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

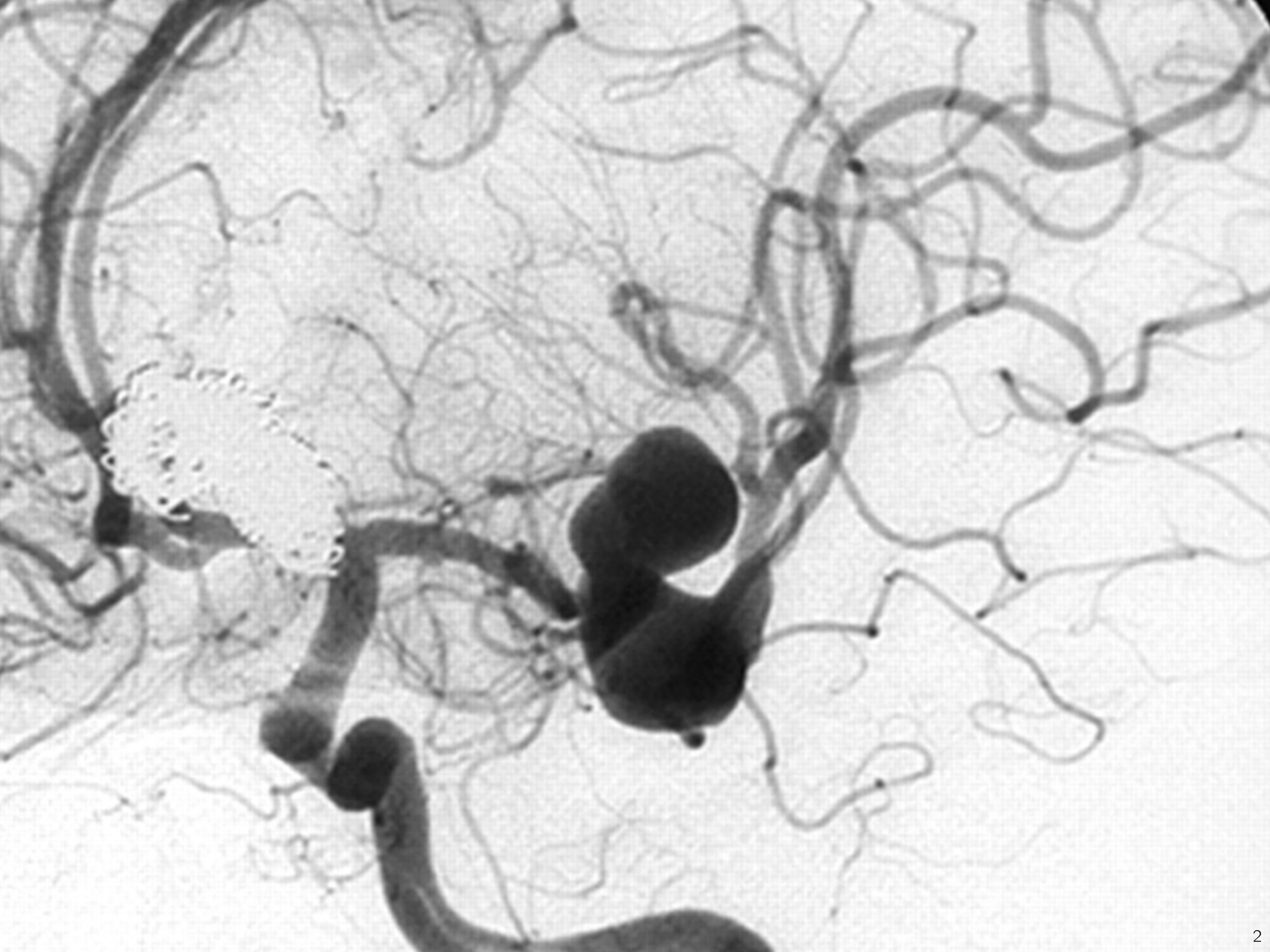
Alban Gaignard, PhD, CNRS

Colloque "Réplicabilité et Reproductibilité de la Recherche"

MITI, CNRS

Paris, 08 septembre 2023

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES l'institut du thorax



Predict the evolution ?

Understand the pathology ?

Prevention strategies ?

ICAN cohort: 34 univ. hospitals / 3000 subjects

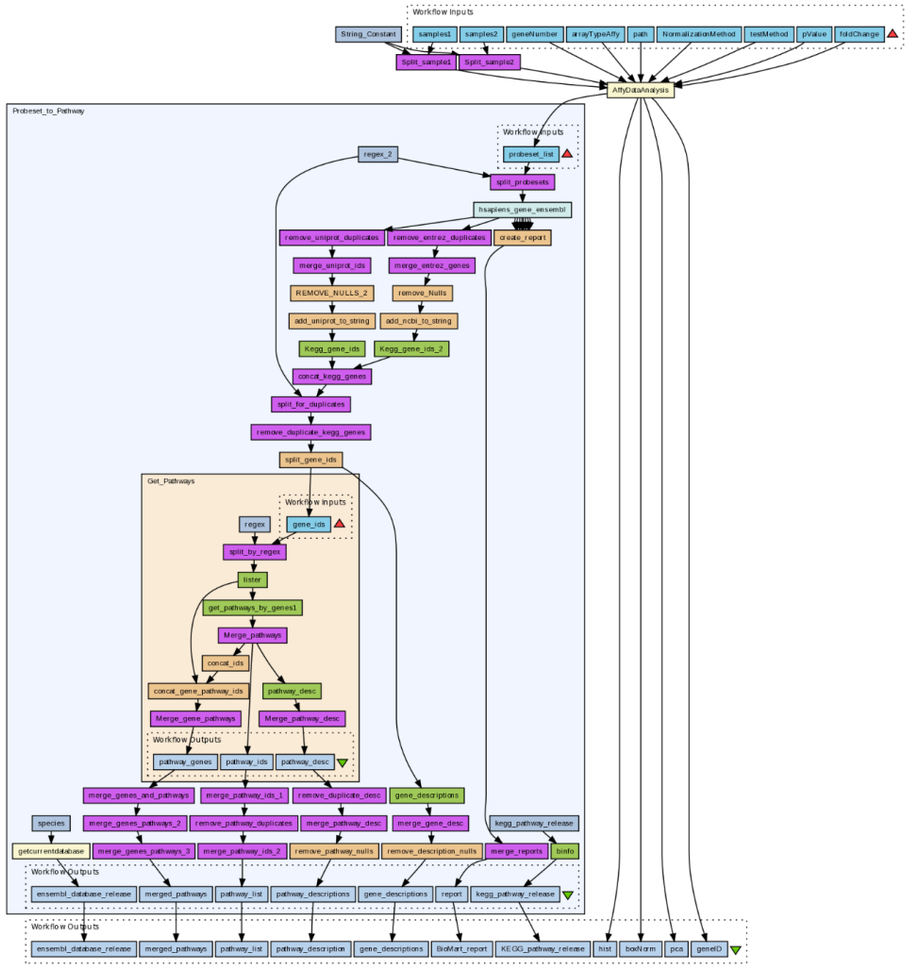
3.000 ToF MRIs
600 whole genomes (analysis in progress)

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. https://doi.org/10.1093/neuros/nyw135



Workflows to enhance trust in scientific results :

→ automation (scalability)

→ abstraction (methods sharing)

→ traceability (provenance)

nextflow

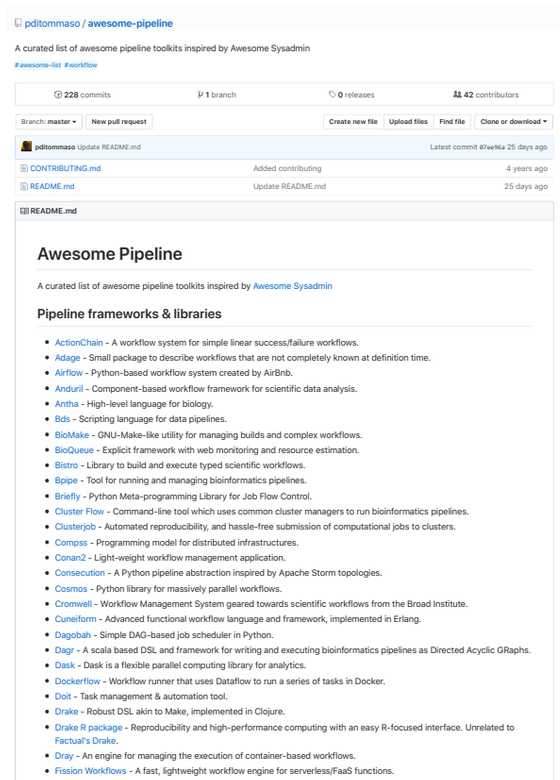
snakemake

Taverna

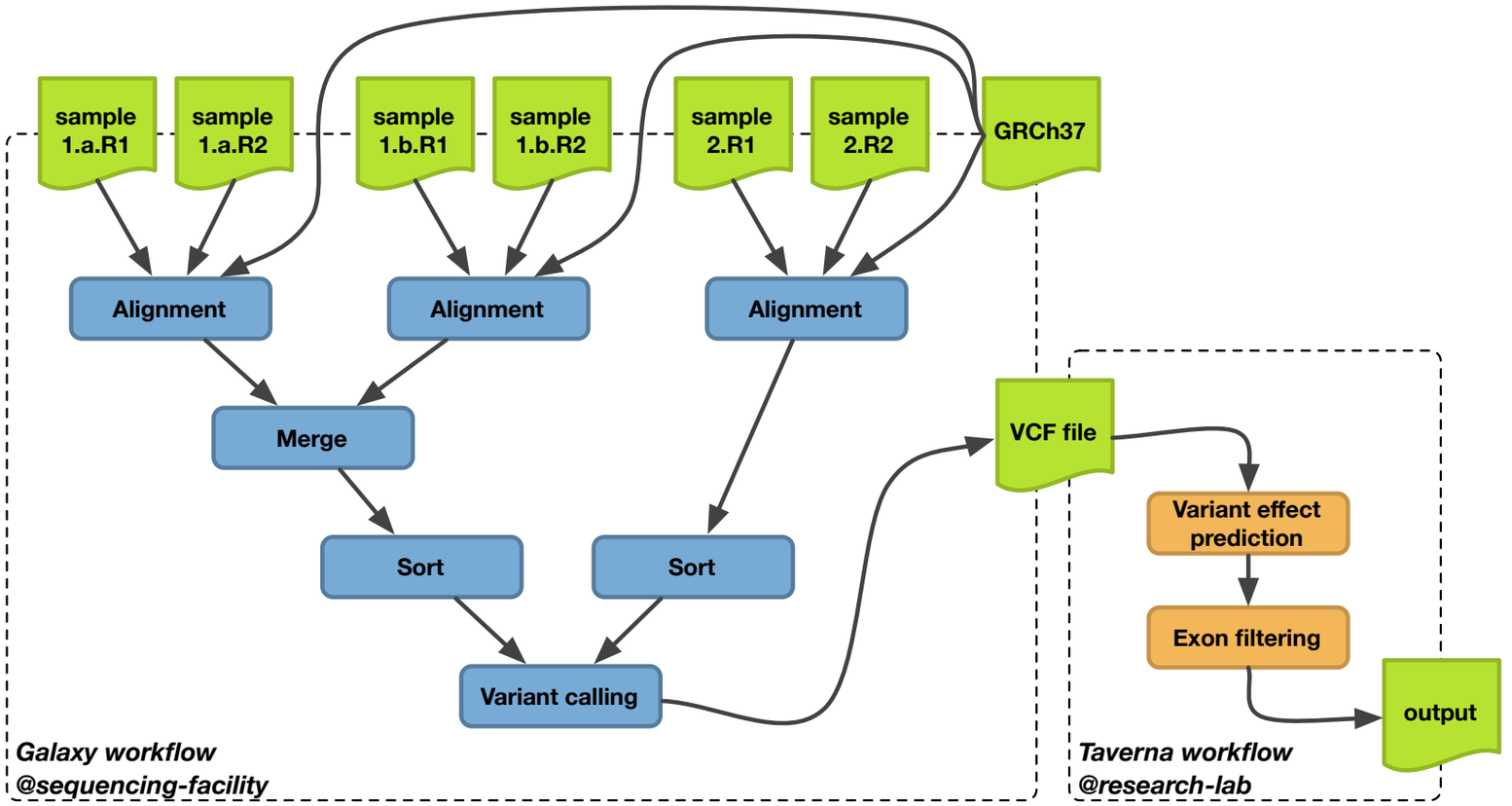
TensorFlow

COMMON WORKFLOW LANGUAGE

Galaxy COMMUNITY HUB



Interpreting bioinformatics analysis results ?



« Which alignment algorithm was used when predicting this pathogenic score ? »

« 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=http://dx.doi.org/10.1561/1900000006

20 years of provenance …

Foundations and Trends® in
Databases
Vol. 1, No. 4 (2007) 379–474

DOI: 10.1561/1900000006

now
the essence of knowledge

Provenance in Databases: Why, How, and Where

James Cheney[1], Laura Chiticariu[2] and Wang-Chiew Tan[3]

[1] *University of Edinburgh, UK, jcheney@inf.ed.ac.uk*
[2] *IBM Almaden Research Center, San Jose, CA, USA, chiti@almaden.ibm.com*
[3] *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.

W3C Recommendation

W3C®

PROV-O: The PROV Ontology

W3C Recommendation 30 April 2013

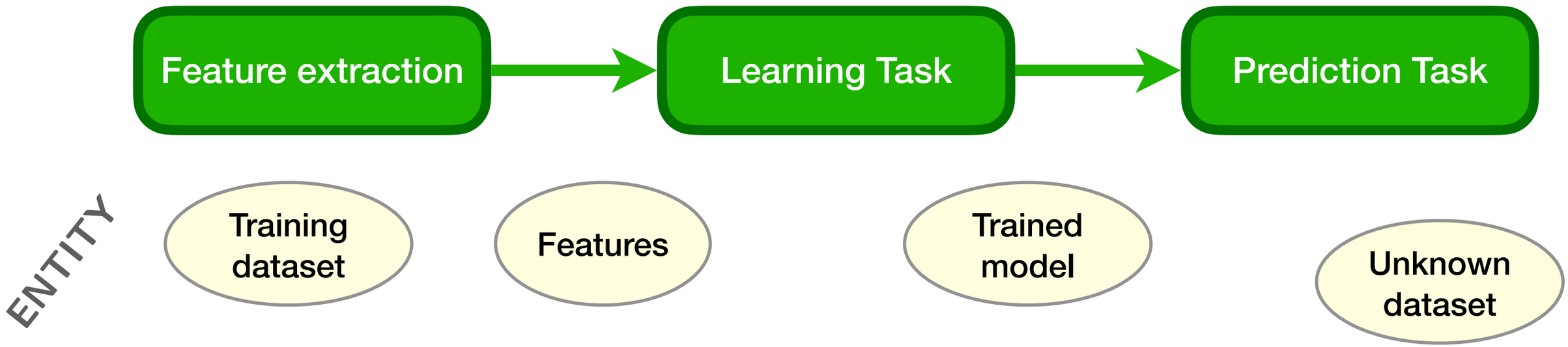
This version:
http://www.w3.org/TR/2013/REC-prov-o-20130430/
Latest published version:
http://www.w3.org/TR/prov-o/
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-o-20130312/
Editors:
Timothy Lebo, Rensselaer Polytechnic 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

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

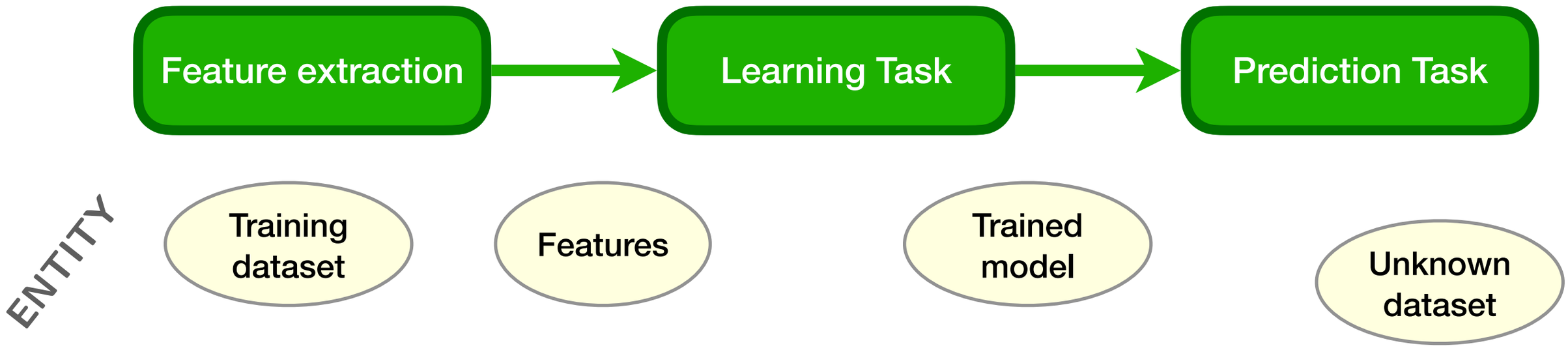
A standardized lightweight, extensible model to represent Provenance on the Web.

Data / Algorithms / Researchers

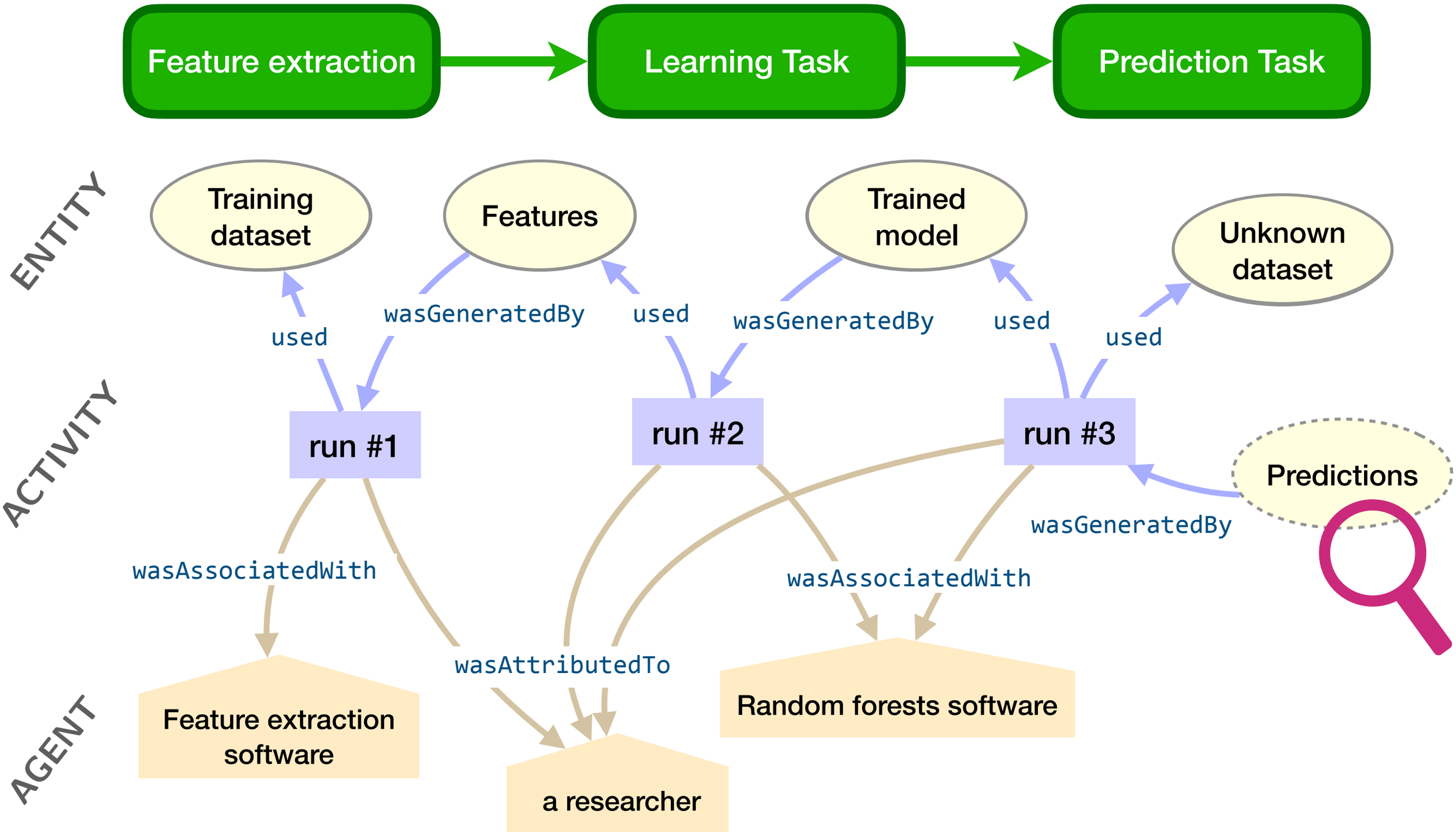
Data / Algorithms / Researchers



Data / Algorithms / Researchers



Data / Algorithms / Researchers



PROV extensions

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

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

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

Research | Open Access | Published: 31 January 2022

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

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

TaPP 2021 PAPERS A USENIX Publication ACCEPTED PAPERS

Astronomical Pipeline Provenance: A Use Case Evaluation

Authors:
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

Towards a Provenance Management System for Astronomical Observatories

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

Conference paper | First Online: 09 July 2021

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

https://doi.org/10.1007/978-3-030-80960-7_20

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

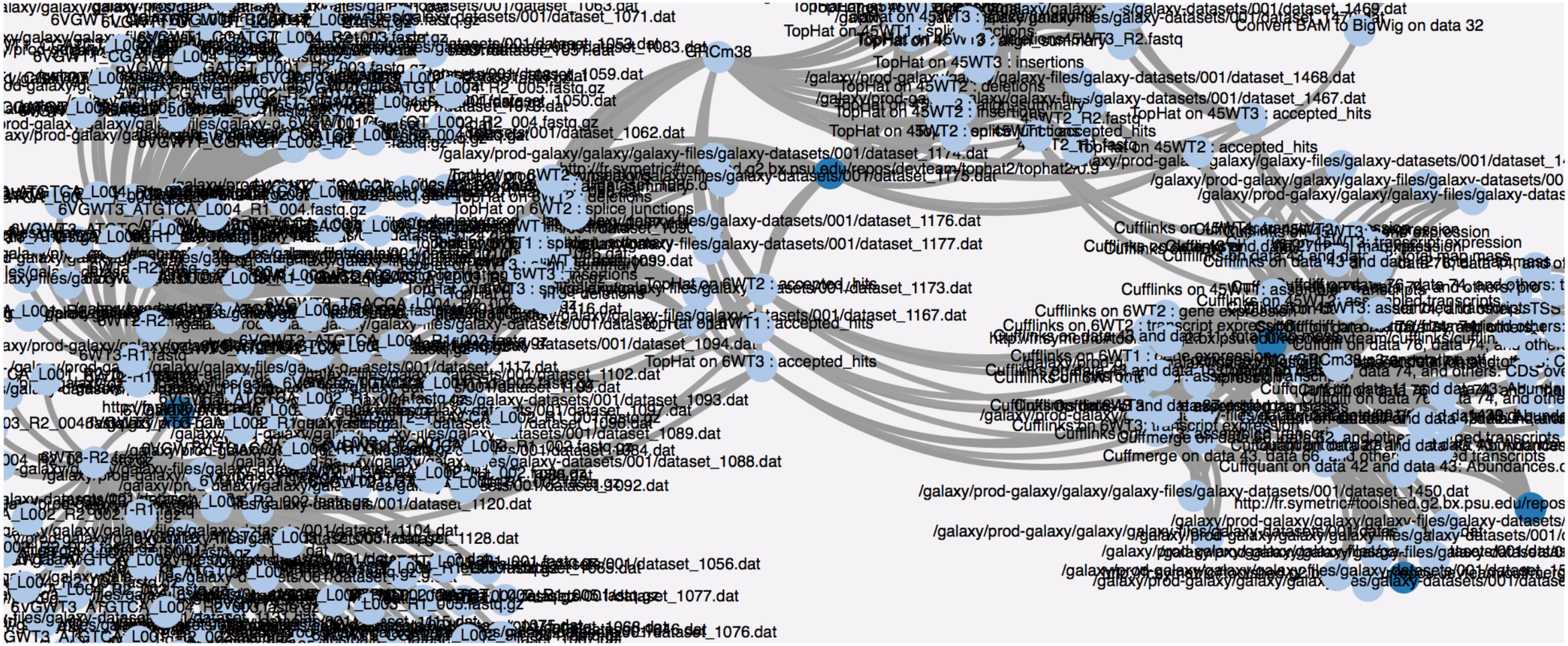
Reuse instead of re-execution ?

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES l'institut du thorax

Is PROV enough for reuse ?

```
   a prov:Bundle, prov:Entity;
   prov:wasAttributedTo <#galaxy2prov>;
   prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#72486b583fe152f0>
    a prov:Activity ;
    prov:wasAssociatedWith <#cat1> ;
    prov:startedAtTime "2015-12-15T12:54:50.749845"^^xsd:dateTime;
    prov:endedAtTime "2015-12-15T12:55:57.016799"^^xsd:dateTime;
```

Visualise



Is PROV enough for reuse ?

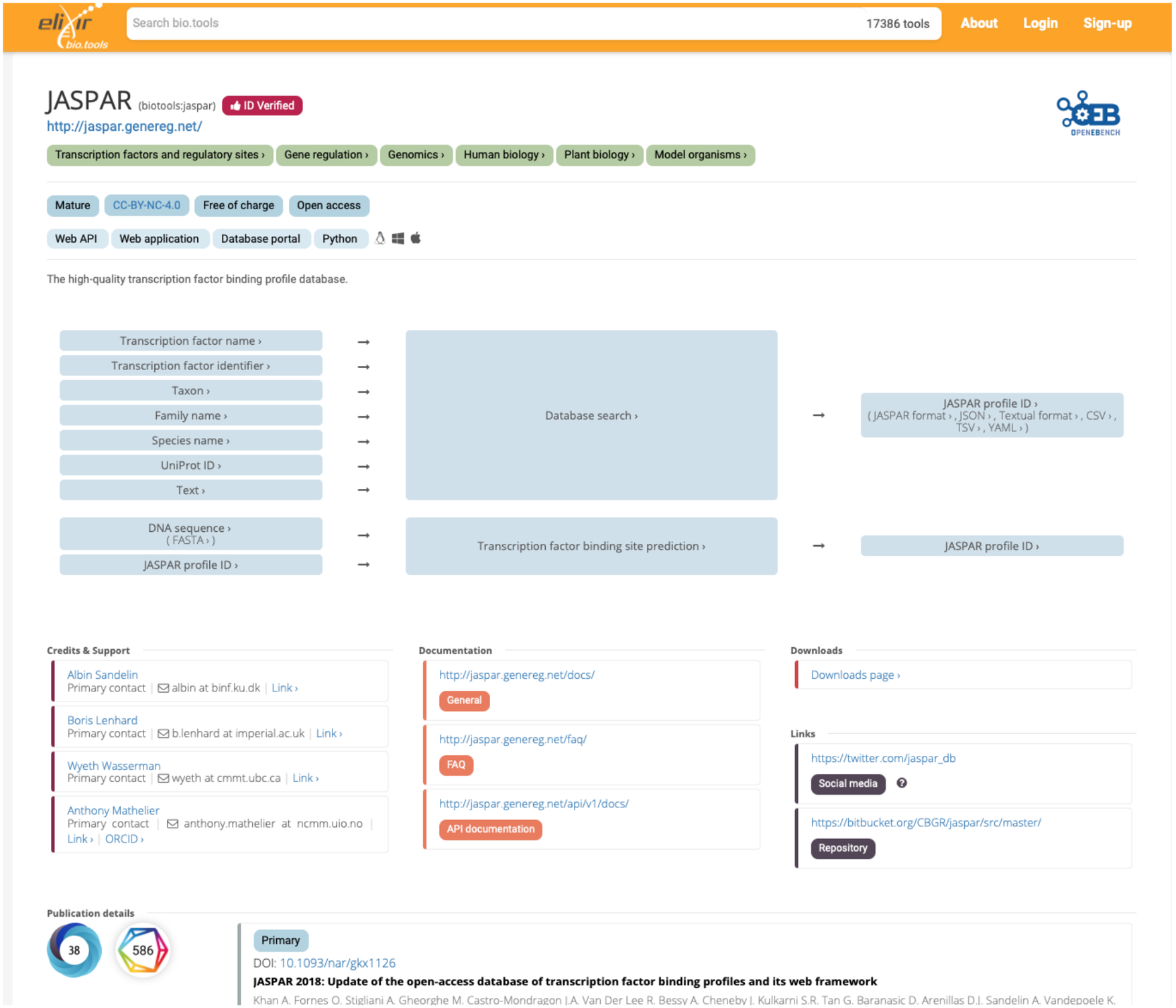
```
   a prov:Bundle, prov:Entity;
   prov:wasAttributedTo <#galaxy2prov>;
   prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#72486b583fe152f0>
    a prov:Activity ;
    prov:wasAssociatedWith <#cat1> ;
    prov:startedAtTime "2015-12-15T12:54:50.749845"^^xsd:dateTime;
    prov:endedAtTime "2015-12-15T12:55:57.016799"^^xsd:dateTime;
```

Visualise

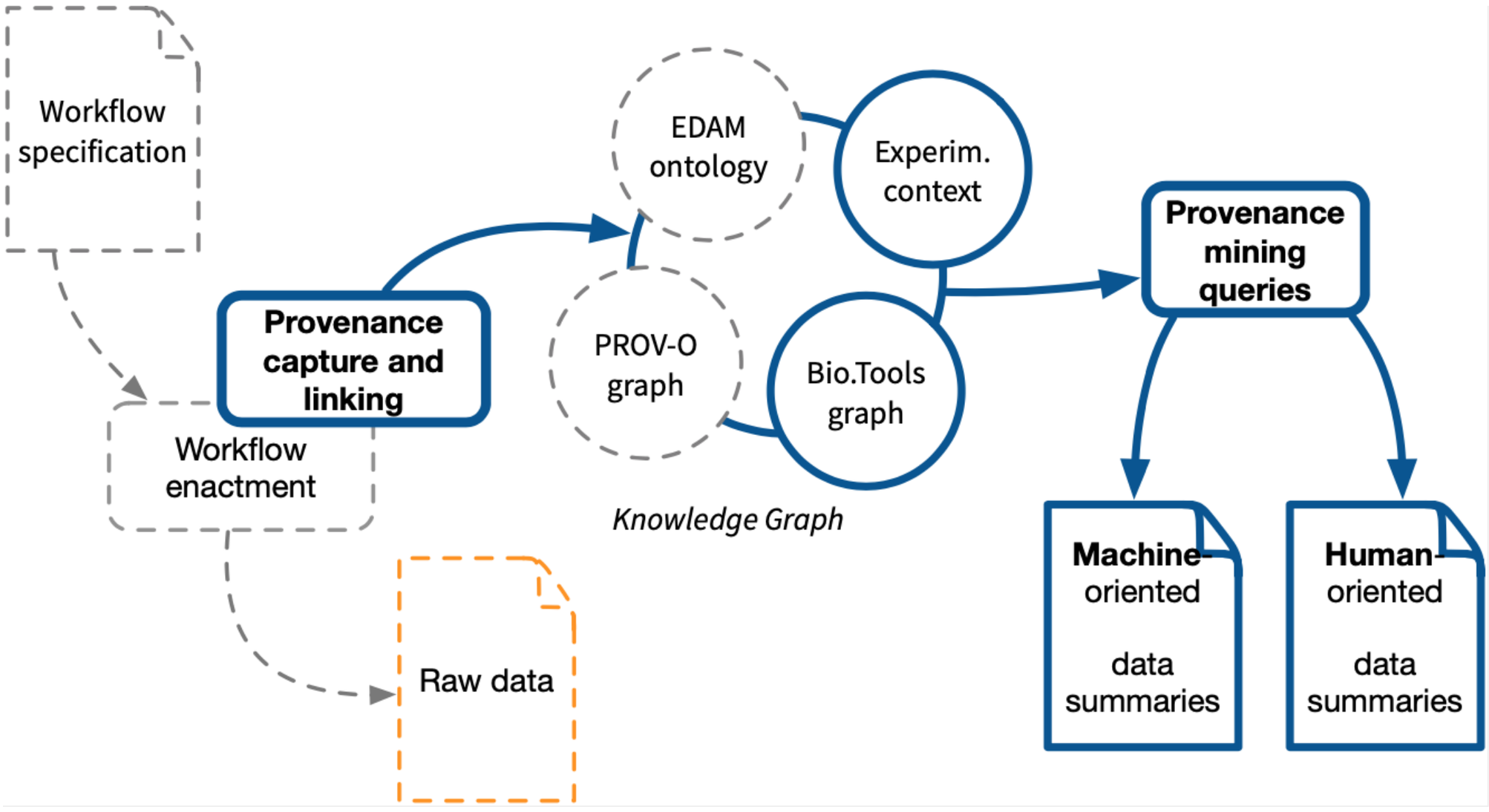
Annotated bioinformatics tools catalog

Ontology terms (EDAM) to "tag"

- topic
- input data
- processing
- output data



Approach



"Which was the reference genome used to produce this VCF file ?"

"A new tool is available, which raw data should I reprocess ?"

Methods and tools

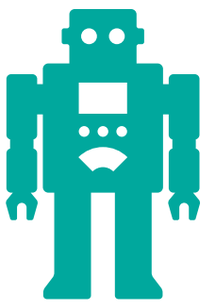
graph pattern matching, inference rules, SPARQL, Python, Jupyter notebooks

A. Gaignard, H. Skaf-Molli, K. Belhajjame **Findable and reusable workflow data products: A genomic workflow case study**. Semantic Web Journal 11 (5), 751-763
https://doi.org/10.3233/SW-200374

Machine-Human readable experiment summaries

https://github.com/albangaignard/fresh-toolbox

Machine-Human readable experiment summaries

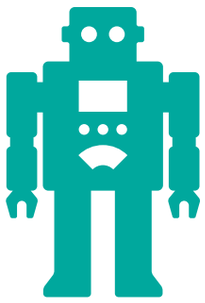


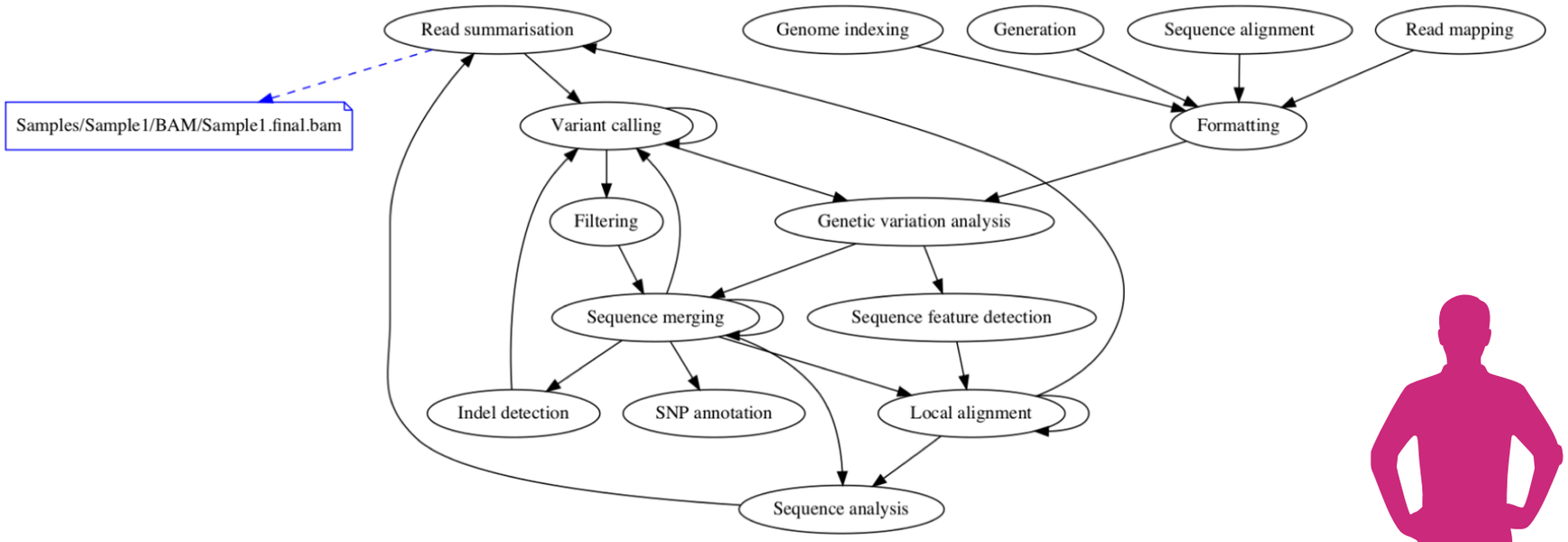
```
[...]
:head {
    _:np1 a np:Nanopublication .
    _:np1 np:hasAssertion :assertion .
    _:np1 np:hasProvenance :provenance .
    _:np1 np:hasPublicationInfo :pubInfo .
}

: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> .
}
[...]
```

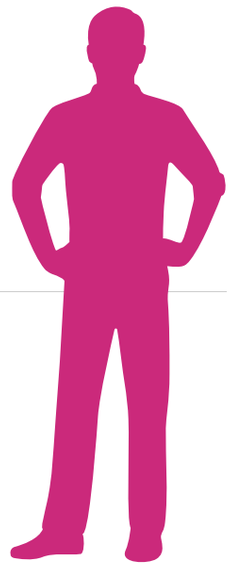
1. It's possible to automatically annotate produced data with ontology terms

https://github.com/albangaignard/fresh-toolbox

Machine-Human readable experiment summaries







```
[...]
:head {
    _:np1 a np:Nanopublication .
    _:np1 np:hasAssertion :assertion .
    _:np1 np:hasProvenance :provenance .
    _:np1 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> .
}
[...]
```

```
...
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
...
```

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

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)

https://github.com/albangaignard/fresh-toolbox

Wrap-up

Inserm

cnrs

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

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
- FAI**R**
  - ⚠ **R1.2** criteria poorly validated[1]
  - ⚠ domain-specific annotations
    → scientific context ? DMPs ?

—— Future works ——

PEPR Santé-Numérique
(ShareFAIR, NeuroVasc)

[1] FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. J Biomed Semant 14, 7 (2023). https://doi.org/10.1186/s13326-023-00289-5

Acknowledgments

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

Hala Skaf-Molli, LS2N, University of Nantes

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

GDR MaDICS
action **ReproVirtuFlow**

contact: alban.gaignard@univ-nantes.fr