MANCHESTER 1824
The University of Manchester

PSNC

Data Archiving and Networked Services
DANS

Welcome

researchobject 2018

Twitter hashtags: #eScience2018 #ResearchObject

Sponsored by

bioexcel
Center of Excellence for Computational Biomolecular Research

Centre of Excellence for Computational Biomolecular Research; H2020 grant 675728.

researchobject.org

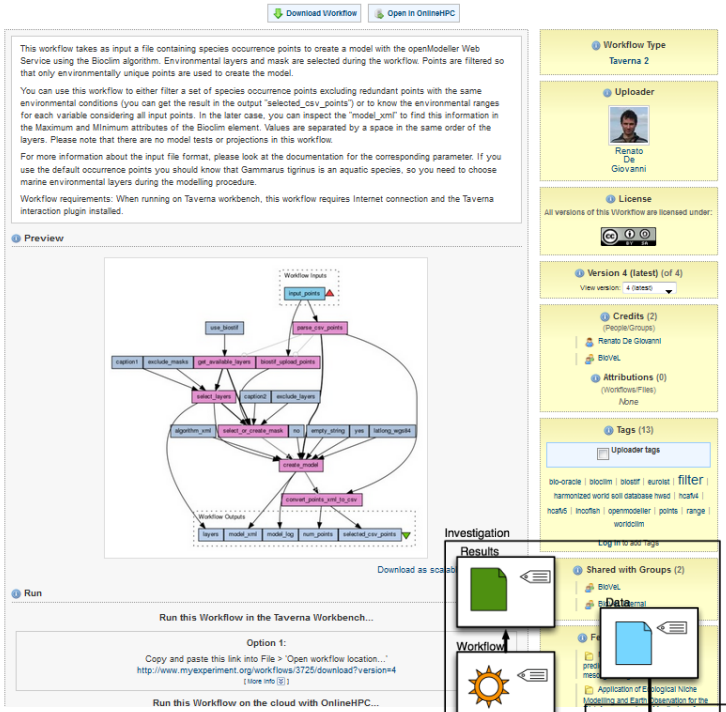
Research Object Community Update

Carole Goble, Stian Soiland-Reyes, Sean Bechhofer
The University of Manchester, UK
carole.goble@manchester.ac.uk

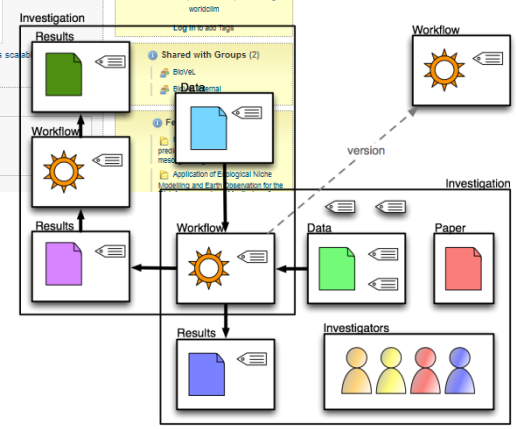
RO2018, 29 October 2018, Amsterdam, 2018.
Satellite workshop of IEEE 14th International Conference on e-Science 2018

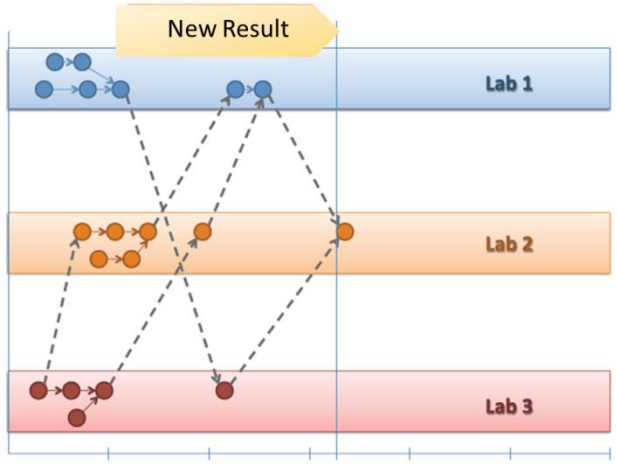
2010 – Research Objects not PDFs

myexperiment



Research has many components, of many types





Exchange of *all* the components of an investigation

Computational instruments break or need to be maintained

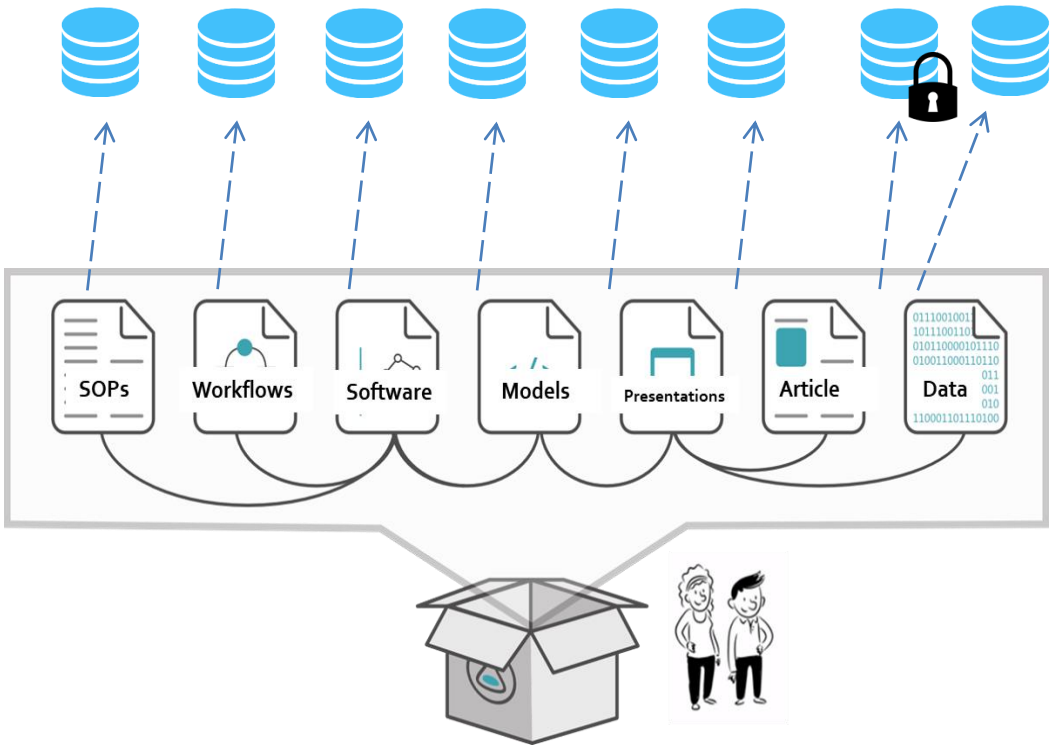
Science, and its products, evolve

Workflow 4Ever

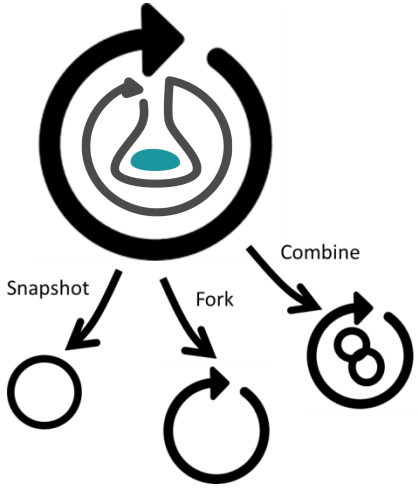
Workflow 4Ever

Overcome fragmentation

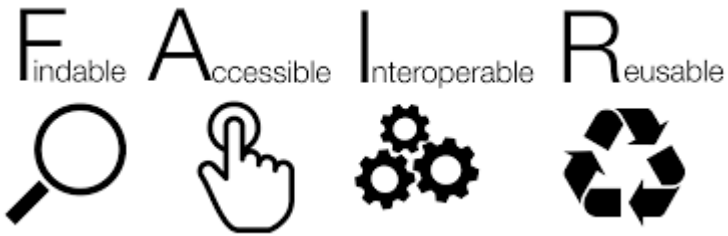
Bundle and relate components



Release evolving research
Accelerate exchange



Support preservation, reproducibility, reuse



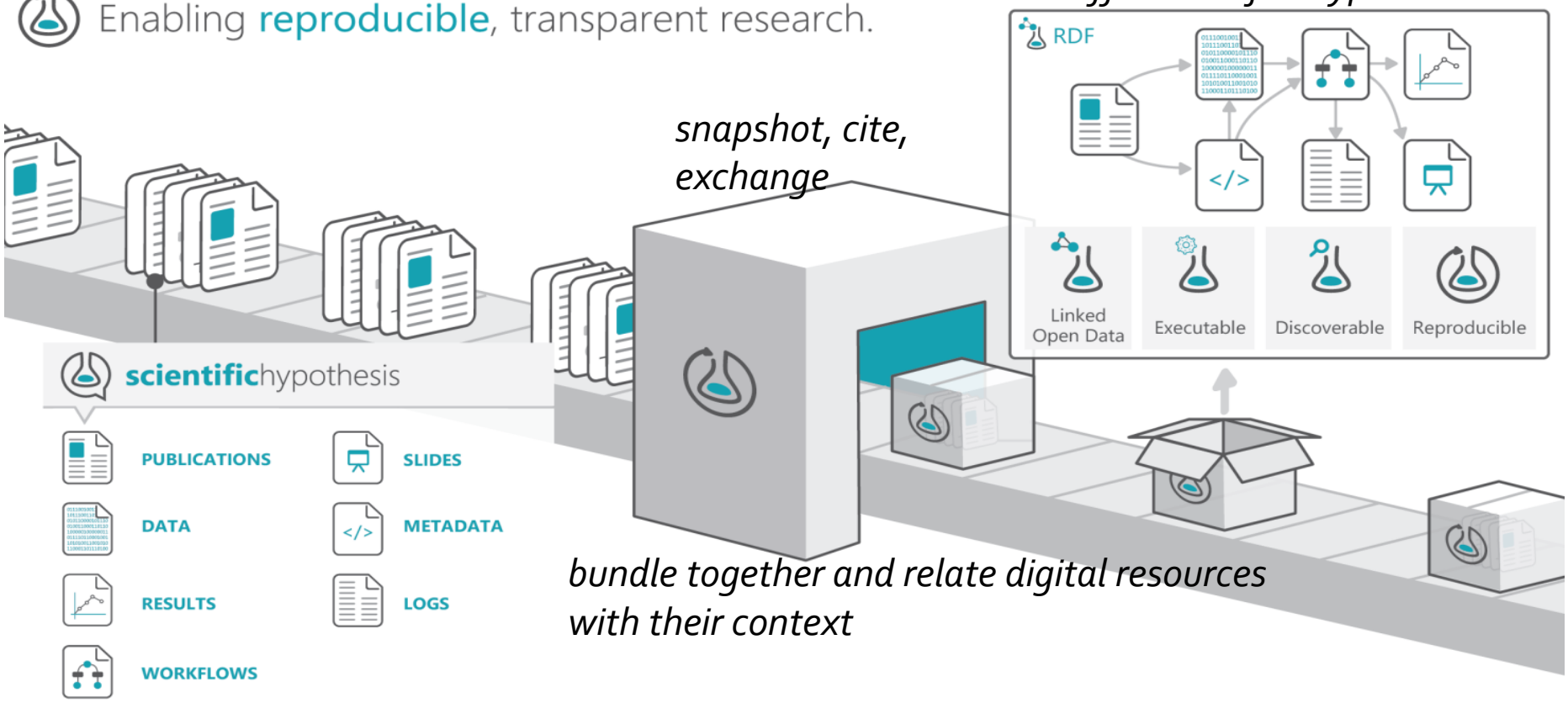
Research Object Framework

Standards-based generic metadata framework

Bechhofer et al (2013) https://doi.org/10.1016/j.future.2011.08.004
Bechhofer et al (2010) https://eprints.soton.ac.uk/268555/

carry machine processable metadata common and specific to different object types

Enabling **reproducible**, transparent research.



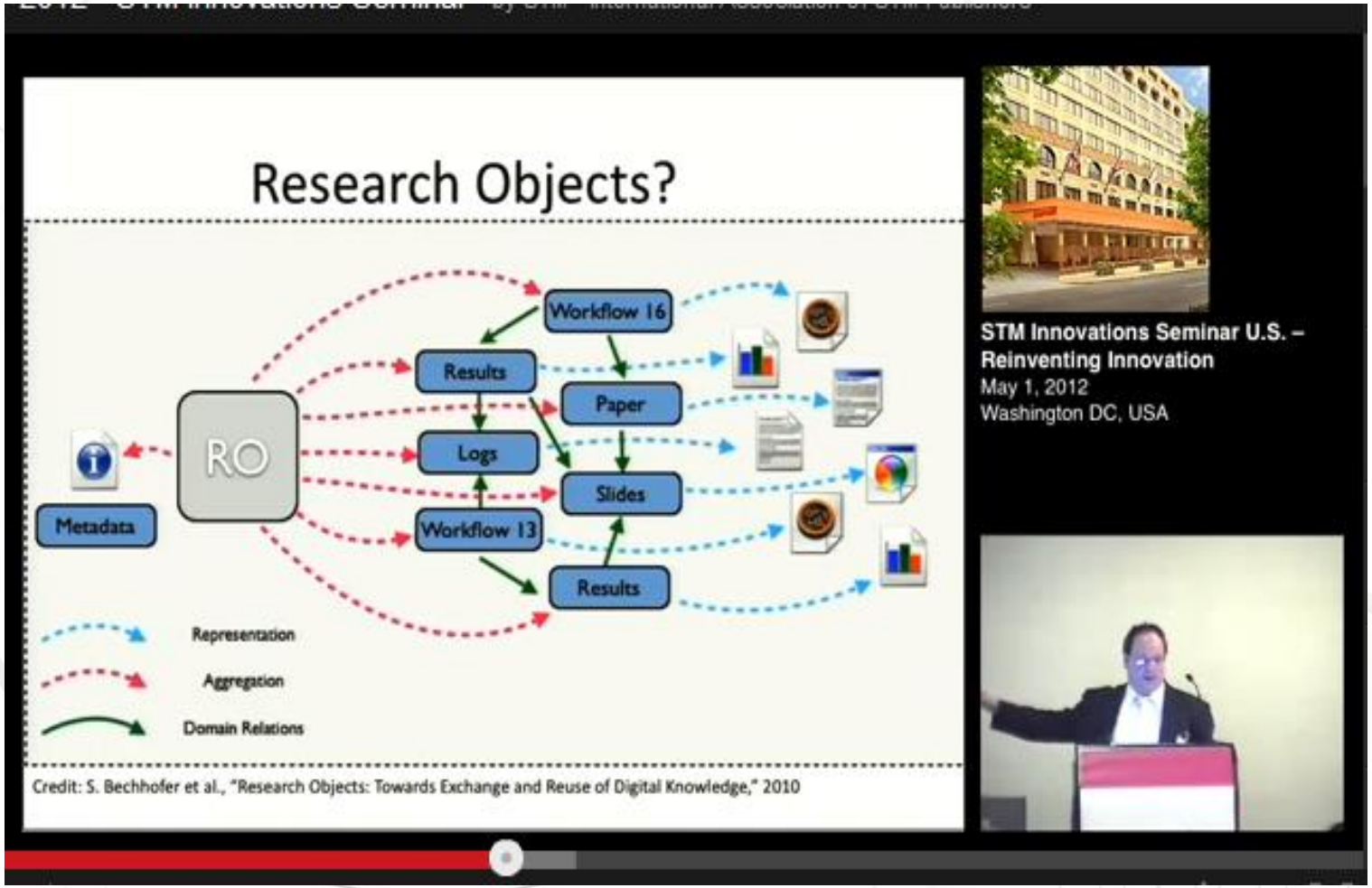
snapshot, cite, exchange

bundle together and relate digital resources with their context

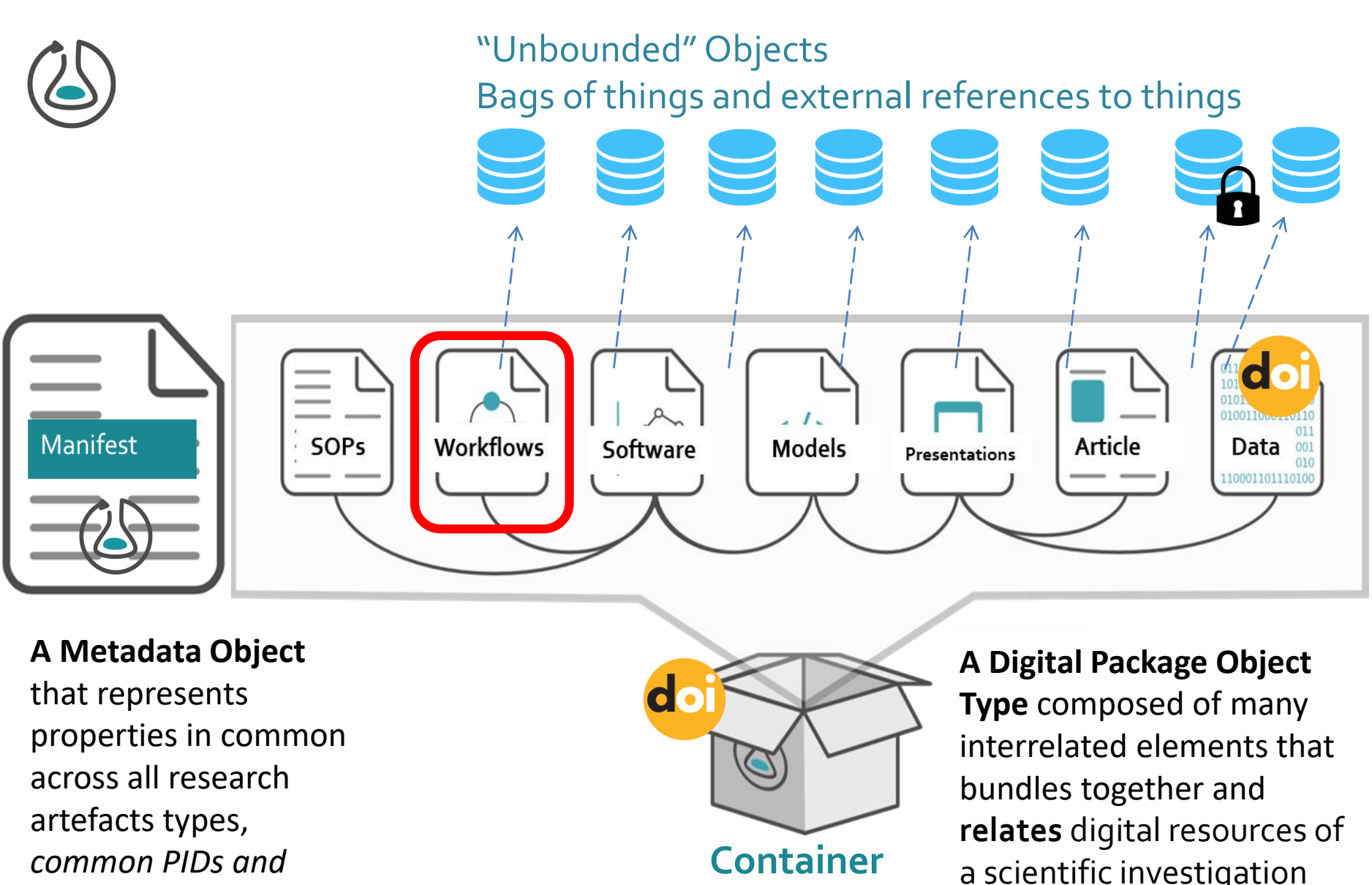
Data used and results produced
Methods used to produce /analyse that data

Provenance and settings, **People** involved,
Annotations understanding & interpretation

Howard Ratner,
Chair STM Future Labs Committee, CEO EVP Nature Publishing Group
Director of Development for CHORUS (Clearinghouse for the Open Research of US) STM Innovations Seminar 2012

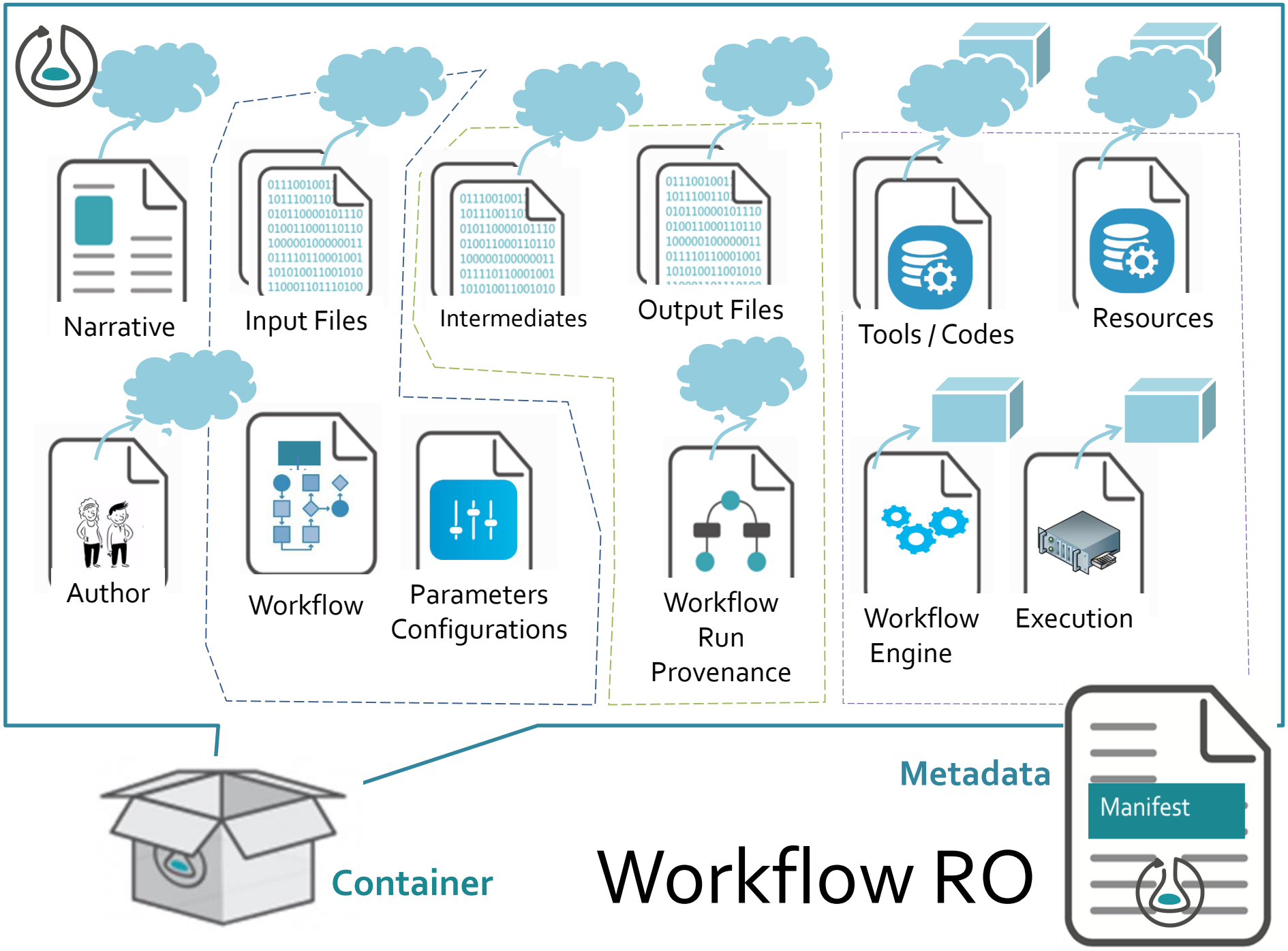


http://www.youtube.com/watch?v=p-W4iLjLTrQ&list=PLC44A300051D052E5



A Metadata Object that represents properties in common across all research artefacts types, *common PIDs and metadata*

A Digital Package Object Type composed of many interrelated elements that bundles together and **relates** digital resources of a scientific investigation with context.


Narrative
Input Files
Intermediates
Output Files
Tools / Codes
Resources
Author
Workflow
Parameters
Configurations
Workflow
Run
Provenance
Workflow
Engine
Execution
Container
Metadata
Manifest
Workflow RO

COMMON WORKFLOW LANGUAGE

Describe and run workflows, and the command line tools they orchestrate, supporting containers to be portable, transparent and interoperable .

W3C PROV

Describe the provenance of the workflow

EDAM

Describe the workflow inputs, outputs, tools and data with controlled vocabularies / ontologies

Galaxy PROJECT

Rabix

Open for Innovation KNIME

Workflow systems run the CWL workflow

docker

BIOCONDA

Software components are containerised to be portable

Workflow RO

Metadata

Gather the CWL workflow descriptions + rich context, provenance using multi-tiered descriptions

Snapshot workflow.

Relate it to other objects.

Container

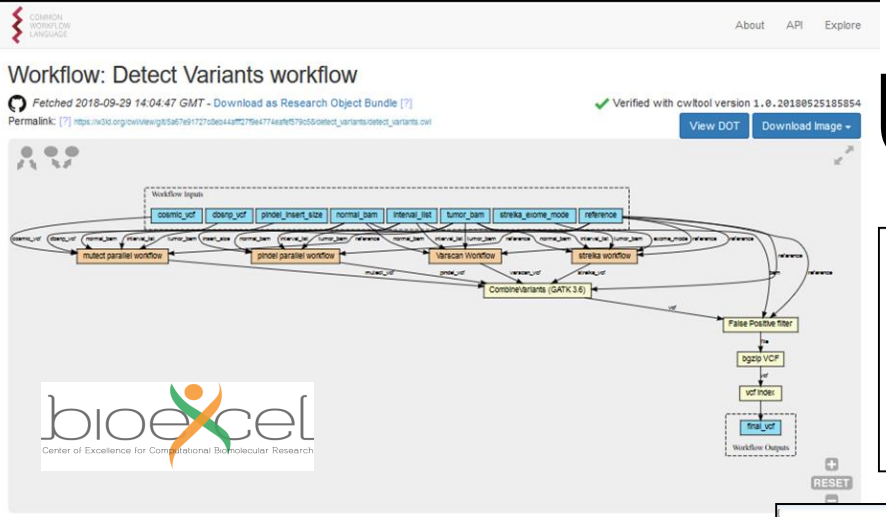
ZIP

BagIt

Archive formats to contain the object

https://www.commonwl.org/

Under the hood



Inputs

ID	Type	Title	Do
reference	File		
pindel_insert_size	Integer		
cosmic_vcf	File (Optional)		
dbsnp_vcf	File (Optional)		
strelka_exome_mode	Boolean		
normal_bam	File		
interval_list	File		
tumor_bam	File		

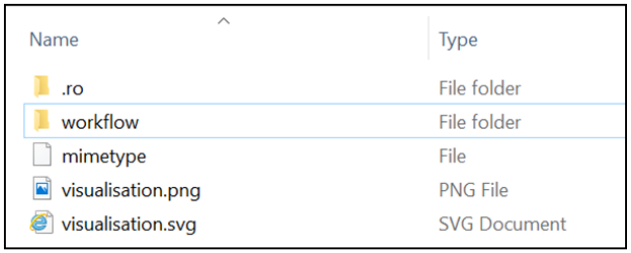
Steps

ID	Runs	Label	Doc
filter	fp_filter.cwl (CommandLineTool)	False Positive filter	
varscan	../varscan/workflow.cwl (Workflow)	Varscan Workflow	
index	index.cwl (CommandLineTool)	vcf index	
bgzip	bgzip.cwl (CommandLineTool)	bgzip VCF	
strelka	../strelka/workflow.cwl (Workflow)	strelka workflow	
mutect	../mutect/workflow.cwl (Workflow)	mutect parallel workflow	
combine	combine_variants.cwl (CommandLineTool)	CombineVariants (GATK 3.6)	
pindel	../pindel/workflow.cwl (Workflow)	pindel parallel workflow	

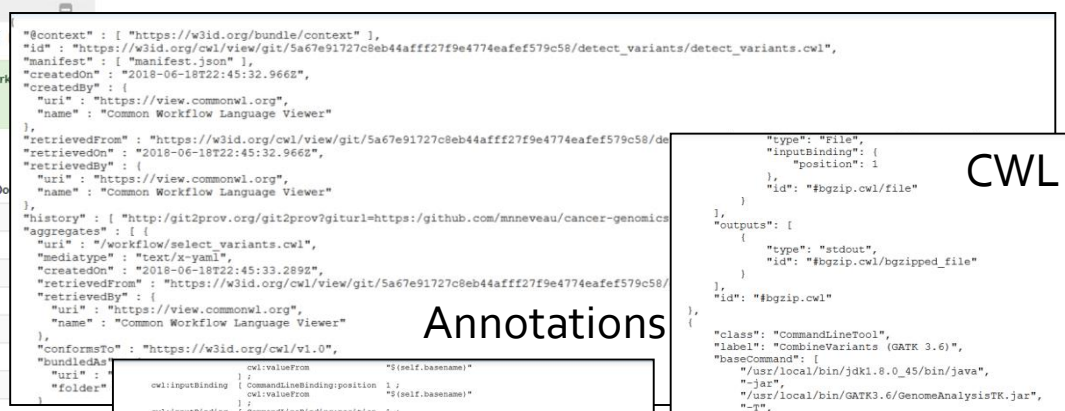
Outputs

ID	Type	Label	Doc
final_vcf	File		

turtle | jsonld | rdfxml | svg | png



Manifest



Annotations

CWL

https://view.commonwl.org/workflows/github.com/mnneveau/cancer-genomics-workflow/blob/master/detect_variants/detect_variants.cwl

BioCompute Objects

U.S. FOOD & DRUG ADMINISTRATION

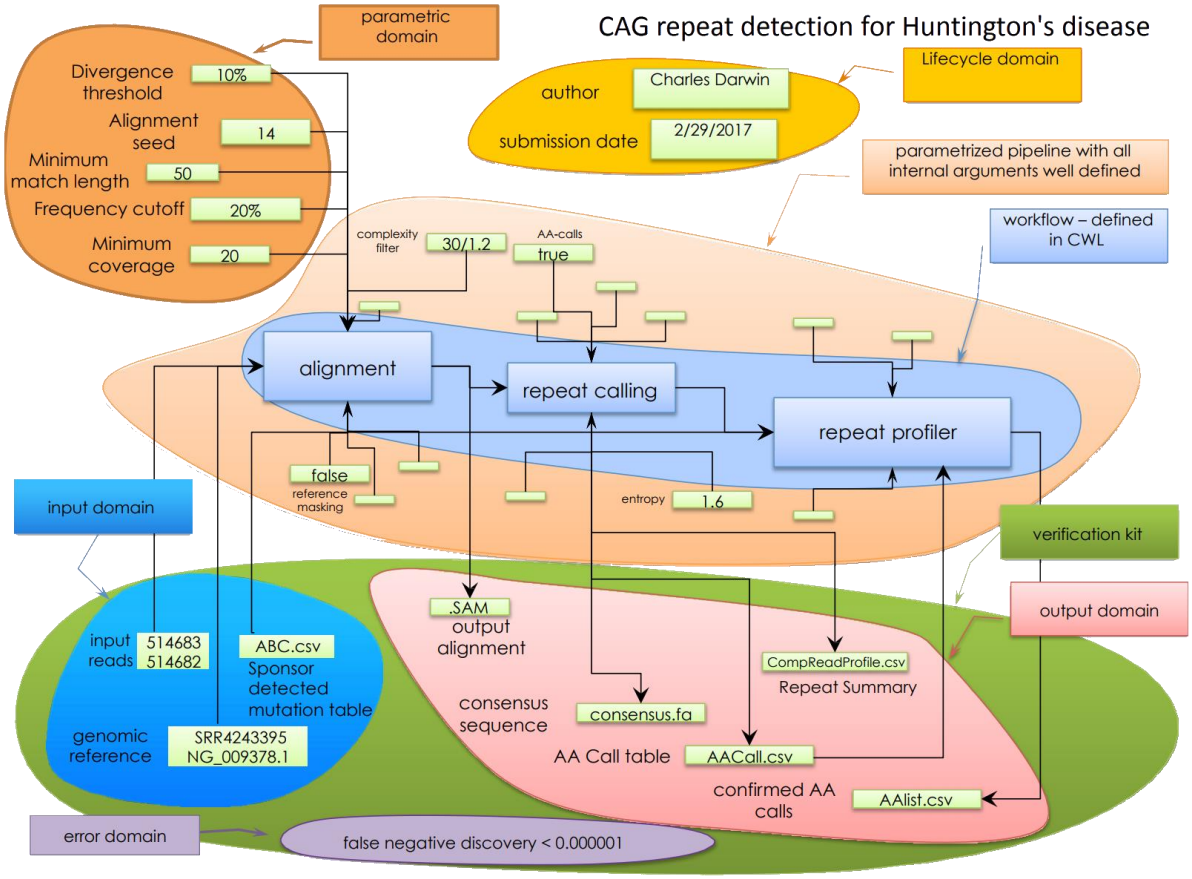
THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

https://osf.io/h59uh/

https://doi.org/10.1101/191783

Standardize exchange of HTS workflows for regulatory submissions between FDA, pharma, bioinformatics platform providers and researchers

Inspect and replicate the computational analytical workflow to review and approve the bioinformatics



researchobject.org

COMMON WORKFLOW LANGUAGE

FHIR®

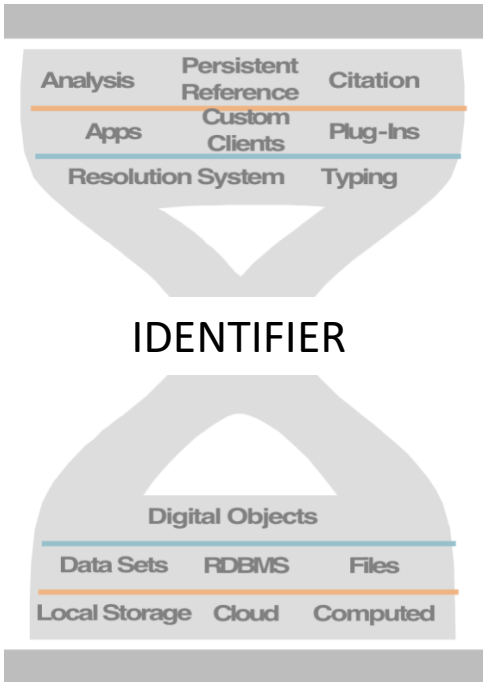
Global Alliance for Genomics & Health

IEEE P2791 BioCompute Working Group (BCOWG)

Desiderata

Use Open Standards.

Incremental.



Graceful degradation.

Machine-processable.

Technology Independent.

Commodity tooling.

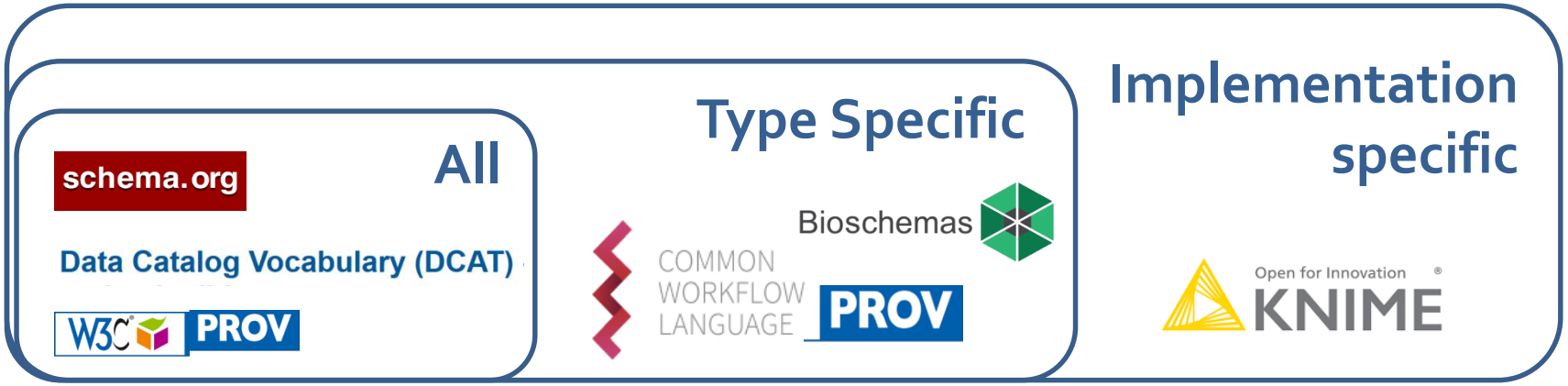
Multi-platform.

The least possible.

The simplest feasible. Low tech.

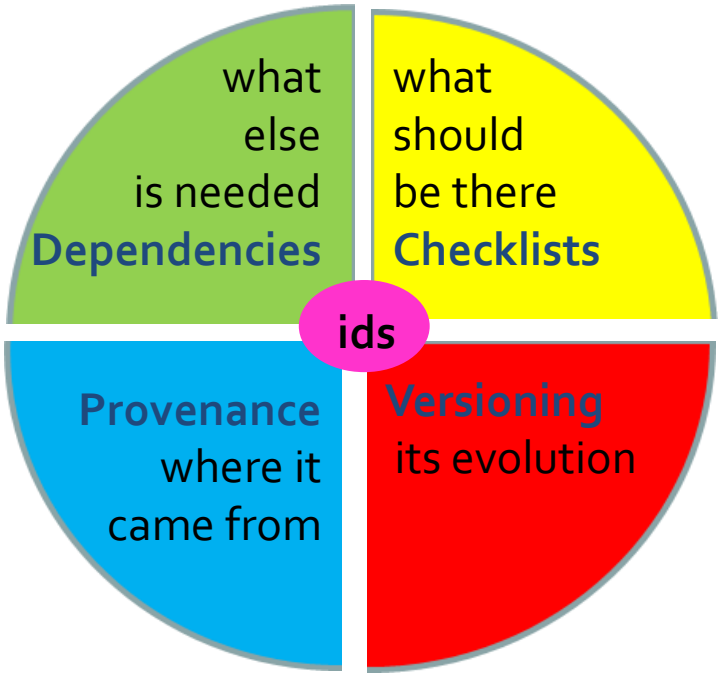
Low user overhead and thin client

Tailored metadata profiles to describe a RO



Manifest Profile Descriptions

Container



Tailored metadata profiles to describe a RO

general purpose to drive scalable infrastructure

Overview | Content | Relations | **Quality** | Notifications | History

Earth Science data-centric RO key requirements ▾ Reload

✗ **Target** Analyzing gene deregulation in Huntington's disease with respect to epigenetic information ***does not satisfy*** **checklist for** ***ready-to-release.***

- ✓ Research Object has title
- ✓ Research Object has description
- ✓ Research Object has creator
- ✗ Research Object does not have editor
- ✗ Research Object does not have indicated purpose
- ✓ Research Object has subject
- ✗ Research Object does not have copyright holder
- ✓ Data artifact is present (e.g., dataset, document, file, image)
- ✗ One or more data artifact has no format indicated, including *http://sandbox.rohub.org/rodl/ROs/HD_chromatin_analysis/broad_hmm_3_Poised_Promoter.txt*
- ✗ One or more data artifact has no filesize indicated, including *http://sandbox.rohub.org/rodl/ROs/HD_chromatin_analysis/broad_hmm_3_Poised_Promoter.txt*
- ✗ Research Object does not have access level specified
- ✗ Research Object does not have DOI

See quality history with RO Monitoring Tool

Validate

ROHUB

everest

Manifest Profile Descriptions

Container

towards generic approaches …

Tailored metadata profiles to describe a RO

general purpose to drive scalable infrastructure

Manifest
Profile
Description

Manifest
Construction

Container
Profile

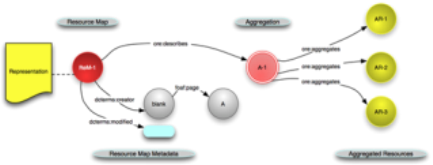
Identity

Identifiers: URI, RRI, DOI, ORCID

Identifiers.org minid

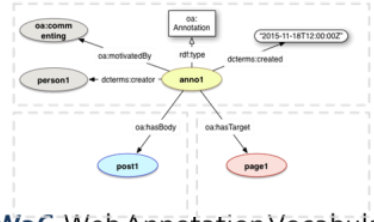
Research Object
Manifest Model

Aggregation



Open Archives Initiative
Object Exchange and Reuse
Aggregation *to enumerate and link together*

Annotation



W3C Web Annotation Vocabulary
Annotation *about RO, elements and their relationships*

Manifest Construction

https://w3id.org/ro/2016-01-28

Wf4Ever Specification

Workflow 4Ever

Research Object ontology

Release 2016-01-28

This version:
https://w3id.org/ro/2016-01-28/
Latest version:
https://w3id.org/ro/
Previous version:
https://w3id.org/ro/2013-11-30/
Revision:
1.0.0-SNAPSHOT
Editors:
Stian Soiland-Reyes, University of Manchester
Sean Bechhofer, University of Manchester
Authors:
Oscar Corcho, UPM
Graham Klyne, University of Oxford
Khalid Belhajjame, University of Manchester
Daniel Garijo, UPM
Esteban García Cuesta, iSOCO
Raul Palma, PSNC
Imported Ontologies:
OAI-ORE
Metadata

DOI 10.5281/zenodo.12744

Includes wfdesc and wfprov
Basis of CWL and CWLProv

Time for a review….?
e.g. RO type to help tools..

Container Profiles

Research Object Bundle 1.0

Specification for a structured ZIP-file, based on the ePub and Adobe UCF specifications

https://researchobject.github.io/specifications/bundle/

BagIt

Specifies a file system structure for transferring and archiving a collection of files, **including their checksums to verify and validate content** and brief metadata.

https://github.com/ResearchObject/bagit-ro

mechanism for serialization and transport consistency,

+

capture identity, annotations and provenance of the resources

→

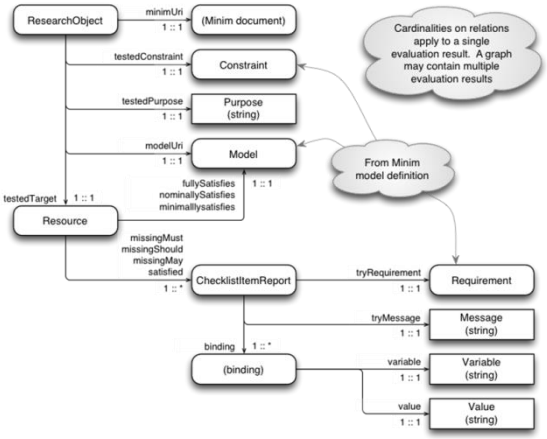
BIG DATA BAG

Big Data collections of arbitrary referenced content

https://github.com/fair-research/bdbag

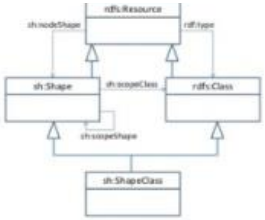
Manifest Profile Description

general construction & validation tooling



Minim model for defining checklists

Shapes Constraint Language



Linked Data and RDF Shapes
Validate graph-based data against a set of conditions

Gamble, Zhao, Klyne, Goble. IEEE eScience 2012, http://dx.doi.org/10.1109/eScience.2012.6404489

ro-show

bundle.zip

validation

FAILED MANIFEST VALIDATION

validated against the following shapes:

workflow-model:

provenance-shape.ttl

aggregate-shape.ttl

- Value for manifest.json property with path 'mediatype' and value 'img/svg+xml' is invalid. (Should be either img/png or img/jpg)

annotation-shape.ttl

basic-model:

preview

info provenance aggregates annotations

context

https://w3id.org/bundle/context

creator

bundle.zip was created by Common Workflow Language Viewer

- RO pre-processing to merge to single graph
- RDF Shape that indicates to follow links
- Bespoke validators / unpackers to iterate over the RO

profile

Workflow-centric Research Object

Validated against workflow shape-graph model

[other profile specific info to be included in future releases!!!]

structure

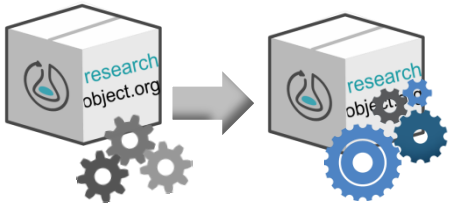
```
+--cac747e8-68f5-40d0-b53d-78e68483ed1c/
|  +--visualisation.svg
|  +--.ro/
|  |  +--annotations/
|  |  |  +--merged.cwl
|  |  |  +--workflow.ttl
|  |  +--manifest.json
|  +--visualisation.png
|  +--mimetype
|  +--workflow/
|  |  +--bam_to_trimmed_fastq_and_hisat_alignments.cwl
|  |  +--merge.cwl
```

[Lilian Gorea, Oluwatomide Fasugba 2018]

Research Object drivers

Exchange & Commons

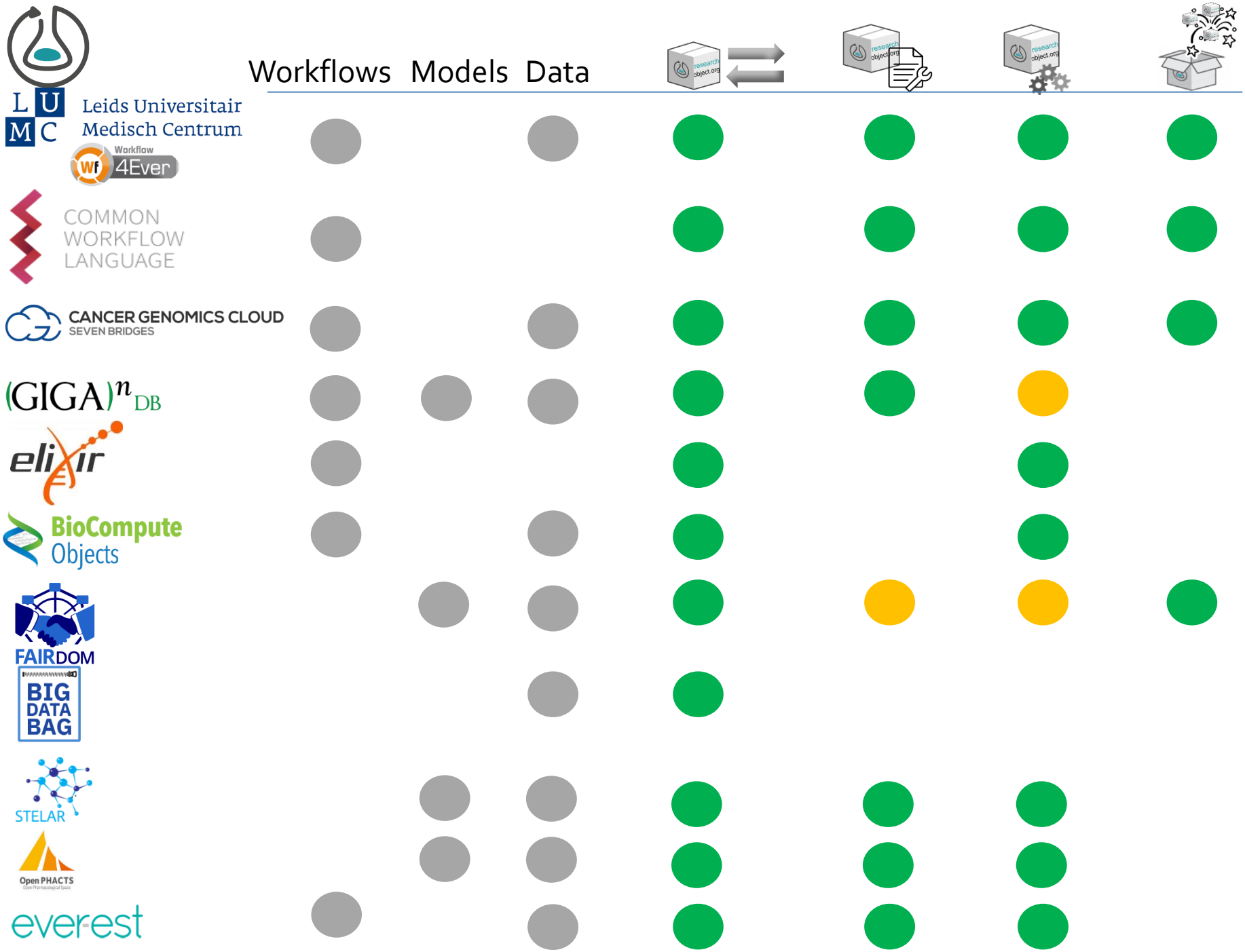
Preservation and fixed point publishing



Reproducibility and execution

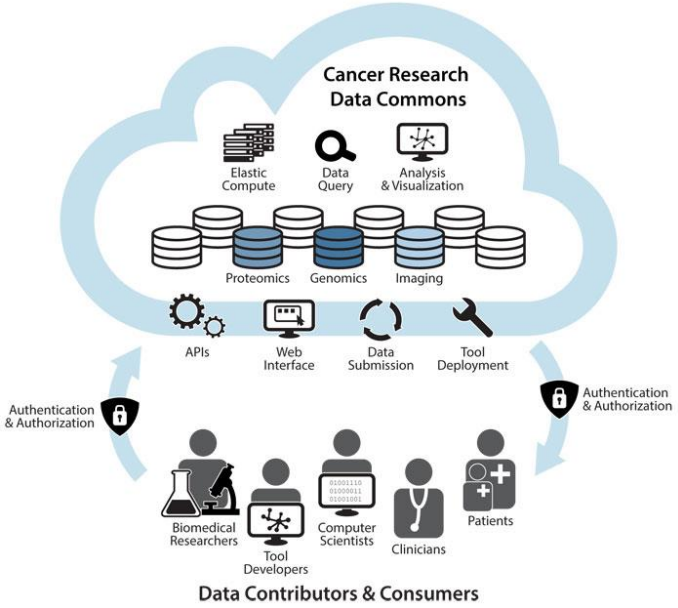
Active "release" research

Goble, De Roure, Bechhofer, Accelerating Knowledge Turns, DOI: 10.1007/978-3-642-37186-8_1


Workflows
Models
Data
Leids Universitair Medisch Centrum
Workflow 4Ever
COMMON WORKFLOW LANGUAGE
CANCER GENOMICS CLOUD SEVEN BRIDGES
(GIGA)n DB
elixir
BioCompute Objects
FAIRDOM
BIG DATA BAG
STELAR
Open PHACTS
everest

NIH Data Commons

CANCER GENOMICS CLOUD
SEVEN BRIDGES



ELSEVIER

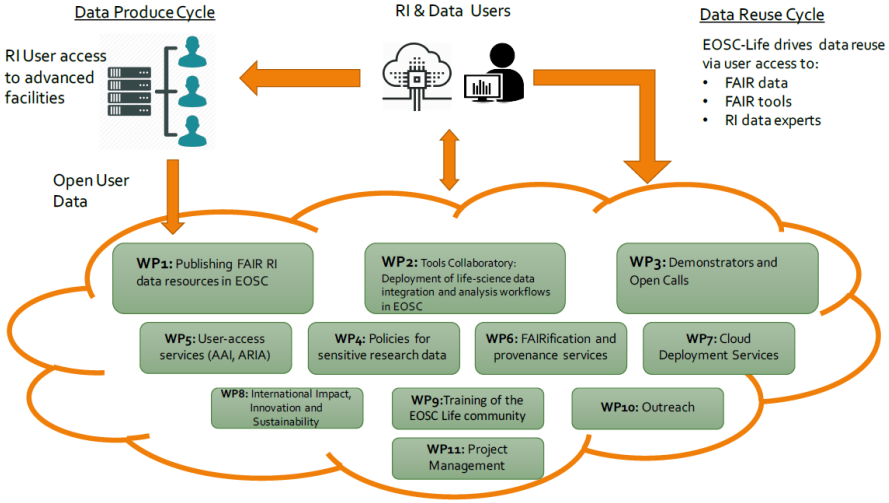
FAIR4CURES

MENDELEY DATA
Put your research data online today

European Open Science Cloud

bioexcel
Center of Excellence for Computational Biomolecular Research

elixir



CWL Workflow Collaboratory

IBISBA 1.0
Industrial Biotechnology Innovation and Synthetic Biology Accelerator

FAIRDOM

EOSC pilot
The European Open Science Cloud for Research Pilot Project

Research Object Gaps

research object.org

RO Profiles
- Templates & Standard types

General tooling
- Construction, Validation, Viewing

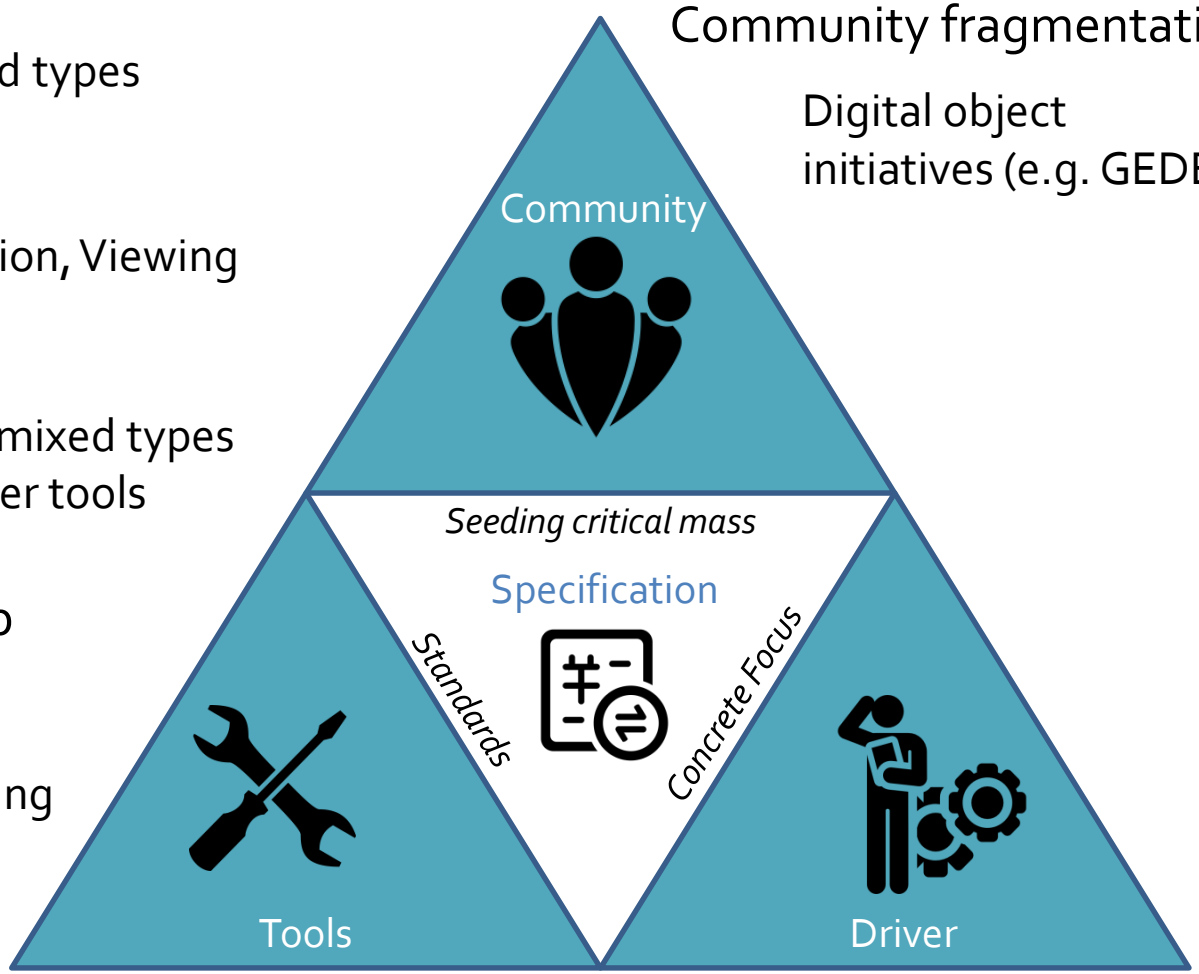
Handling RO composition
- Nesting, Complex & mixed types
- Less flexibility -> easier tools

RO Life cycles stewardship
- Fixed snapshot
- Living objects
- Rot, mutations, cloning
- References

RO Circulation
- Credit, tracking

Community fragmentation

Digital object initiatives (e.g. GEDE)



researchobject.org

Build a Community

Acknowledgements

Barend Mons
Sean Bechhofer
Matthew Gamble
Raul Palma
Jun Zhao
Mark Robinson
Alan Williams
Norman Morrison
Stian Soiland-Reyes
Tim Clark
Alejandra Gonzalez-Beltran
Philippe Rocca-Serra
Ian Cottam
Susanna Sansone
Kristian Garza
Daniel Garijo
Catarina Martins
Iain Buchan
Michael Crusoe
Rob Finn

Carl Kesselman
Ian Foster
Kyle Chard
Vahan Simonyan
Ravi Madduri
Raja Mazumder
Gil Alterovitz,
Denis Dean II
Durga Addepalli
Wouter Haak
Anita De Waard
Paul Groth
Oscar Corcho
Josh Sommer

The University of Manchester

FAIRDOM

FAIR4CURES

BioCompute Objects

IBISBA 1.0

bioexcel
Center of Excellence for Computational Biomolecular Research

Project ID: 675728

COMMON WORKFLOW LANGUAGE

Workflow 4Ever

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

elixir

BBSRC
bioscience for the future

Federal Ministry of Education and Research

NIH Big Data to Knowledge (BD2K)

MRC Medical Research Council