myExperiment Research Objects: Beyond Workflows and Packs

Stian Soiland-Reyes

myGrid, University of Manchester

BOSC 2013, ISMB, Berlin, 2013-07-18

Motivation: Scientific workflows

Coordinated **execution** of *services* and linked *resources*

Dataflow between services

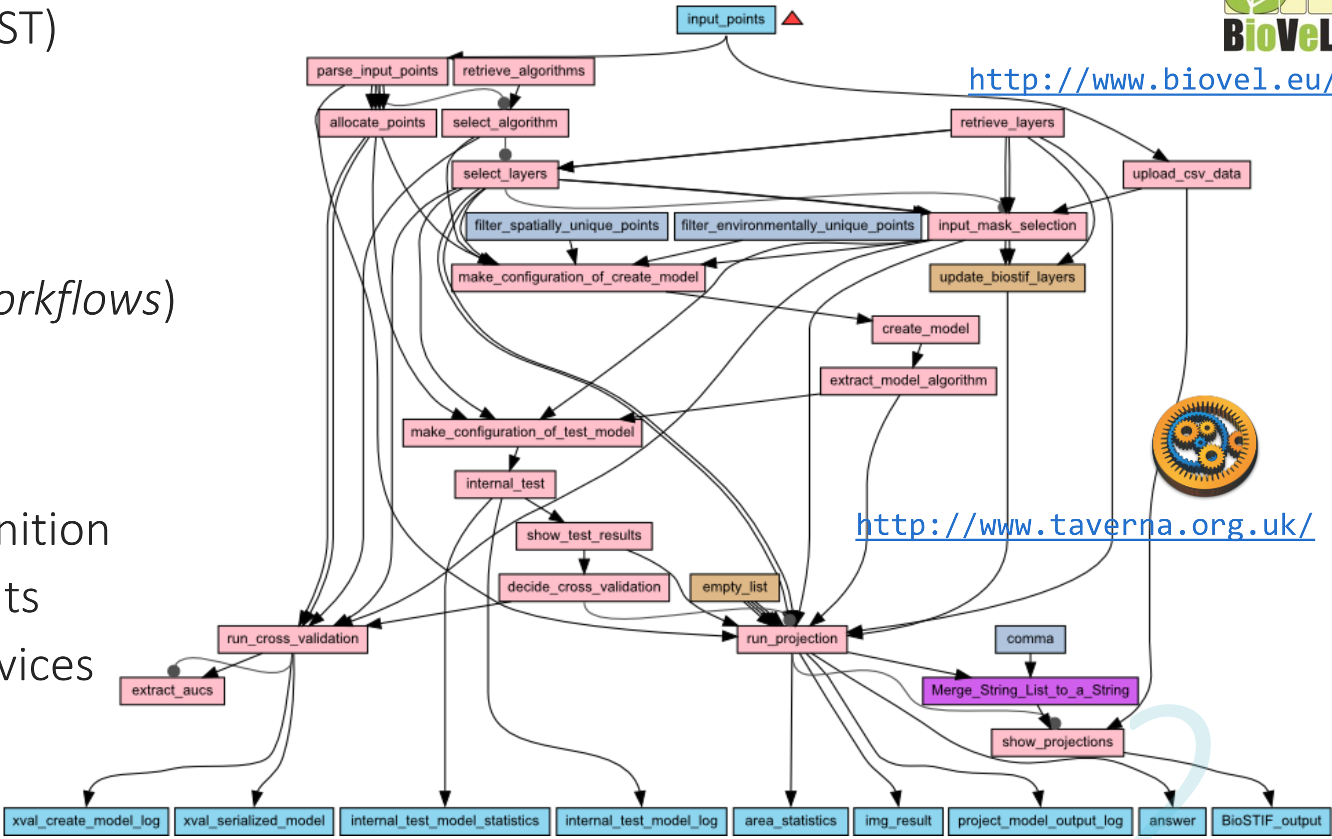
- Web services (SOAP, REST)
- Command line tools
- Scripts
- User interactions
- Components (*nested workflows*)

Method becomes:

- **Documented** visually
- **Shareable** as single definition
- **Reusable** with new inputs
- **Repurposable** other services
- **Reproducible**?

http://www.myexperiment.org/workflows/3355

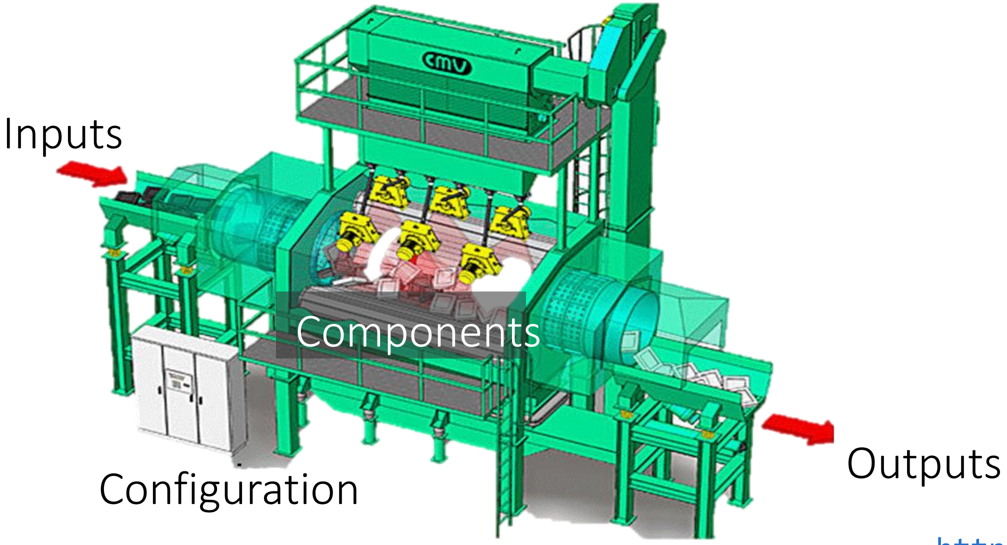
http://www.biovel.eu/

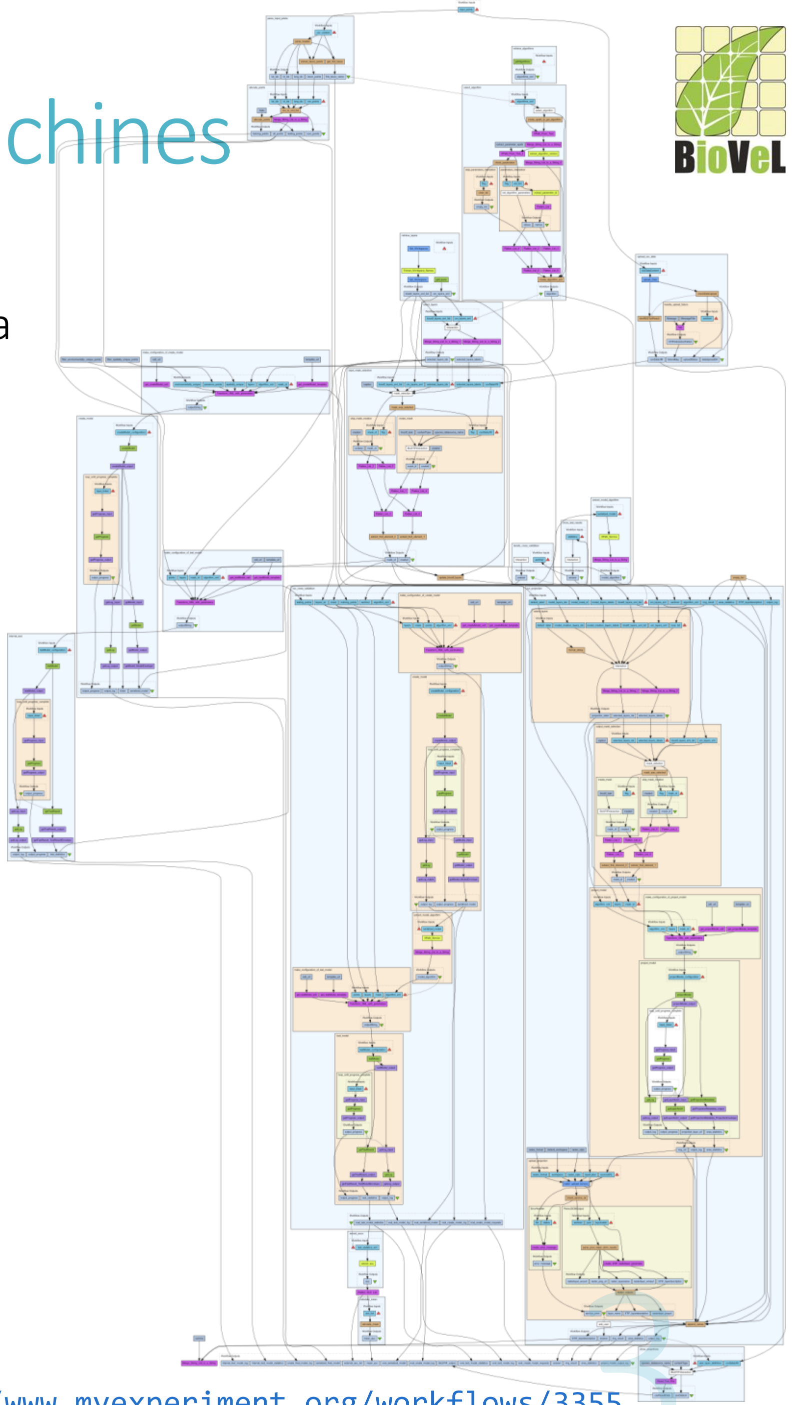


http://www.taverna.org.uk/

But workflows are complex machines

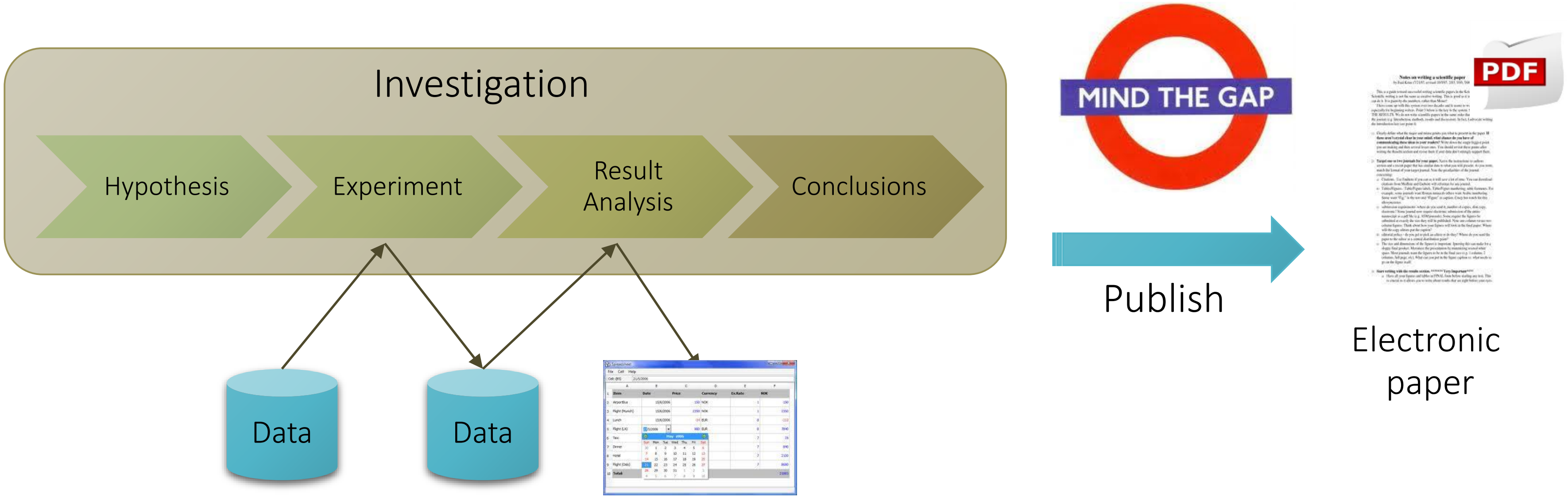
- Will it **still work** after a year? 10 years?
- Expanding *components*, we see a workflow involves a series of specific **tools and services** which
  - **Depend** on datasets, software libraries, other tools
  - Are often poorly **described** or **understood**
  - Over time *evolve*, ***change***, *break* or are *replaced*
- **User interactions** are not reproducible
  - But can be *tracked* and *replayed*





http://www.myexperiment.org/workflows/3355

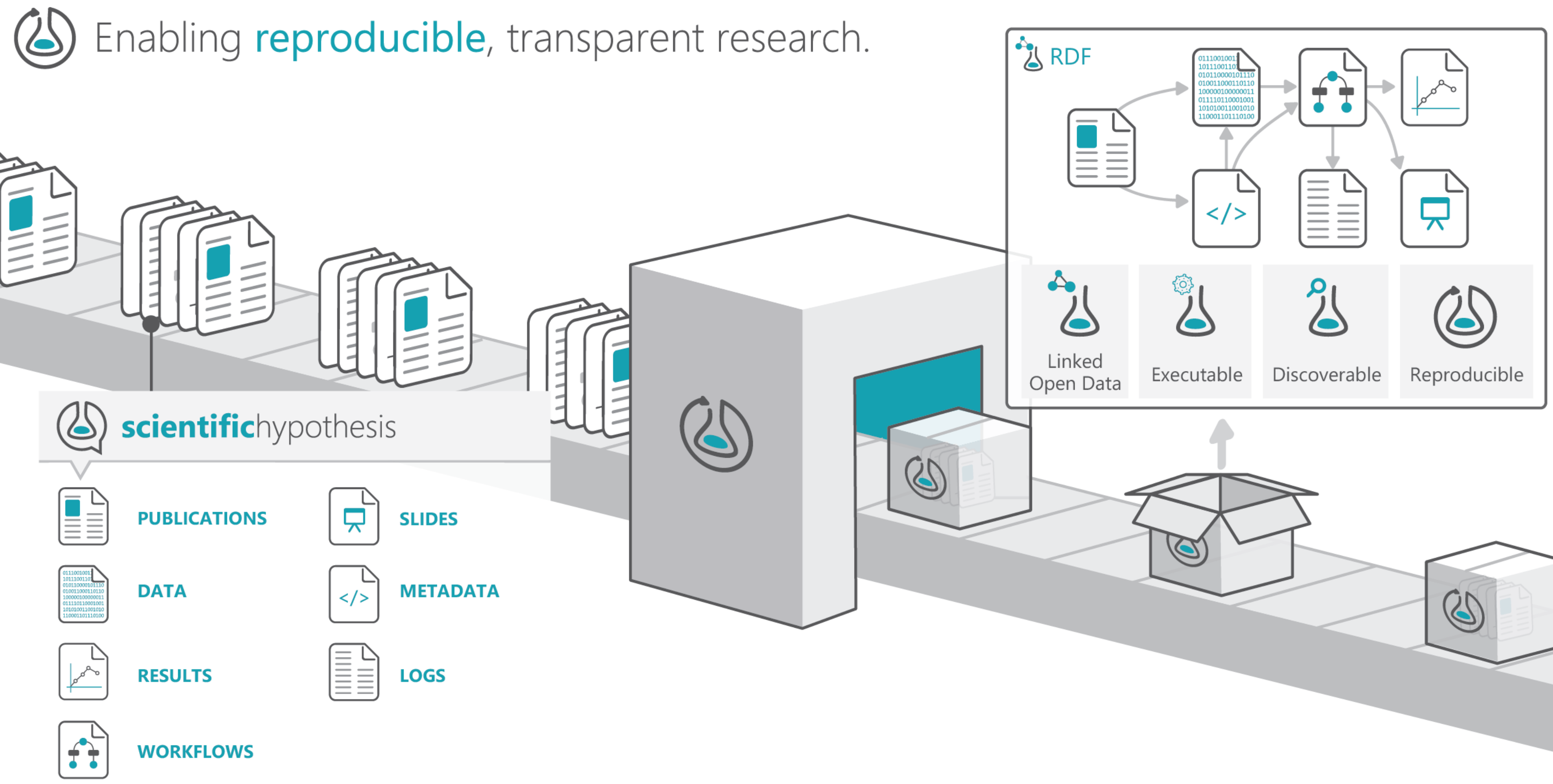
Electronic Paper Not Enough



Open Research movement: Openly share the data of your experiments

http://figshare.com/ http://datadryad.org/ http://www.force11.org/beyondthepdf2

RESEARCH OBJECT (RO)



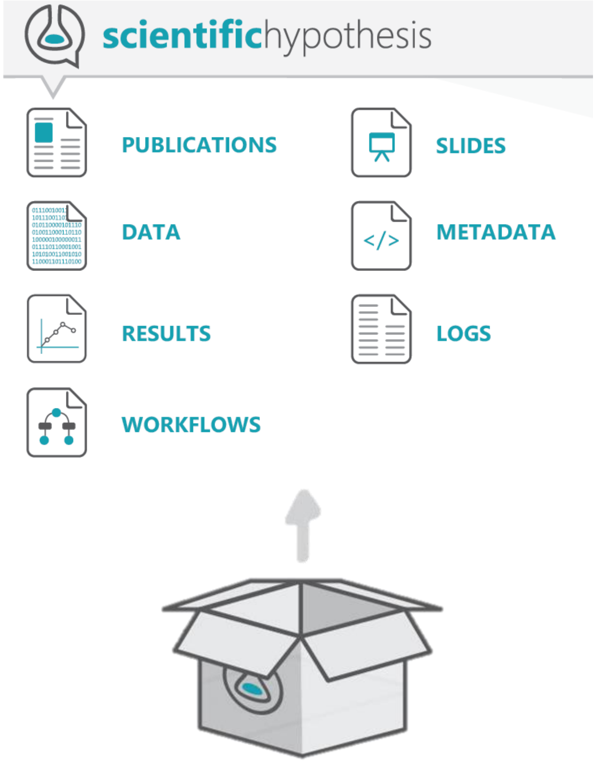
Research objects goal: Openly share *everything* about your experiments, including how those things are related

http://www.researchobject.org/

What is in a research object?

A Research Object **bundles** and **relates** digital resources of a scientific experiment or investigation:

- **Data** used and results produced in experimental study
- **Methods** employed to produce and analyse that data
- **Provenance** and settings for the experiments
- **People** involved in the investigation
- **Annotations** about these resources, that are essential to the understanding and interpretation of the scientific outcomes captured by a research object



http://www.researchobject.org/

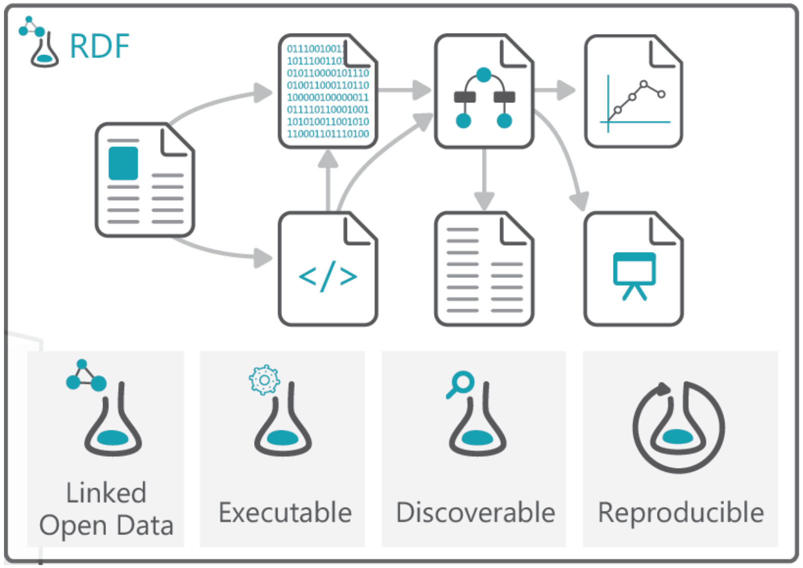
Gathering everything

Research Objects (RO) *aggregate* related resources, their *provenance* and *annotations*

Conveys “everything you need to know” about a study/experiment/analysis/dataset/workflow

Shareable, **evolvable**, contributable, **citable**

ROs have their own provenance and lifecycles



Why Research Objects?

i. To **share** your research materials
(RO as a social object)

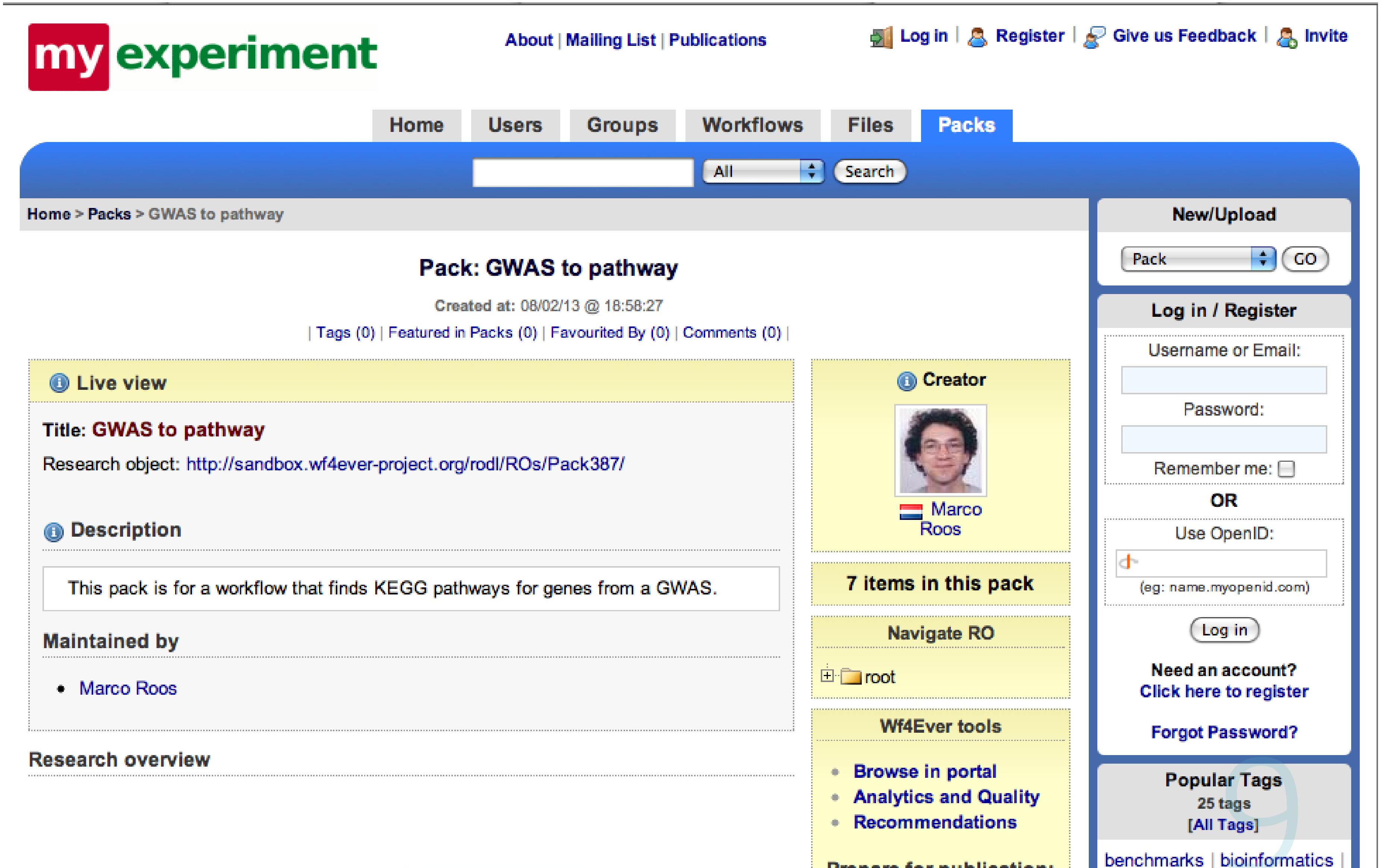
ii. To facilitate **reproducibility** and **reuse** of methods

iii. To be **recognized** and **cited**
(even for constituent resources)

iv. To **preserve** results and **prevent decay**
(curation of workflow definition;
using provenance for partial rerun)

A Research object

http://alpha.myexperiment.org/packs/387



About | Mailing List | Publications

Log in | Register | Give us Feedback | Invite

Home | Users | Groups | Workflows | Files | Packs

All | Search

Home > Packs > GWAS to pathway

Pack: GWAS to pathway

Created at: 08/02/13 @ 18:58:27

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

Live view

Title: GWAS to pathway

Research object: http://sandbox.wf4ever-project.org/rodl/ROs/Pack387/

Description

This pack is for a workflow that finds KEGG pathways for genes from a GWAS.

Maintained by

- Marco Roos

Research overview

Creator

Marco Roos

7 items in this pack

Navigate RO

root

Wf4Ever tools

- Browse in portal
- Analytics and Quality
- Recommendations

Prepare for publication

New/Upload

Pack | GO

Log in / Register

Username or Email:

Password:

Remember me: □

OR

Use OpenID:

(eg: name.myopenid.com)

Log in

Need an account?

Click here to register

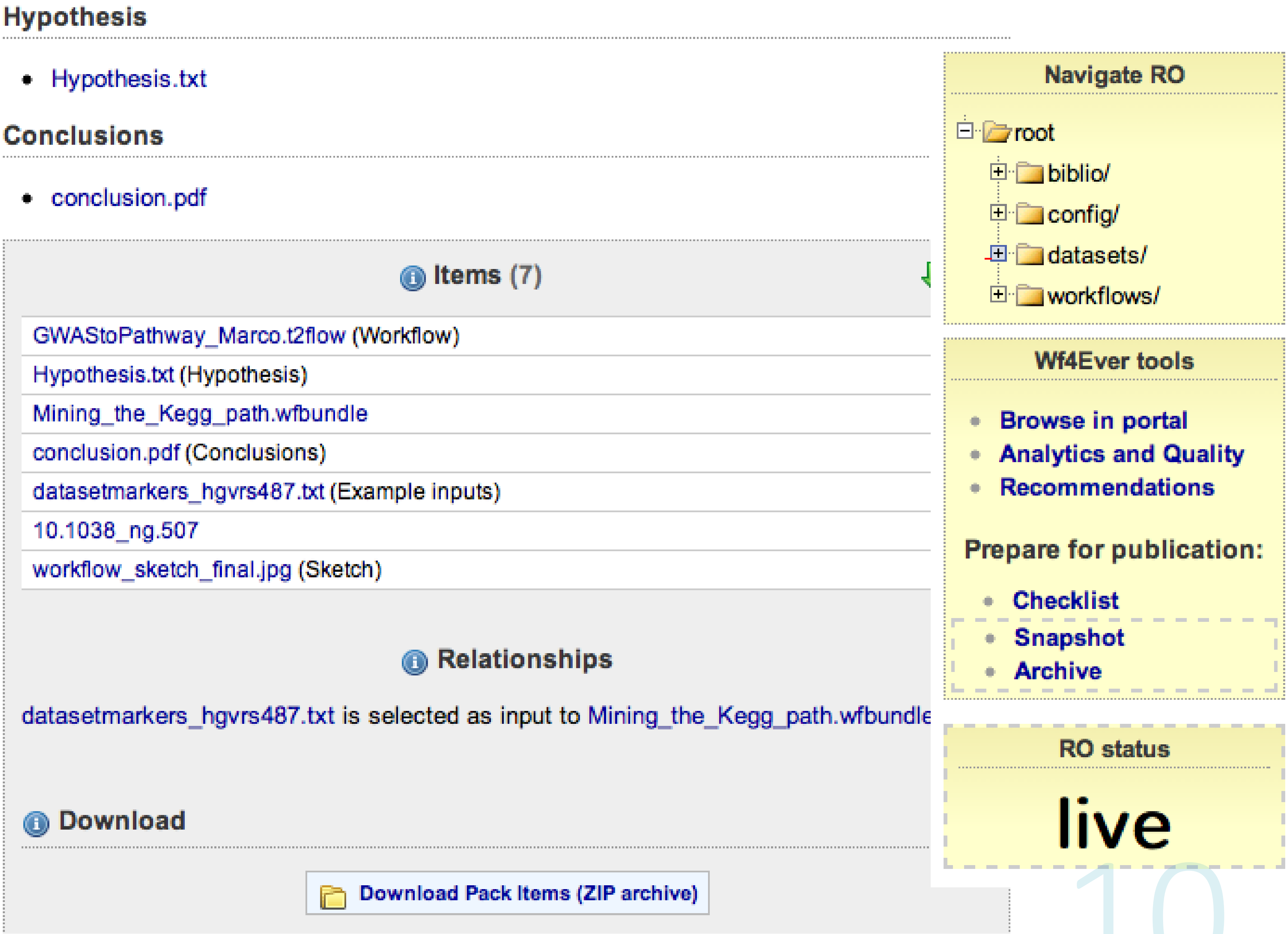
Forgot Password?

Popular Tags

25 tags

[All Tags]

benchmarks | bioinformatics |



Hypothesis

- Hypothesis.txt

Conclusions

- conclusion.pdf

Items (7)

GWAStoPathway_Marco.t2flow (Workflow)

Hypothesis.txt (Hypothesis)

Mining_the_Kegg_path.wfbundle

conclusion.pdf (Conclusions)

datasetmarkers_hgvrs487.txt (Example inputs)

10.1038_ng.507

workflow_sketch_final.jpg (Sketch)

Relationships

datasetmarkers_hgvrs487.txt is selected as input to Mining_the_Kegg_path.wfbundle

Download

Download Pack Items (ZIP archive)

Navigate RO

- root
  - biblio/
  - config/
  - datasets/
  - workflows/

Wf4Ever tools

- Browse in portal
- Analytics and Quality
- Recommendations

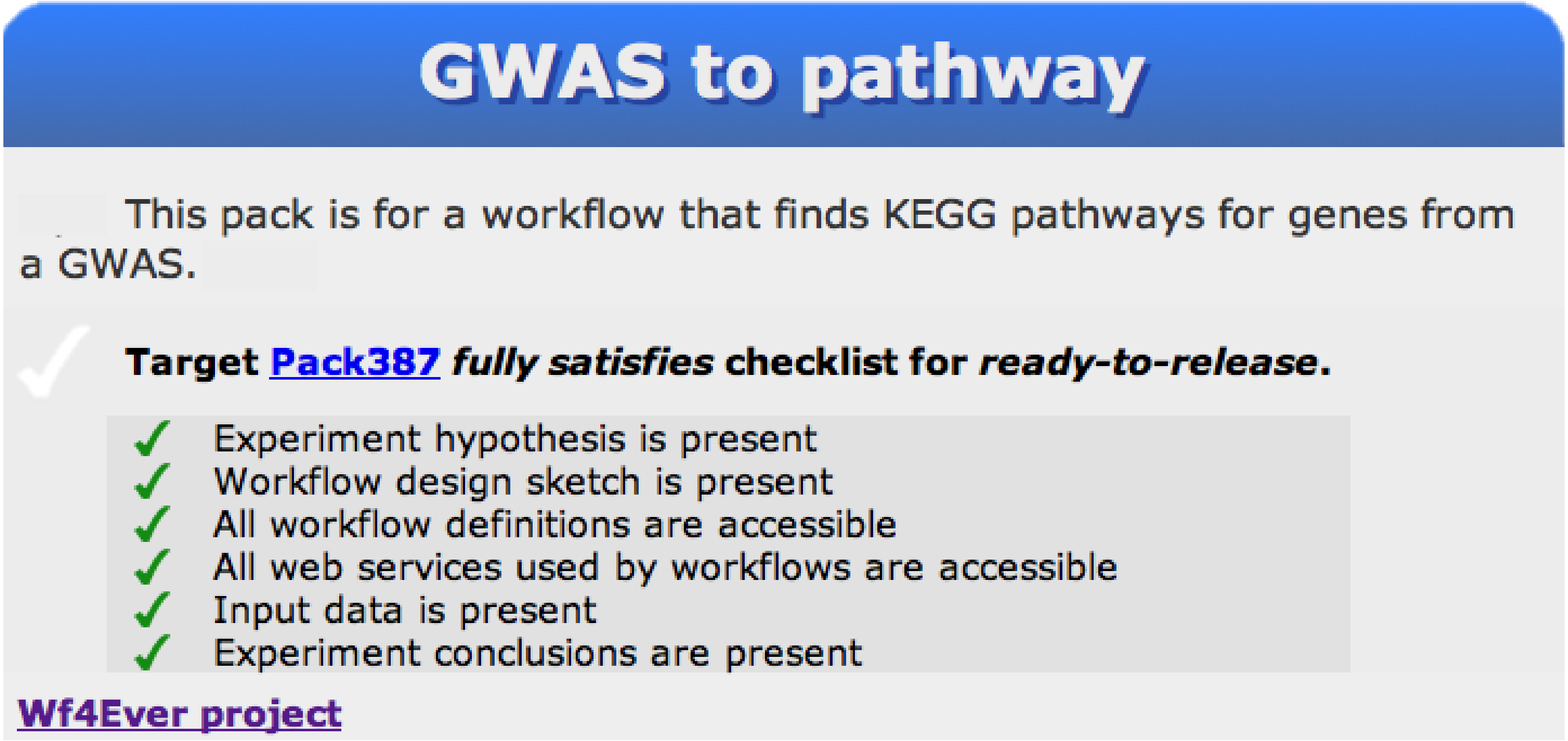
Prepare for publication:

- Checklist
- Snapshot
- Archive

RO status

live

Quality Assessment of a research object



GWAS to pathway

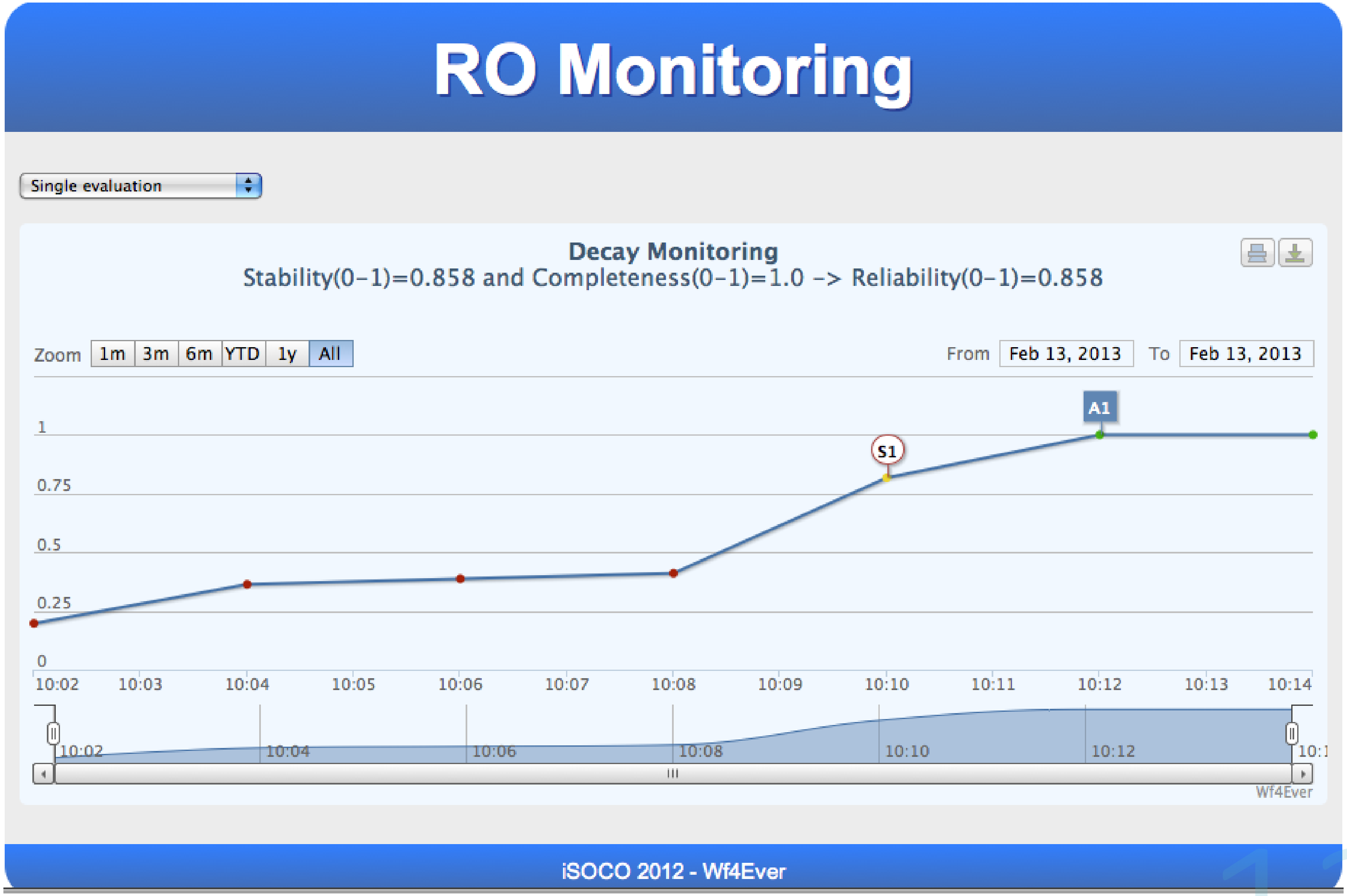
This pack is for a workflow that finds KEGG pathways for genes from a GWAS.

✓ **Target Pack387 *fully satisfies* checklist for *ready-to-release*.**

- ✓ Experiment hypothesis is present
- ✓ Workflow design sketch is present
- ✓ All workflow definitions are accessible
- ✓ All web services used by workflows are accessible
- ✓ Input data is present
- ✓ Experiment conclusions are present

Wf4Ever project

Quality Monitoring



RO Monitoring

Single evaluation

Decay Monitoring

Stability(0-1)=0.858 and Completeness(0-1)=1.0 -> Reliability(0-1)=0.858

iSOCO 2012 - Wf4Ever

Annotations in research objects

Types: “This document contains an *hypothesis*”

Relations: “These datasets are *consumed* by that tool”

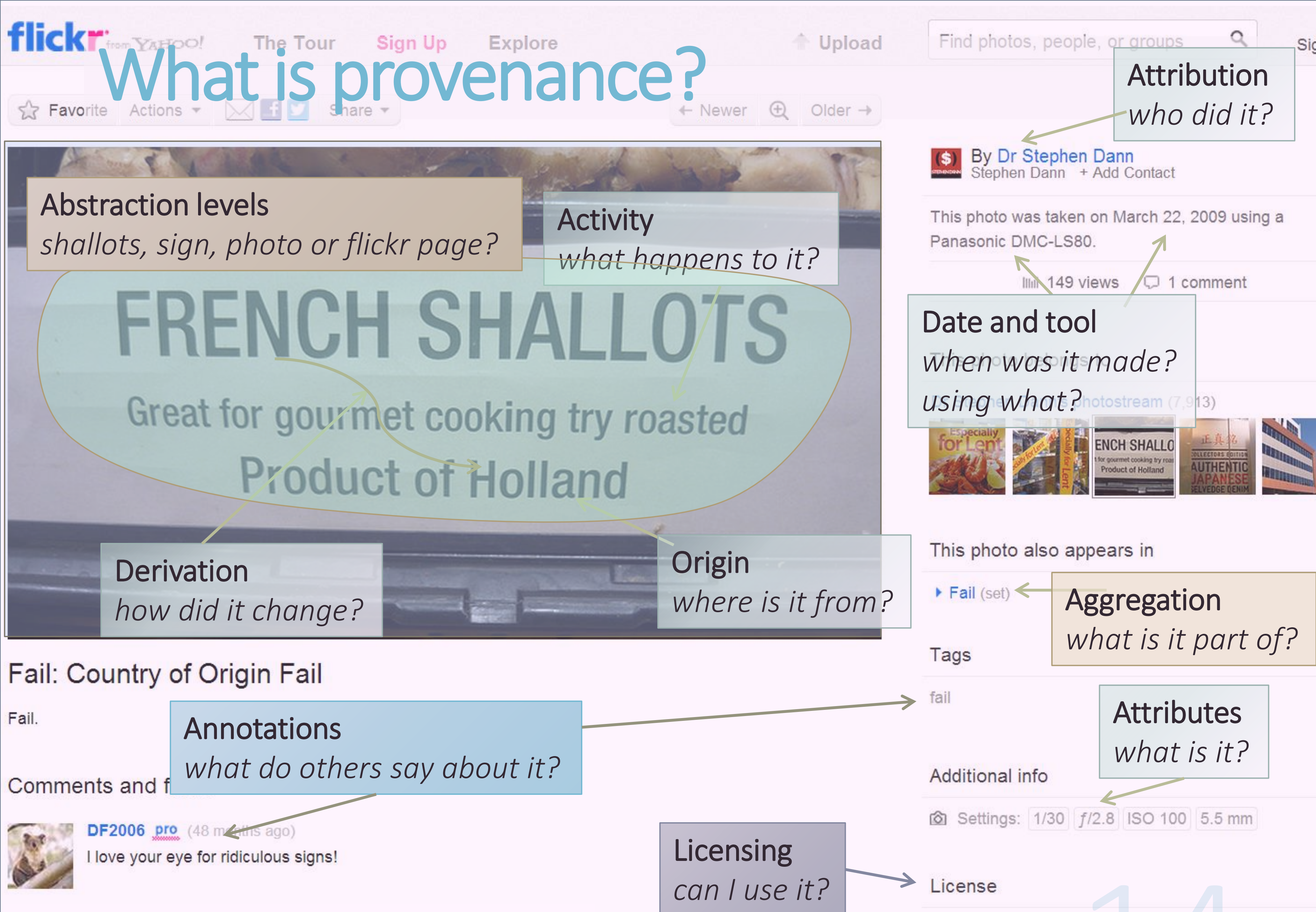
Provenance: “These results *came from* this workflow run”

Descriptions: “*Purpose* of this step is to filter out invalid data”

Comments: “This method looks useful, but how do I install it?”

Examples: “This is how you could use it”

What is provenance?



Attribution
who did it?

By Dr Stephen Dann
Stephen Dann + Add Contact

This photo was taken on March 22, 2009 using a Panasonic DMC-LS80.

149 views 1 comment

Abstraction levels
shallots, sign, photo or flickr page?

Activity
what happens to it?

Date and tool
when was it made? using what?

Derivation
how did it change?

Origin
where is it from?

This photo also appears in

Fail (set)

Aggregation
what is it part of?

Fail: Country of Origin Fail

Fail.

Annotations
what do others say about it?

Comments and f

DF2006 pro (48 months ago)
I love your eye for ridiculous signs!

Tags

fail

Attributes
what is it?

Additional info

Settings: 1/30 ƒ/2.8 ISO 100 5.5 mm

Licensing
can I use it?

License

Some rights reserved

Request to license Dr Stephen Dann's photos via Getty Images

By Dr Stephen Dann
licensed under Creative Commons Attribution-ShareAlike 2.0 Generic
http://www.flickr.com/photos/stephendann/3375055368/

Attribution

Who *collected* this sample? Who helped?

Which lab performed the *sequencing*?

Who did the *data analysis*?

Who wrote the *analysis* workflow?

Who made the *data set* used by analysis?

Who *curated* the results?

Why do I need this?

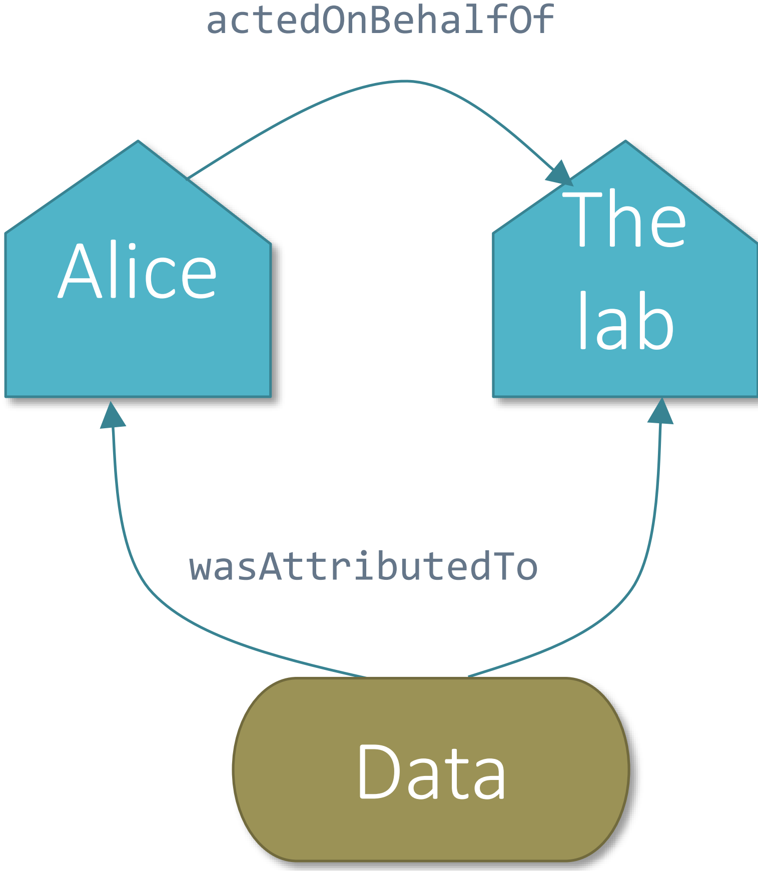
i. To be **recognized** for my work

ii. Who should I give **credits** to?

iii. Who should I **complain** to?

iv. Can I **trust** them?

v. Who should I make **friends** with?



Derivation

Which sample was this metagenome sequenced from?

Which meta-genomes was this sequence extracted from?

Which sequence was the basis for the results?

What is the previous revision of the new results?

Why do I need this?

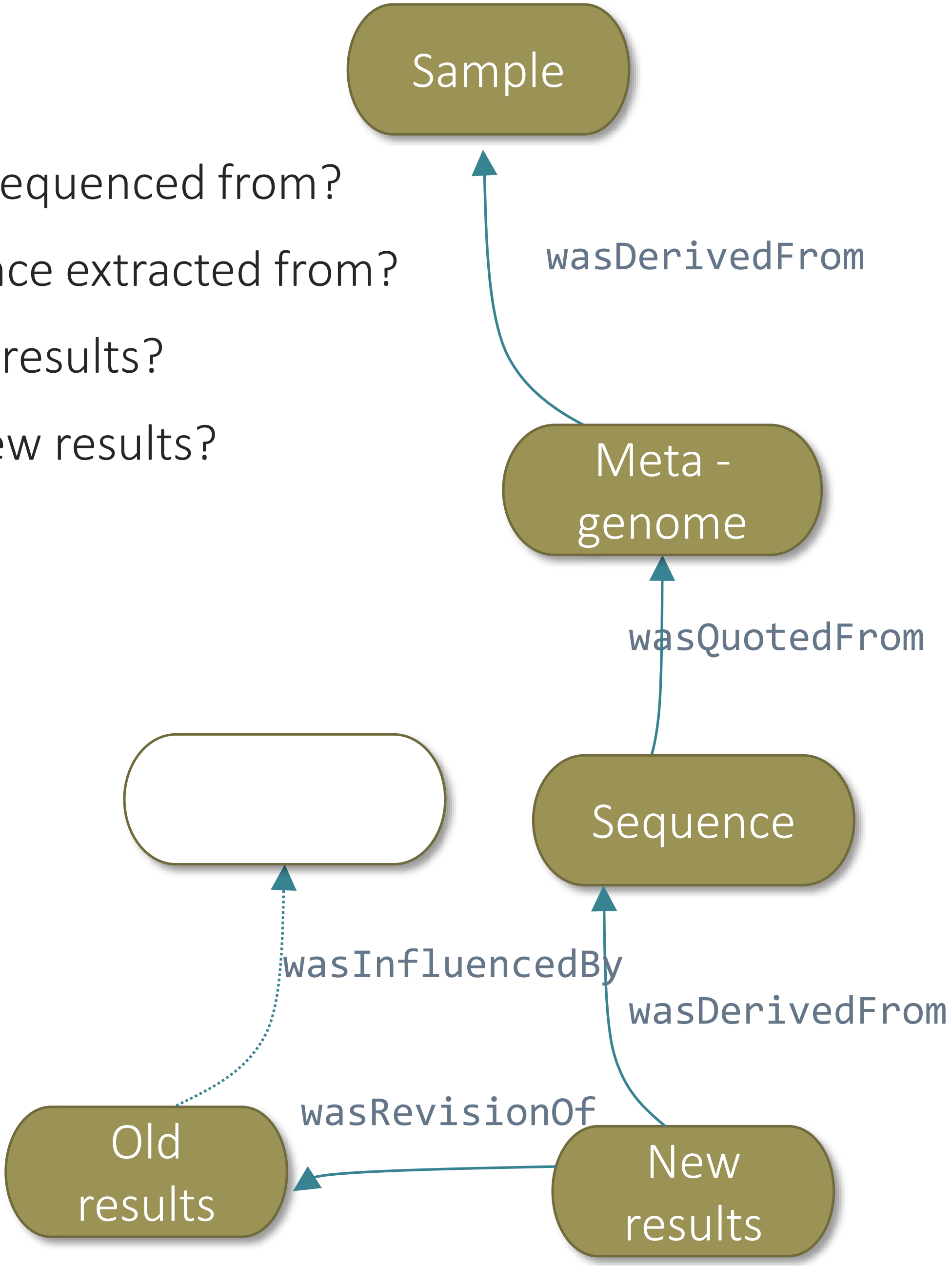
i. To **verify** consistency (did I use the correct sequence?)

ii. To find the latest **revision**

iii. To **backtrack** where a diversion appeared after a change

iv. To **credit** work I depend on

v. **Auditing** and defence for peer review



Activities

What happened? When? Who?

What was used and generated?

Why was this workflow started?

Which workflow ran? Where?

Why do I need this?

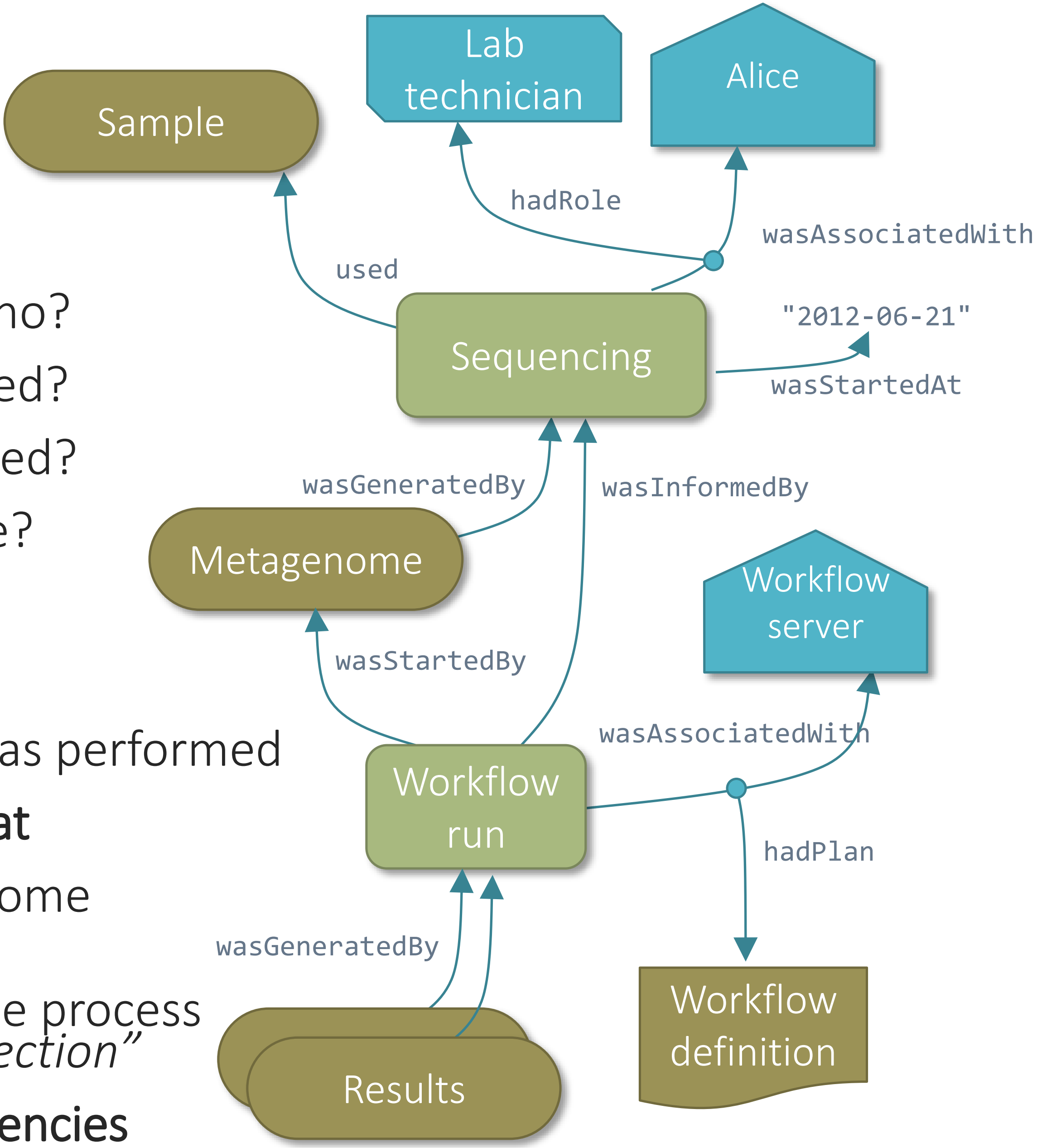
i. To see which **analysis** was performed

ii. To find out **who** did **what**

iii. What was the metagenome **used** for?

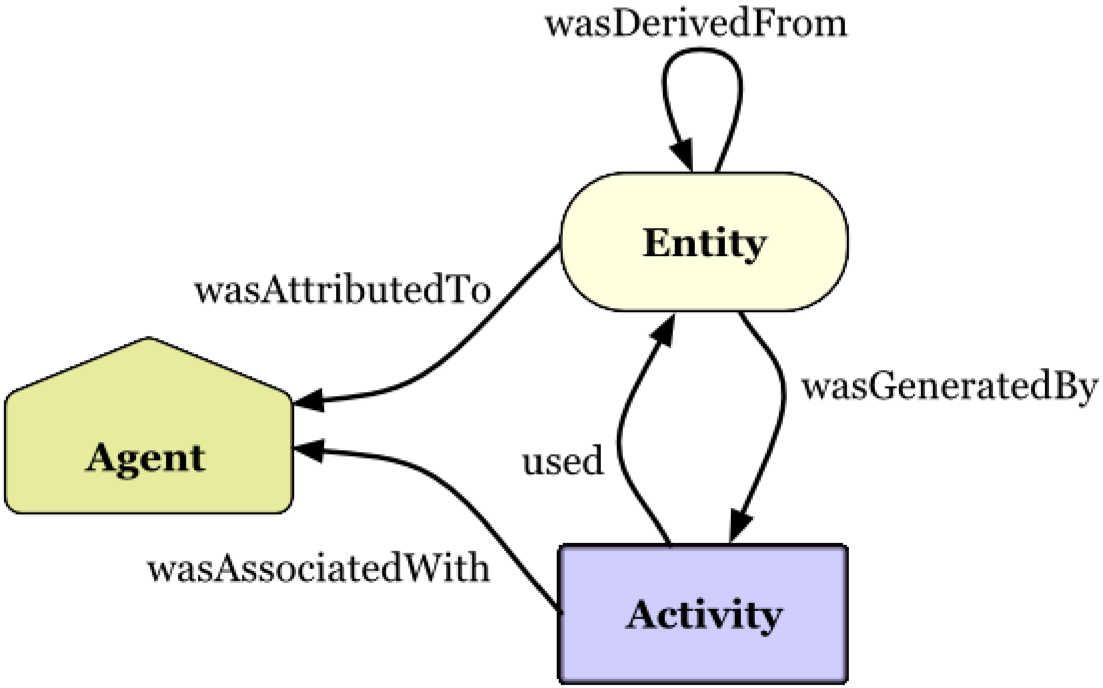
iv. To **understand** the whole process *"make me a Methods section"*

v. To track down **inconsistencies**



Core PROV model

Provenance Working Group



http://www.w3.org/TR/prov-primer/

Provenance of what?

Who **made** the (content of) research object? Who **maintains** it?

Who **wrote** this document? Who **uploaded** it?

Which CSV was this Excel file **imported from**?

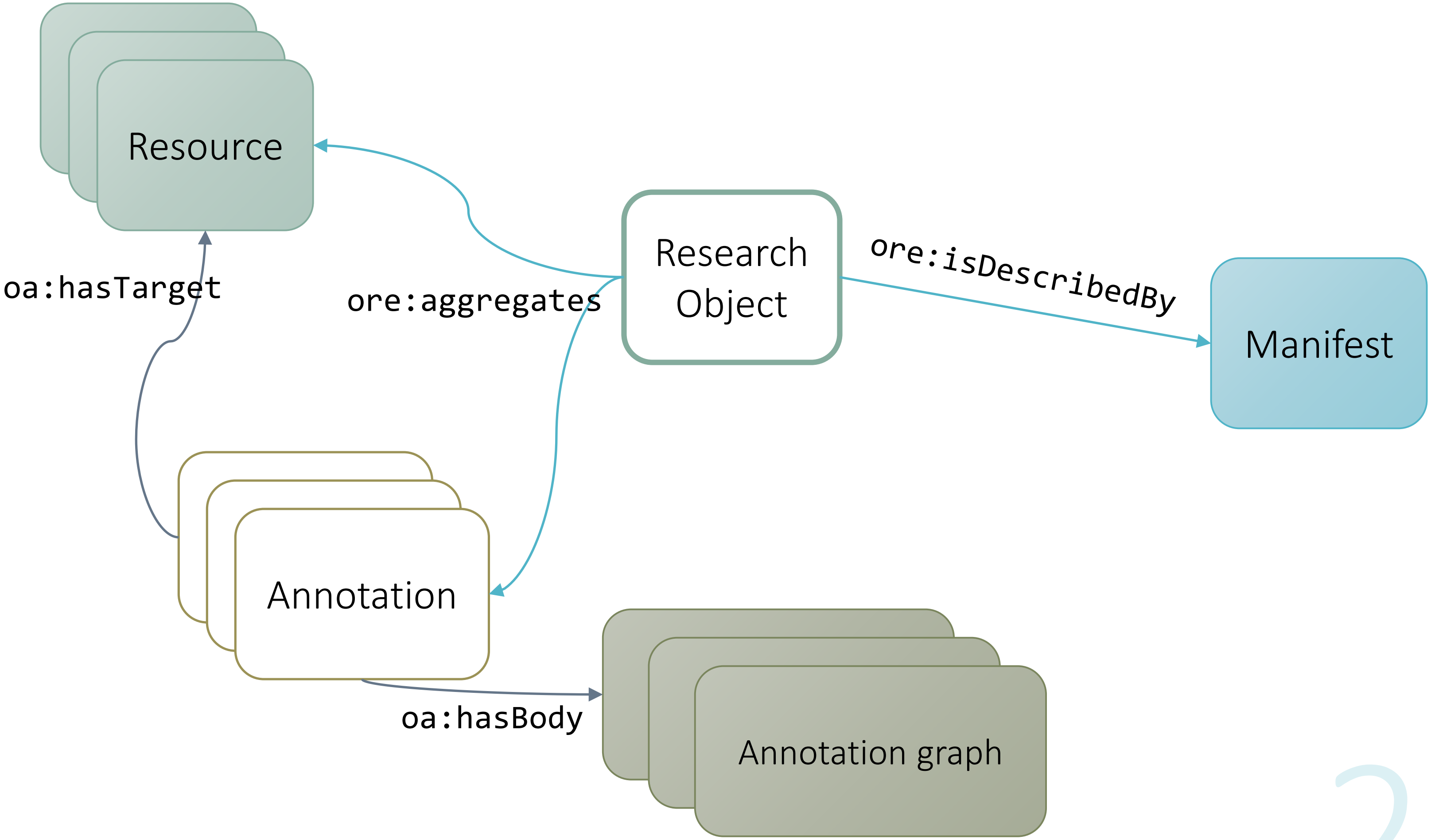
Who wrote this **description**? When? How did we get it?

What is the **state** of this RO? (Live or Published?)

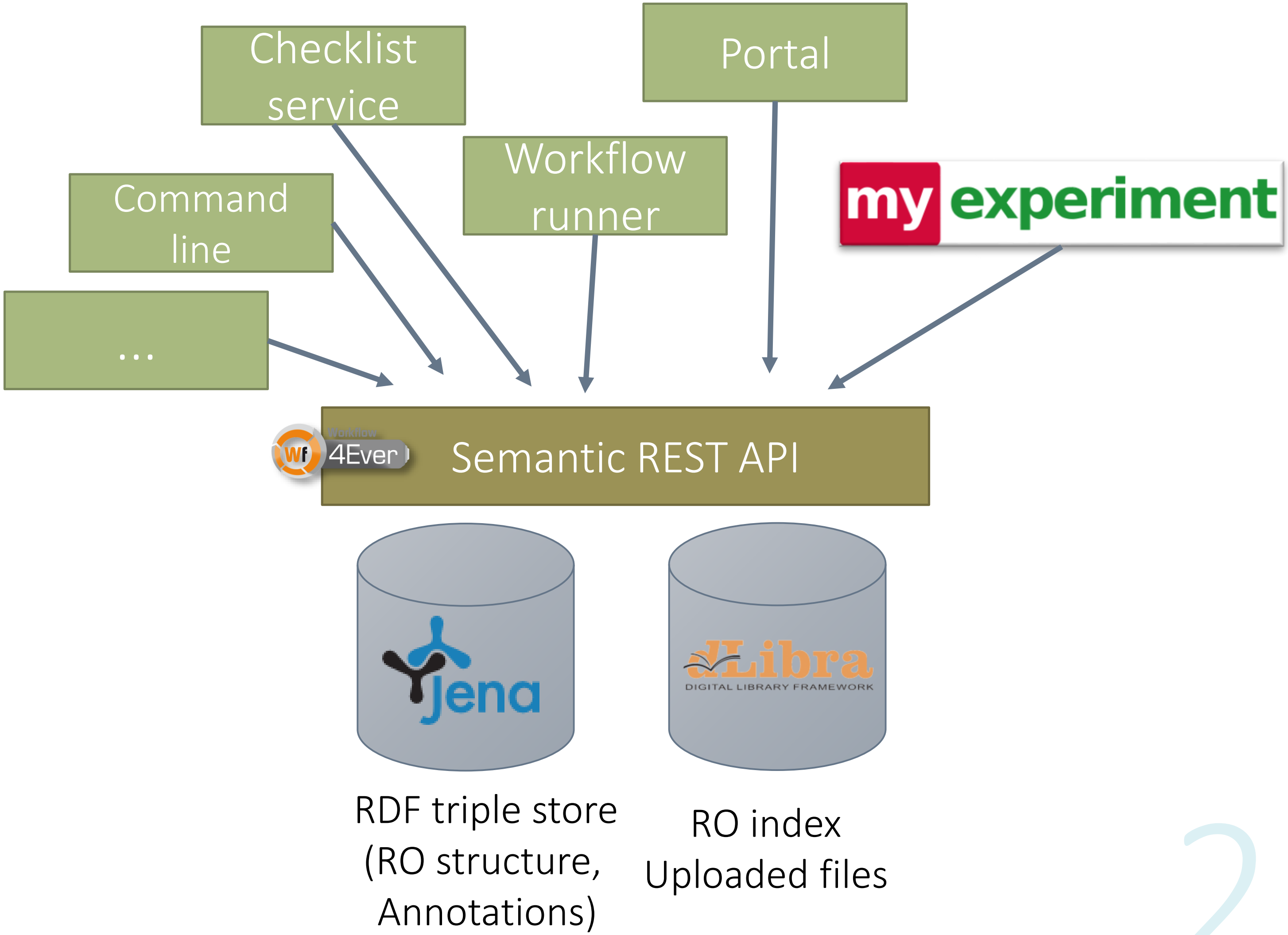
What did the research object look like before? (**Revisions**) – are there newer versions?

Which research objects are **derived** from this RO?

Research object model at a glance



Wf4Ever architecture


Checklist service

Portal

Workflow runner

Command line

...

myexperiment

Wf 4Ever

Semantic REST API

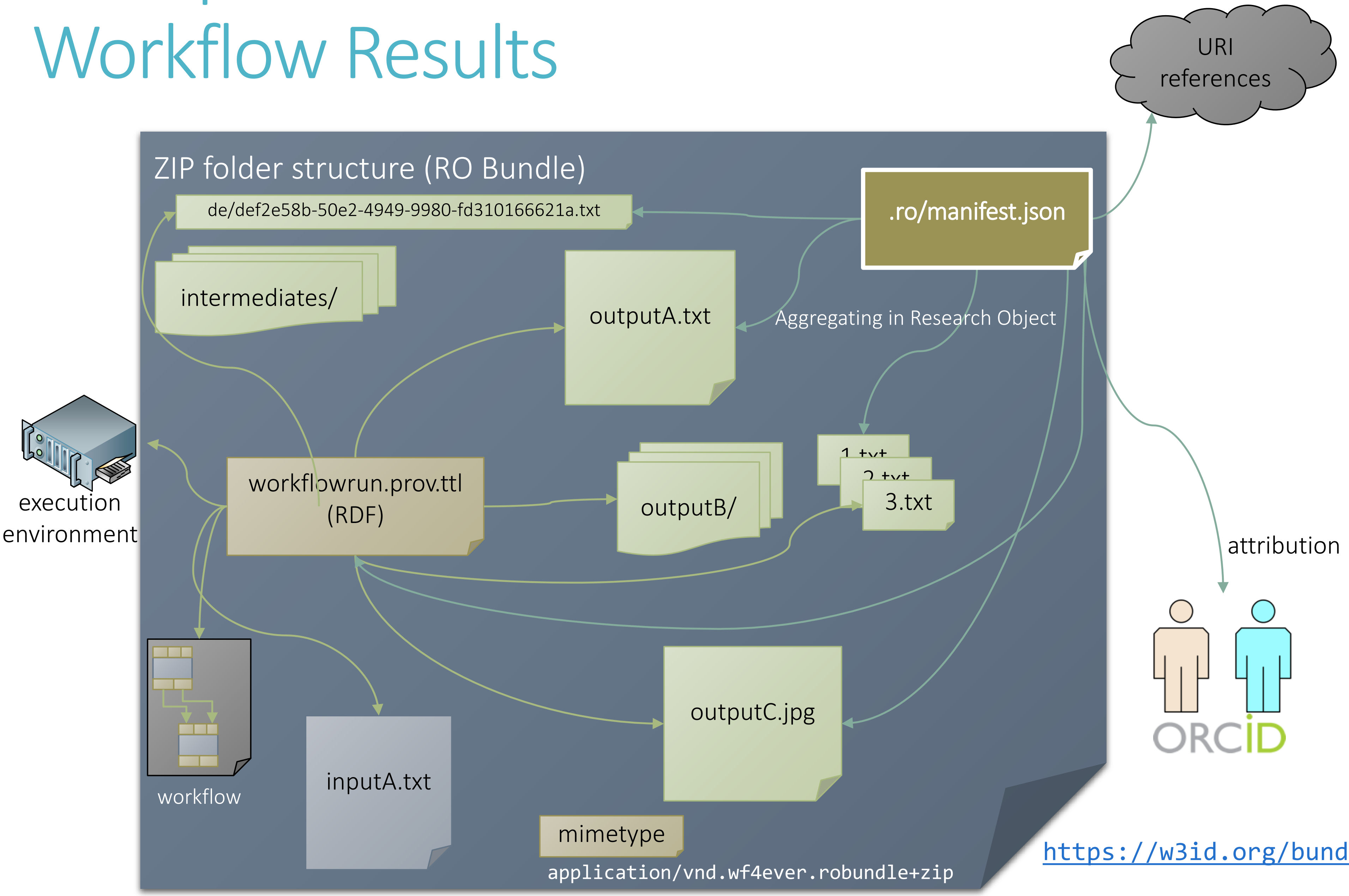
RDF triple store (RO structure, Annotations)

RO index
Uploaded files

Saving a research object: RO bundle

- Single, **transferrable** research object
  - Self-contained **snapshot**
  - Which files in ZIP, which are URIs? (Up to user/application)
- Regular **ZIP file**, explored and unpacked with standard tools
- **JSON manifest** is programmatically accessible without RDF understanding
- Works **offline** and in desktop applications – no REST API access required
- Basis for RO-enabled **file formats**, e.g. Taverna run bundle
- **Exchanged** with myExperiment and RO tools

Example RO bundle: Workflow Results



URI references

ZIP folder structure (RO Bundle)

de/def2e58b-50e2-4949-9980-fd310166621a.txt

.ro/manifest.json

intermediates/

outputA.txt

Aggregating in Research Object

execution environment

workflowrun.prov.ttl (RDF)

outputB/

1.txt

2.txt

3.txt

attribution

ORCID

workflow

inputA.txt

outputC.jpg

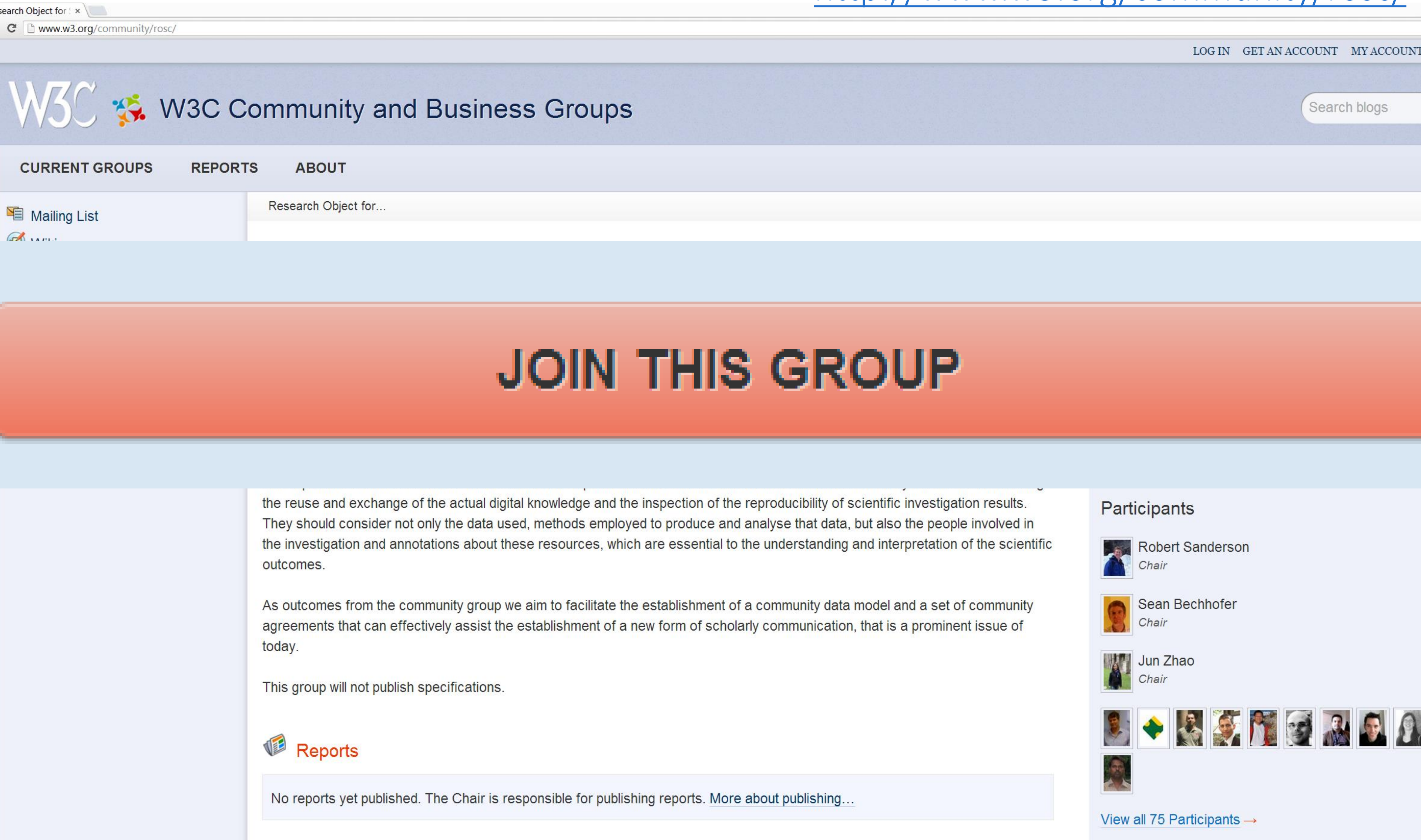
mimetype

application/vnd.wf4ever.robundle+zip

https://w3id.org/bund

W3C community group for RO

http://www.w3.org/community/rosc/



JOIN THIS GROUP

the reuse and exchange of the actual digital knowledge and the inspection of the reproducibility of scientific investigation results. They should consider not only the data used, methods employed to produce and analyse that data, but also the people involved in the investigation and annotations about these resources, which are essential to the understanding and interpretation of the scientific outcomes.

As outcomes from the community group we aim to facilitate the establishment of a community data model and a set of community agreements that can effectively assist the establishment of a new form of scholarly communication, that is a prominent issue of today.

This group will not publish specifications.

Reports

No reports yet published. The Chair is responsible for publishing reports. More about publishing…

Participants

Robert Sanderson
Chair

Sean Bechhofer
Chair

Jun Zhao
Chair

View all 75 Participants →

Summary

Provide provenance records: Use (and extend) PROV.

Share your methods (tools, workflows), not just your data

Make research objects: Collect resources and relate them

Pick your RO integration level to adapt:

1. *Conceptual model*
2. *RO Core ontology (aggregation and annotation)*
3. *Workflow description and provenance*
4. *Wf4Ever architecture for APIs*
5. *myExperiment as user interface*