# RSECon22 Walkthrough: A FAIR Data Pipeline: provenance-driven data management for traceable scientific workflows

7th September 2022

Ryan Field: Ryan.Field@glasgow.ac.uk
Richard Reeve: Richard.Reeve@glasgow.ac.uk

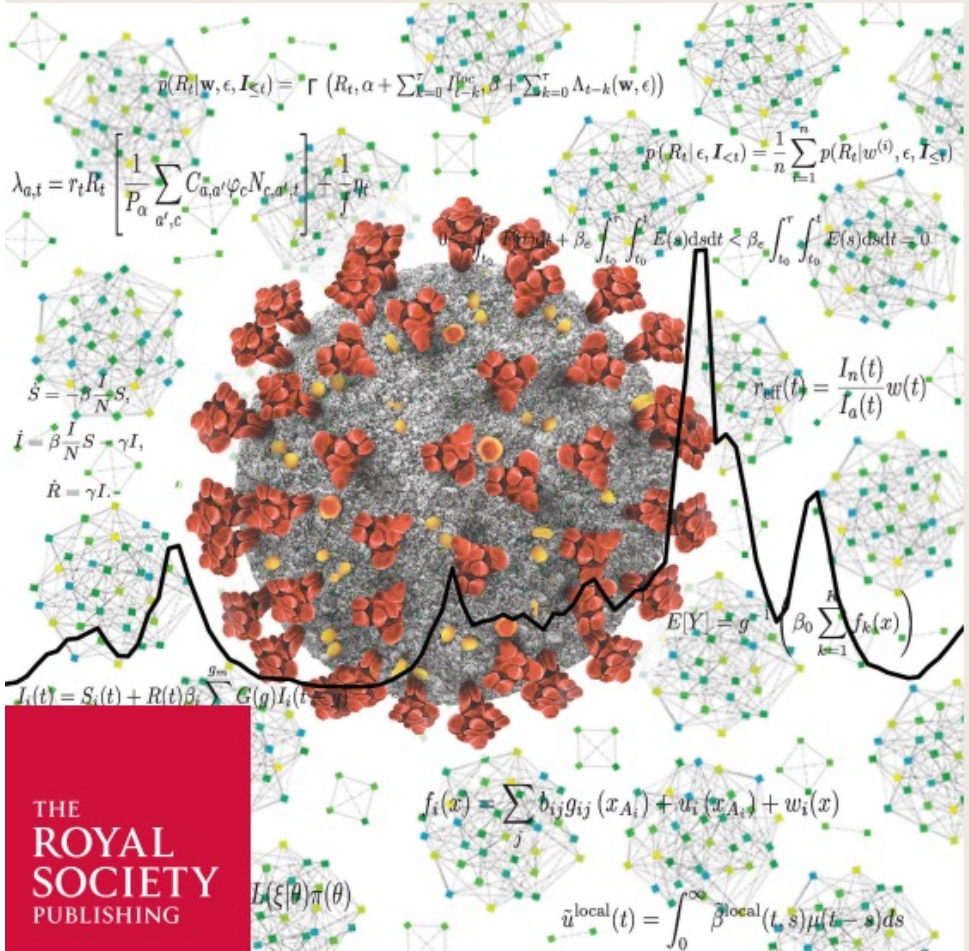
ISSN 1364-503X | Volume 380 | Issue 2233 | 3 October 2022

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES

## Technical challenges of modelling real-life epidemics and examples of overcoming these

Theme issue compiled and edited by Dr Jasmina Panovska-Griffiths, Dr William Waites, and Professor Graeme J. Ackland



THE ROYAL SOCIETY PUBLISHING



PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

**Cite this article:** Mitchell SN *et al.* 2022 FAIR data pipeline: provenance-driven data management for traceable scientific workflows. *Phil. Trans. R. Soc. A* **380**: 20210300. https://doi.org/10.1098/rsta.2021.0300

Received: 11 October 2021
Accepted: 19 April 2022

One contribution of 18 to a theme issue 'Technical challenges of modelling real-life epidemics and examples of overcoming these'.

**Subject Areas:**
e-science, software, bioinformatics, computer modelling and simulation

**Keywords:**
FAIR, provenance, data management, epidemiology, modelling, COVID-19

**Author for correspondence:**
Richard Reeve
e-mail: richard.reeve@glasgow.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6070465.

THE ROYAL SOCIETY PUBLISHING

# FAIR data pipeline: provenance-driven data management for traceable scientific workflows

Sonia Natalie Mitchell[1,2], Andrew Lahiff[6], Nathan Cummings[6], Jonathan Hollocombe[6], Bram Boskamp[7], Ryan Field[3], Dennis Reddyhoff[8], Kristian Zarebski[6], Antony Wilson[9], Bruno Viola[6], Martin Burke[7], Blair Archibald[4], Paul Bessell[10], Richard Blackwell[11], Lisa A. Boden[10], Alys Brett[6], Sam Brett, Ruth Dundas[3], Jessica Enright[2,4], Alejandra N. Gonzalez-Beltran[9], Claire Harris[2,7], Ian Hinder[12], Christopher David Hughes[11], Martin Knight[7], Vino Mano[11], Ciaran McMonagle[2,3], Dominic Mellor[2,5], Sibylle Mohr[1,2], Glenn Marion[2,7], Louise Matthews[1,2], Iain J. McKendrick[2,7], Christopher Mark Pooley[7], Thibaud Porphyre[13], Aaron Reeves[14], Edward Townsend, Robert Turner[8], Jeremy Walton[15] and Richard Reeve[1,2]

[1]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK
[2]Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, Glasgow, G12 8QQ, UK
[3]MRC/CSO Social and Public Health Sciences Unit, Institute of Health and Wellbeing, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK

github.com/FAIRDataPipeline/RSECon22

Search or jump to... / Pull requests Issues Marketplace Explore

FAIRDataPipeline / RSECon22 Public

Edit Pins | Watch 3 | Fork 0 | Star 1

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main | 3 branches | 0 tags | Go to file | Add file | Code

**Your main branch isn't protected**
Protect this branch from force pushing, deletion, or require status checks before merging.

Protect this branch

**Ryan Field** and **Ryan Field** fix line ending — 728dee0 2 days ago — 28 commits

| | | |
|---|---|---|
| .github/workflows | Container does not need to run | 16 days ago |
| Notebooks | Update to Julia 1.8, use ro_crate | 3 days ago |
| Dockerfile | fix line ending | 2 days ago |
| README.md | Update README.md | 2 days ago |

README.md

## RSECon22 Walkthrough: A FAIR Data Pipeline:

About

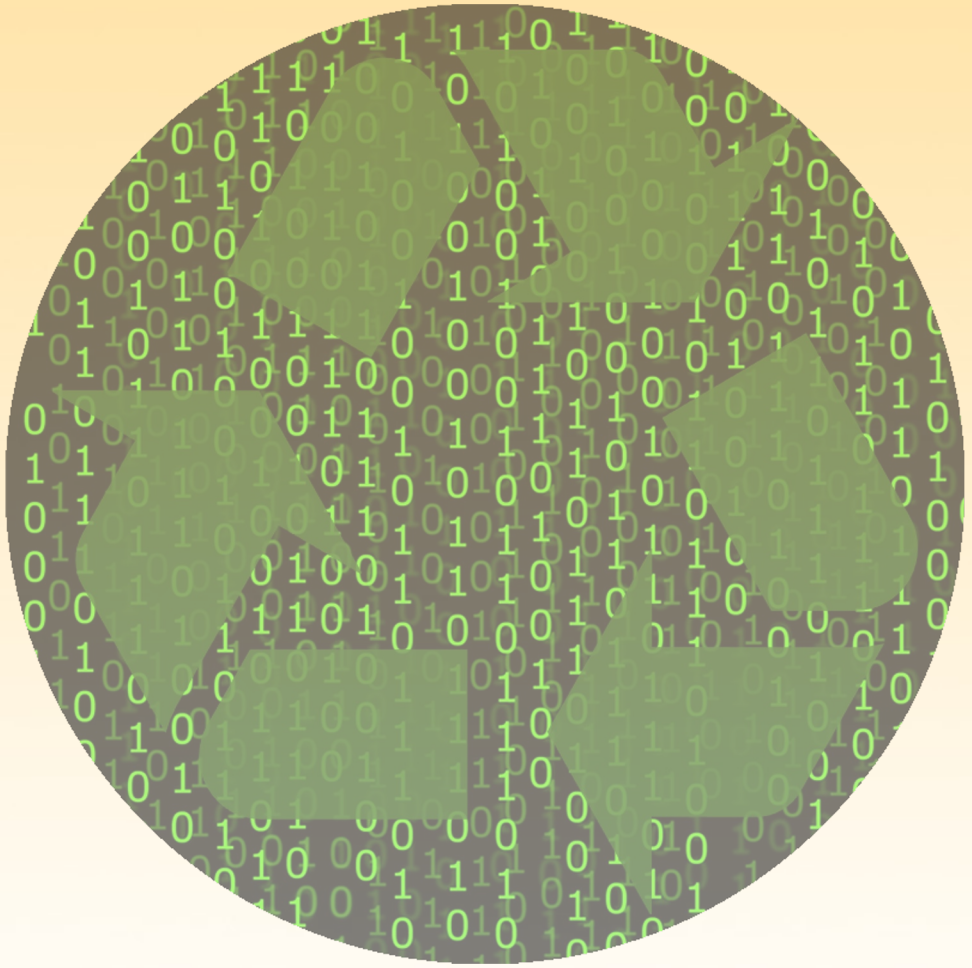
Packages 1

rsecon

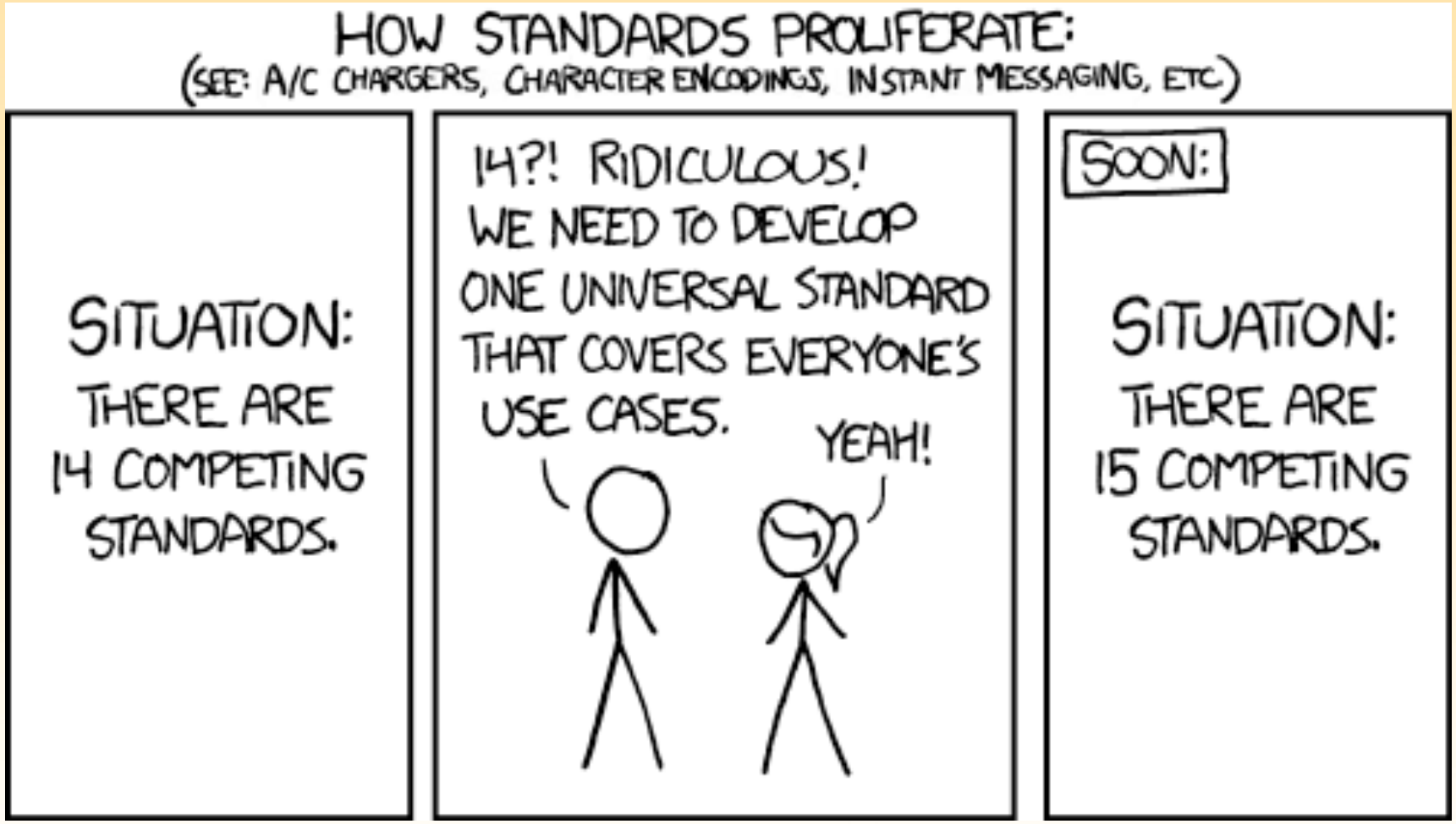
https://github.com/FAIRDataPipeline/RSECon22

# WHY?

- FAIR
  - Findable
  - Accessible
  - Interoperable
  - Reusable
- Traceability / Provenance
- Validity



# Reinventing the Wheel?



# git-annex



git-annex allows managing files with git, without checking the file contents into git. While that may seem paradoxical, it is useful when dealing with files larger than git can currently easily handle, whether due to limitations in memory, time, or disk space.

git-annex is designed for git users who love the command line. For everyone else, the git-annex assistant turns git-annex into an easy to use folder synchroniser.


To get a feel for git-annex, see the walkthrough.

**key concepts**

- git-annex man page
- how it works
- special remotes
- workflows
- sync

**the details**

- encryption
- key-value backends
- bare repositories
- submodules
- internals
- scalability
- design

**other stuff**

- testimonials
- privacy
- what git annex is not
- related software
- public git-annex repos
- thanks
- sitemap

- install
- assistant
- walkthrough
- tips
- bugs
- todo
- forum
- comments
- contact
- thanks

**use case: The Archivist**

Bob has many drives to archive his data, most of them kept offline, in a safe place.

With git-annex, Bob has a single directory tree that includes all his files, even if their content is being stored offline. He can reorganize his files using that tree, committing new versions to git, without worry about accidentally deleting anything.

**use case: The Nomad**

Alice is always on the move, often with her trusty netbook and a small handheld terabyte USB drive, or a smaller USB keydrive. She has a server out there on the net. She stores data, encrypted in the Cloud.

All these things can have different files on them, but Alice no longer has to deal with the tedious process of keeping them manually in sync, or remembering where she put a file. git-annex manages all these data sources as if they were git remotes.

# Open-source Version Control System for Machine Learning Projects

```
$ dvc add images
```

```
$ dvc run -d images -o model.p cnn.py
```

```
$ dvc remote add -d myrepo s3://mybucket
```

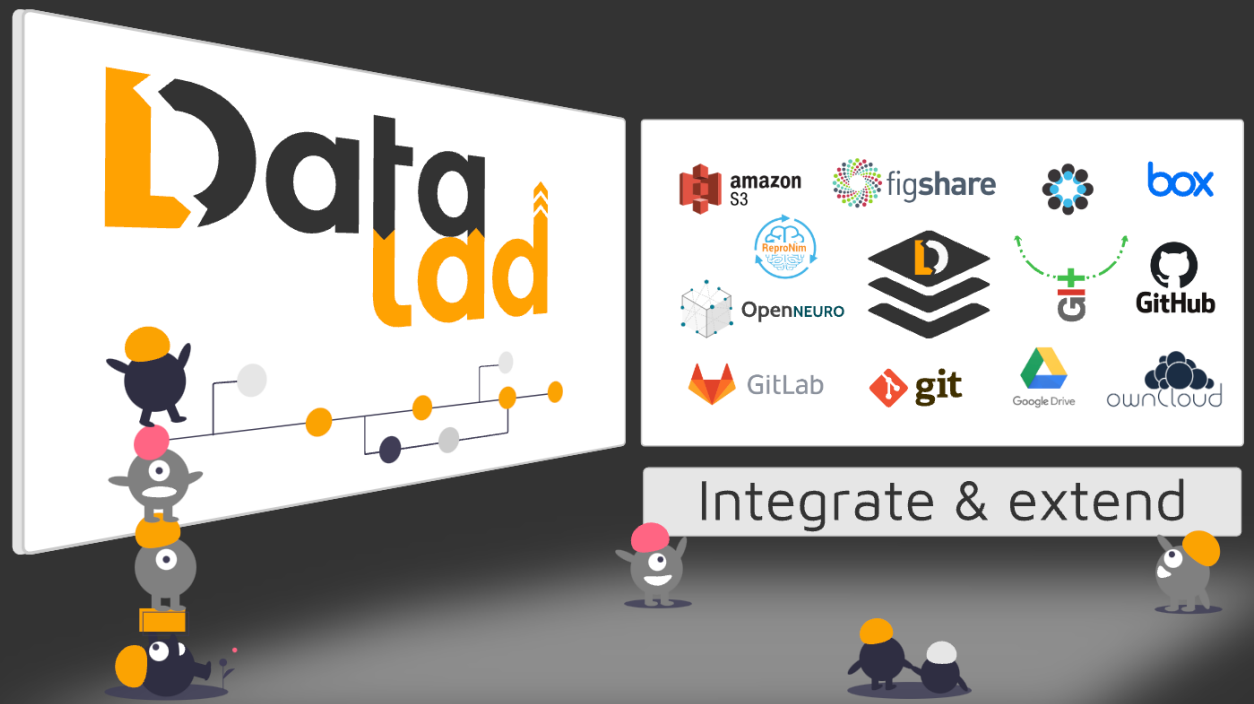
```
$ dvc push
```

Download (macOS)

Watch video
How it works

We're on GitHub ★ 7969

Learn more




# distributed data management

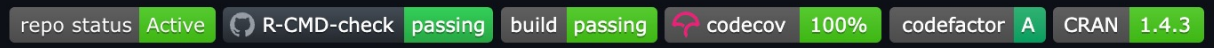
# free and open source

Get DataLad

Star 338

README.md

# orderly




> 1. an attendant in a hospital responsible for the non-medical care of patients and the maintenance of order and cleanliness.
> 2. a soldier who carries orders or performs minor tasks for an officer.

`orderly` is a package designed to help make analysis more reproducible. Its principal aim is to automate a series of basic steps in the process of writing analyses, making it easy to:

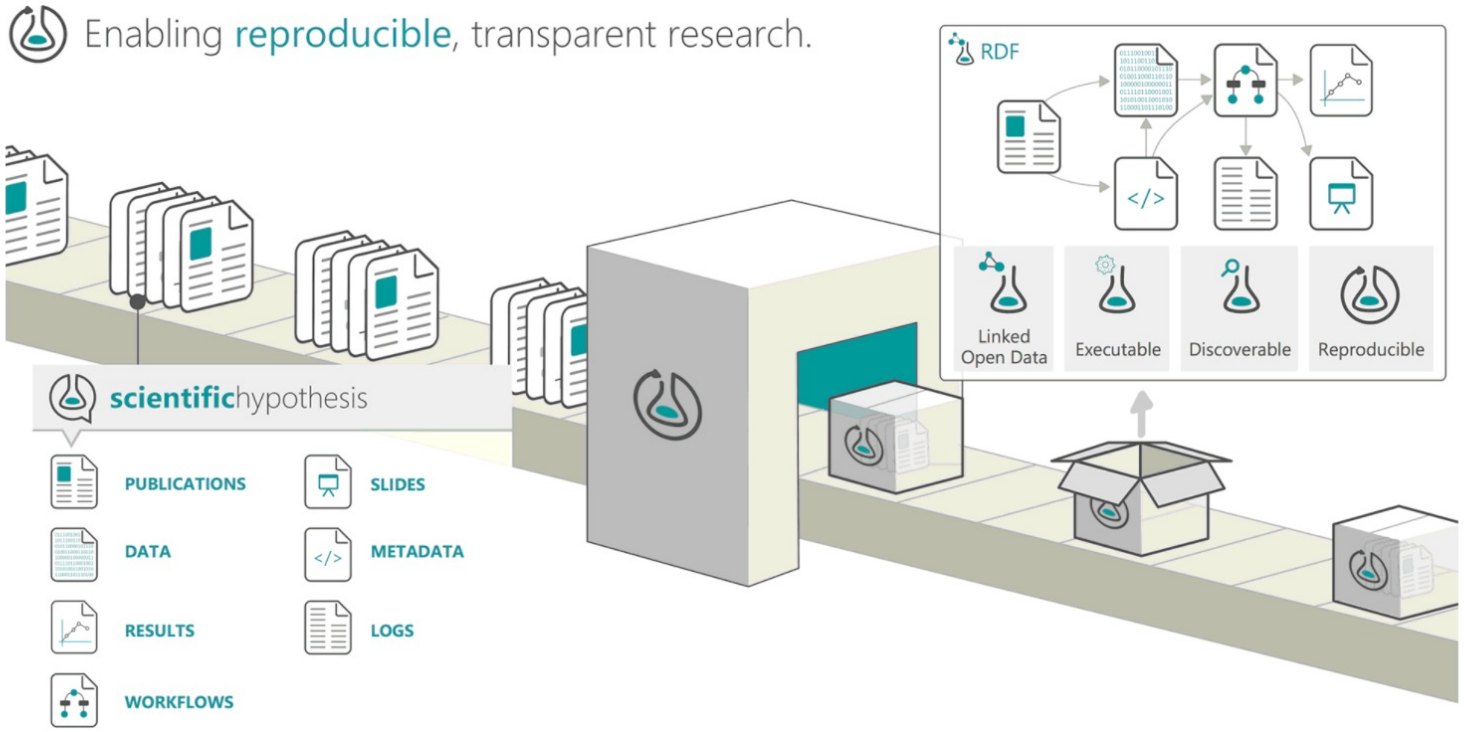
- track all inputs into an analysis (packages, code, and data resources)
- store multiple versions of an analysis where it is repeated
- track outputs of an analysis
- create analyses that depend on the outputs of previous analyses

With `orderly` we have two main hopes:

- analysts can write code that will straightforwardly run on someone else's machine (or a remote machine)
- when an analysis that is run several times starts behaving differently it will be easy to see when the outputs started changing, and what inputs started changing at the same time

`orderly` requires a few conventions around organisation of a project, and after that tries to keep out of your way. However, these requirements are designed to make collaborative development with git easier by minimising conflicts and making backup easier by using an append-only storage system.

## The problem




# Research Object Crate

RO-Crate has been developed as a schema.org-based JSON lightweight approach to the next generation Research Object serialization.

# Open Provenance

## The rationale of PROV☆

Luc Moreau[a,*], Paul Groth[b], James Cheney[c], Timothy Lebo[d], Simon Miles[e]

[a] Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK
[b] Elsevier Labs, Amsterdam, Netherlands
[c] Laboratory for Foundations of Computer Science, University of Edinburgh, Edinburgh EH8 9AB, UK
[d] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA
[e] Department of Informatics, King's College London, Strand, London, WCR2 2LS, UK

ARTICLE INFO

*Article history:*
Received 28 February 2014
Received in revised form
19 March 2015
Accepted 3 April 2015
Available online 20 April 2015

*Keywords:*
Provenance
PROV
Standardization
Requirement
Design decision
Rationale

ABSTRACT

The PROV family of documents are the final output of the World Wide Web Consortium Provenance Working Group, chartered to specify a representation of provenance to facilitate its exchange over the Web. This article reflects upon the key requirements, guiding principles, and design decisions that influenced the PROV family of documents. A broad range of requirements were found, relating to the key concepts necessary for describing provenance, such as resources, activities, agents and events, and to balancing PROV's ease of use with the facility to check its validity. By this retrospective requirement analysis, the article aims to provide some insights into how PROV turned out as it did and why. Benefits of this insight include better inter-operability, a roadmap for alternate investigations and improvements, and solid foundations for future standardization activities.



# PROV-Overview

## An Overview of the PROV Family of Documents

## W3C Working Group Note 30 April 2013

**This version:**
http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
**Latest published version:**
http://www.w3.org/TR/prov-overview/
**Previous version:**
http://www.w3.org/TR/2013/WD-prov-overview-20130312/

**Editors:**
Paul Groth, VU University Amsterdam
Luc Moreau, University of Southampton



## Abstract

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web. This document provides an overview of this family of documents.

# Open Metadata

W3C Recommendation

TABLE OF CONTENTS


1. **Introduction**
2. **Motivation for change**
3. **Namespaces**

3.1 Normative namespaces

3.2 Non-normative namespaces


## Data Catalog Vocabulary (DCAT) - Version 2

W3C Recommendation 04 February 2020


**This version:**
https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/

**Latest published version:**
https://www.w3.org/TR/vocab-dcat-2/

**Latest editor's draft:**
https://w3c.github.io/dxwg/dcat/

**Implementation report:**
https://w3c.github.io/dxwg/dcat-implementation-report/

**Previous version:**
https://www.w3.org/TR/2019/PR-vocab-dcat-2-20191119/

**Previous Recommendation:**
https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/



**Editors:**

Riccardo Albertoni (CNR - Consiglio Nazionale delle Ricerche, Italy)
David Browning (Refinitiv)
Simon Cox (CSIRO)
Alejandra Gonzalez Beltran (Scientific Computing Department, Science and Technology Facilities Council, UK) (Previously at the University of Oxford)
Andrea Perego (European Commission, Joint Research Centre)
Peter Winstanley (Scottish Government)

**Former editors:**

Fadi Maali (DERI)
John Erickson (Tetherless World Constellation (RPI))


## Abstract

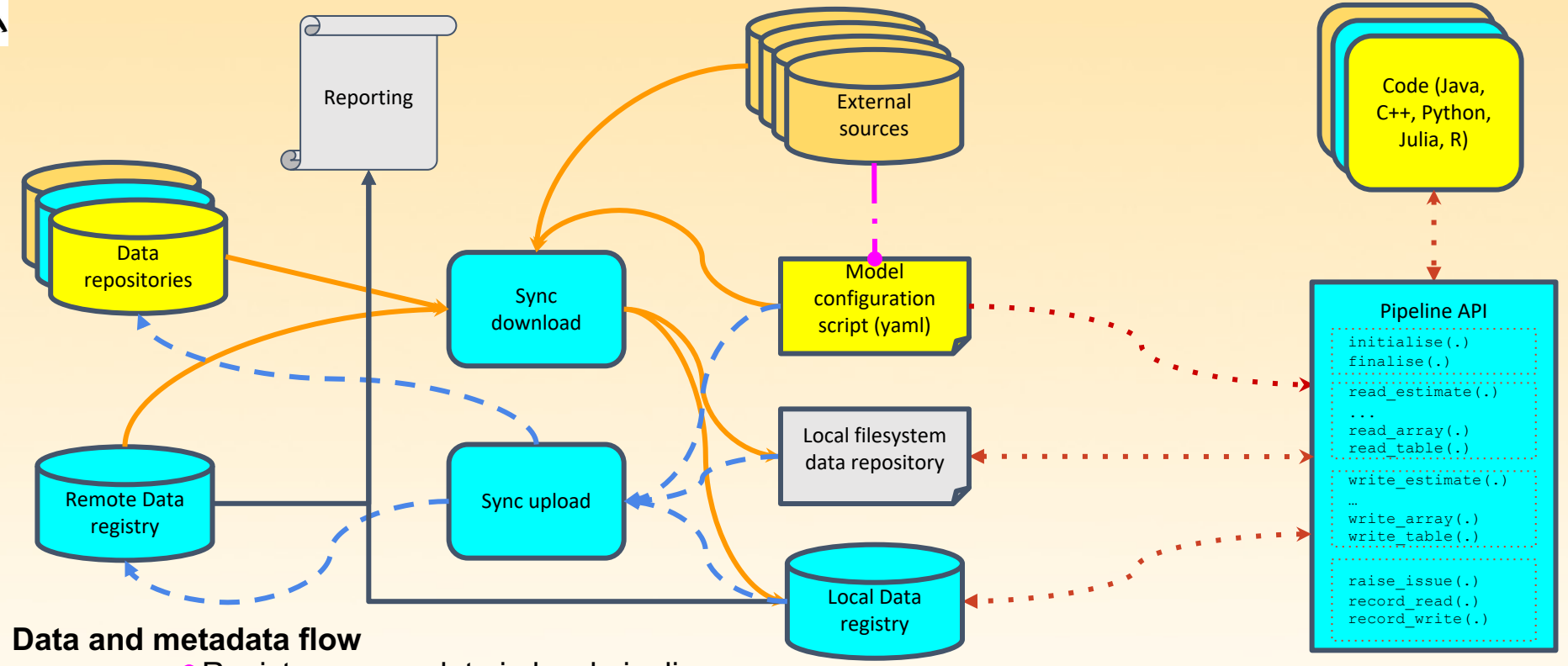

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

DCAT enables a publisher to describe datasets and data services in a catalog using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple catalogs. This can increase the discoverability of datasets and data services. It also makes it possible to have a decentralized approach to publishing data catalogs and makes federated search for datasets across catalogs in multiple sites possible using the same query mechanism and structure. Aggregated DCAT metadata can serve as a manifest file as part of the digital preservation process.

The namespace for DCAT terms is `http://www.w3.org/ns/dcat#`

The suggested prefix for the DCAT namespace is `dcat`

# FAIR Data pipeline



https://www.fairdatapipeline.org/

Universities of Glasgow, Edinburgh, SRUC, Sheffield, Manchester, St Andrews, Stirling, Strathclyde, Heriot-Watt, Southampton, Cambridge, Oxford, Swansea, Bangor, Nottingham, York, Middlesex, City, U. of London, Warwick, Chester, Exeter, King's College London, Loughborough, EPIC, Biomathematics and Statistics Scotland, UK Atomic Energy Authority, Met Office, Software Sustainability Institute, Man Group, Invenia, Horus Security, Red Sift Ltd, and others


*This work was undertaken in part as a contribution to the Rapid Assistance in Modelling the Pandemic (RAMP) initiative, coordinated by the Royal Society*

UKRI Science and Technology Facilities Council