy-files/galaxy-dataset /galaxy/prod-galaxy/galaxy/galaxy/galaxy-files/galaxy-dataset /galaxy/pgataga/arodegatacy/ggatacy/jbategarfiles/gataceted@catadatata /galaxy g

# Is PROV enough for reuse ?

11 a prov:Bundle, prov:Entity;

- 12 prov:wasAttributedTo <#galaxy2prov>;
- 13 prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
- 14 . 15
- 16 <#72486b583fe152f0>
- 17 a prov:Activity ;
- 18 prov:wasAssociatedWith <#catl> ;
- 19 prov:startedAtTime "2015-12-15T12:54:50.749845"^^xsd:dateTime;
- 20 prov.endedatmime\_"2015\_12\_15m12.55.57\_016799"^^xed.datemime.

### Visualise

6VGW 6VCWATEGOUGH BOODER LOUGH LOUGH LOUGH BOODER SUS RESTORES (001/dataset\_1071.dat setat 059.dat **BCicket** L0021712\_004\_fastq.gz 004001/dataset\_1062.dat rod-galax, alaxy/galaxy

Hanna (R) ..... Too fine-grained !

SX gales/ga axy/prod-gal

92%\_LUwes/001/02taset91084.dat

Es/gales/-tattesgis/001/dataset 1120.dat

4 dat 1128.dat

dataset\_1056.dat iets@601astataszet 1077.dat

1076.dat

at on 45avr 3 : Hand 2 y galaxy galaxy datasets 001 (dataset 1469 dat TopHat ŏ ophat on 450v 13; alterious4514773\_R2.fastq TopHat on 45WT3 : insertions deletions-files/galaxy-datasets/001/dataset\_1468.dat TopHat on 45WT2 : epiterevint target as a second data sets/001/dataset\_1467.dat TopHat on 45WT2 : epiterevint target as topHat on 45WT3 : accepted\_hits

No **domain**-specific **concepts** !

dat gat

/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset\_11/4 dat 4 T2\_fbpfaston 45WT2 : accepted\_hits /fatala//om6WT2t0///r\_synetric#tones/galaxy-datasets/001/dataset\_11/4 dat / Galaxy/prod-gala\_\_\_\_galaxy/galaxy-files/galaxy-datasets/001/dataset\_1/ /fatala//om6WT2t0///r\_synetric#tones/galaxy-datasets/001/dataset\_11/4 dat / Galaxy/prod-gala\_\_\_\_galaxy/galaxy-files/galaxy-datasets/001/dataset\_1/ /galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/00 /galaxy/prod-galaxy/galaxy/galaxy/galaxy-files/galaxy-datas

> Aiexpression tata 49 and data 127 Bote alter paras of

Wuff: as a fight prover the reader of the rs: Chailes and the wist as a standard Cuffmerge on data

/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset\_1450 http://fr.symetric#toolshed.ga /galaxy/prod-galaxy/galaxy/galaxy/galaxy/ /galaxy/pgatagalarodgataaxygataaxyfibataarfiles/gatagetedal

# Annotated bioinformatics tools catalog

Ontology terms (EDAM) to "tag"

- ► topic
- input data
- processing
- output data



# Approach



### Methods and tools

graph pattern matching, inference rules, SPARQL, Python, Jupyter notebooks

### Machine-Human readable experiment summaries

### Machine-Human readable experiment summaries

```
[...]
:head {
    _:npl a np:Nanopublication .
    _:npl np:hasAssertion :assertion .
    _:npl np:hasProvenance :provenance .
    _:npl np:hasPublicationInfo :pubInfo .
}
:assertion {
    <http://snakemake-provenance/Samples/Sample1/
    BAM/Sample1.merged.bai> rdfs:seeAlso
    <http://edamontology.org/operation_3197> .
```

```
<http://snakemake-provenance/VCF/hapcaller.
indel.recal.filter.vcf.gz> rdfs:seeAlso
<http://edamontology.org/operation_3695> .
```

[...]

### 1. It's possible to automatically annotate produced data with **ontology terms**

### Machine-Human readable experiment summaries

Read summarisation





Genome indexing

Generation

Sequence alignment

Read mapping

```
_:np1 a np:Nanopublication .
_:np1 np:hasAssertion :assertion .
_:np1 np:hasProvenance :provenance .
_:np1 np:hasPublicationInfo :pubInfo .
```

```
J
```

[...] :head {

```
:assertion {
    <http://snakemake-provenance/Samples/Sample1/
    BAM/Sample1.merged.bai> rdfs:seeAlso
    <http://edamontology.org/operation_3197> .
```

```
<http://snakemake-provenance/VCF/hapcaller.
indel.recal.filter.vcf.gz> rdfs:seeAlso
<http://edamontology.org/operation_3695> .
```

, [...]

### It's possible to automatically annotate produced data with ontology terms

• • •

. . .

The file Samples/Sample1/BAM/Sample1.realign.bai results from tool gatk2\_indel\_realigner-IP which Locally align two or more molecular sequences.

It was produced in the context of Rare Coding Variants in ANGPTL6 Are Associated with Familial Forms of Intracranial Aneurysm

- 2. It's possible to automatically display the **typical bioinformatics tasks** data originate from
- 3. It's possible to document data with **text** leveraging ontology definitions (EDAM)

# Wrap-up





# Take-home message & perspectives

- PROV = consensual model to represent and share provenance on the web
- ≠ communities/tools → PROV
   heterogeneity
- Computational reproducibility

   → fine-grained capture (hardware ? OS ?)
   → information overload for humans
- ► FAIR
  - **R1.2** criteria poorly validated<sup>1</sup>
  - domain-specific annotations
  - $\rightarrow$  scientific context ? DMPs ?

Future works
 PEPR Santé-Numérique
 (ShareFAIR, NeuroVasc)



# Acknowledgments



Audrey Bihouée, Institut du Thorax, BiRD Bioinformatics facility, University of Nantes



Hala Skaf-Molli, LS2N, University of Nantes

### GDR MaDICS action ReproVirtuFlow



Khalid Belhajjame, LAMSADE, University of Paris-Dauphine, PSL



contact: alban.gaignard@univ-nantes.fr

