researchobject.org

Research Objects

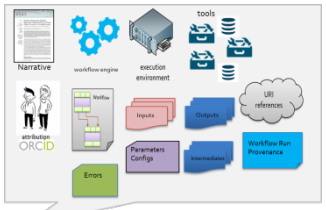
More than the Sum of the Many Parts

Carole Goble
The University of Manchester, UK
EU Infrastructures ELIXIR-UK, FAIRDOM, BioExcel, ISBE …
Software Sustainability Institute UK

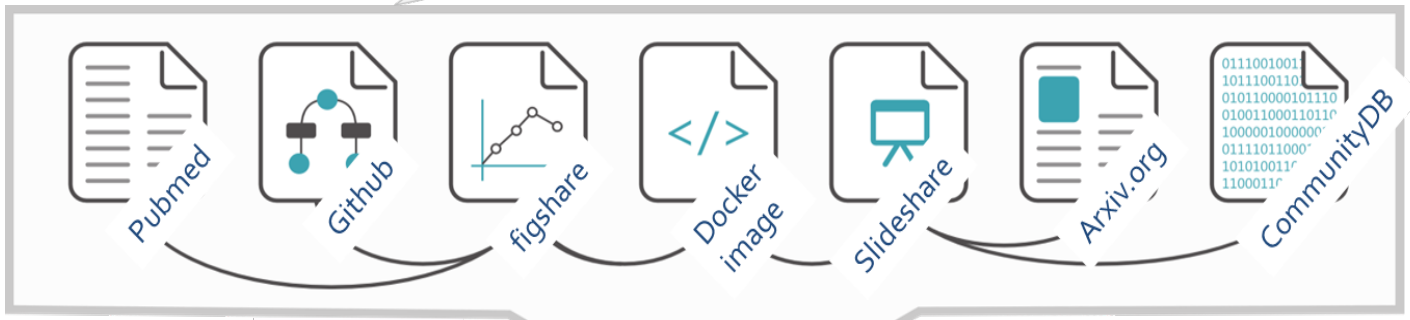
Workshop on Managing Digital Research Objects in an Expanding Science Ecosystem, 15 Nov 2017, Bethesda, USA

MICKE

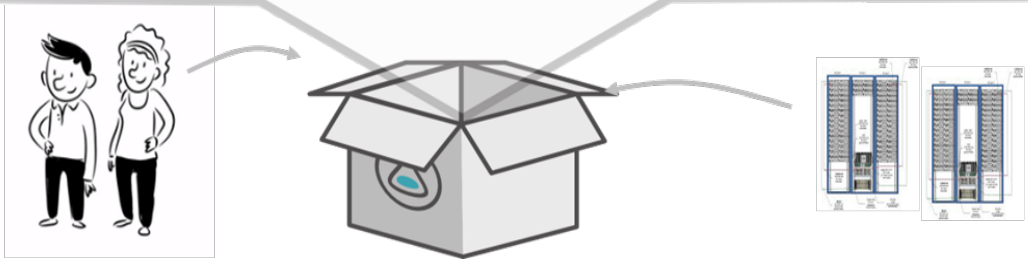
Digital Objects Wholes & Parts in an Expanding Ecosystem



A Digital Package Object Type composed of many interrelated elements



A Digital Object that represents properties in common across all research artefact types, *Common PIDs and Metadata*



A Digital Package Object that bundles together and **relates** digital resources of a scientific investigation with context

Citable Reproducible Packaging

Nested content
Heterogeneous elements.
Distributed and embedded content.
Externally stewarded content.
Checklists + Checksums

Workflow driven Data Analytics:

Research Components are Many and Various

Journal of Biogeography

Original Article

Mapping present and future potential distribution patterns for a meso-grazer guild in the Baltic Sea

Sonja Leidenberger[1,*], Renato De Giovanni[2], Robert Kulawik[3], Alan R. Williams[4] and Sarah J. Bourlat[5]

Article first published online: 5 SEP 2014

DOI: 10.1111/jbi.12395

© 2014 The Authors. *Journal of Biogeography* Published by John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Issue

Journal of Biogeography
Volume 42, Issue 2, pages 241–254, February 2015

SEARCH

In this issue

Advanced > Saved Searches >

ARTICLE TOOLS

Get PDF (1271K)
Save to My Profile
E-mail Link to this Article
Export Citation for this Article
Get Citation Alerts
Request Permissions

Share

Additional Information (Show All)

How to Cite | Author Information | Publication History | Funding Information

Abstract | Article | References | Supporting Information | Cited By

Enhanced Article (HTML) | Get PDF (1271K) | Manchester University

Keywords:
Climate change; Baltic Sea; ecological niche modelling; e-Science; food web; *Fucus radicans*; *Fucus vesiculosus*; *Gasterosteus aculeatus*; *Idotea*; workflows

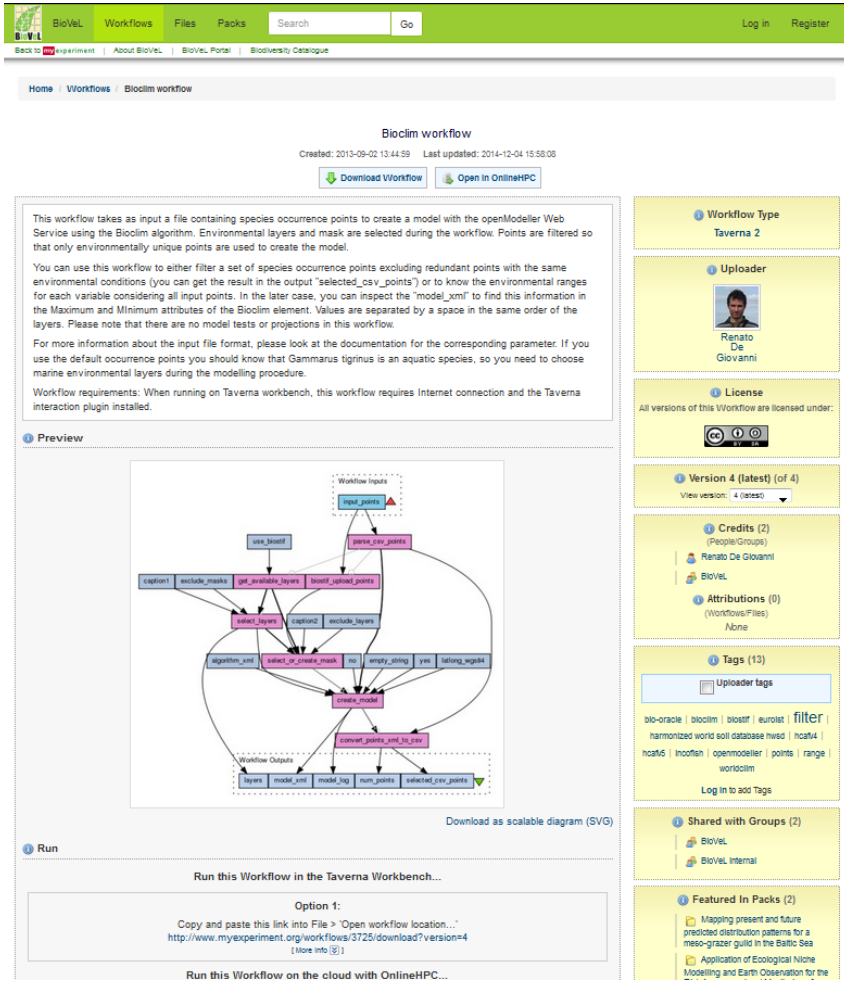
Abstract

Jump to…

Aim

The Baltic Sea is one of the world's largest semi-enclosed brackish water bodies characterized by many special features, including endemic species that may be particularly threatened by climate change. We mapped potential distribution patterns under present and future conditions for a community with three trophic levels. We analysed climate-induced changes in the species' distribution patterns and examined possible consequences for the chosen food web.

…orthern Europe.



BioVeL | Workflows | Files | Packs | Search | Go | Log in | Register

Home / Workflows / Bioclim workflow

Bioclim workflow

Created: 2013-09-02 13:44:59 Last updated: 2014-12-04 15:58:08

Download Workflow | Open in OnlineHPC

This workflow takes as input a file containing species occurrence points to create a model with the openModeller Web Service using the Bioclim algorithm. Environmental layers and mask are selected during the workflow. Points are filtered so that only environmentally unique points are used to create the model.

You can use this workflow to either filter a set of species occurrence points excluding redundant points with the same environmental conditions (you can get the result in the output "selected_csv_points") or to know the environmental ranges for each variable considering all input points. In the later case, you can inspect the "model_xml" to find this information in the Maximum and Minimum attributes of the Bioclim element. Values are separated by a space in the same order of the layers. Please note that there are no model tests or projections in this workflow.

For more information about the input file format, please look at the documentation for the corresponding parameter. If you use the default occurrence points you should know that Gammarus tigrinus is an aquatic species, so you need to choose marine environmental layers during the modelling procedure.

Workflow requirements: When running on Taverna workbench, this workflow requires Internet connection and the Taverna interaction plugin installed.

Preview

Download as scalable diagram (SVG)

Run

Run this Workflow in the Taverna Workbench...

Option 1:

Copy and paste this link into File > 'Open workflow location...'
http://www.myexperiment.org/workflows/3725/download?version=4

Run this Workflow on the cloud with OnlineHPC...

Workflow Type
Taverna 2

Uploader
Renato De Giovanni

License
All versions of this Workflow are licensed under:

Version 4 (latest) (of 4)

Credits (2)
(People/Groups)
Renato De Giovanni
BioVeL

Attributions (0)
(Workflows/Files)
None

Tags (13)
Uploader tags
bio-oracle | bioclim | biostif | eurolst | filter | harmonized world soil database hwsd | hcaf\4 | hcaf\5 | incofish | openmodeller | points | range | worldclim
Log in to add Tags

Shared with Groups (2)
BioVeL
BioVeL Internal

Featured In Packs (2)
Mapping present and future predicted distribution patterns for a meso-grazer guild in the Baltic Sea
Application of Ecological Niche Modelling and Earth Observation for the

BioVeL

myexperiment

Added by **Biodiversity eLaboratory** ... about one hour ago (04/08/14 @ 11:36:52)

more

File: Fucus vesiculosus current distribution (Biodiversity eLaboratory)

A low-resolution PNG image of the predicted current niche of Fucus vesilosus

Added by **Biodiversity eLaboratory** ... about one hour ago (04/08/14 @ 11:36:26)

more

File: Fucus vesiculosus current and 2050 difference (Biodiversity eLaboratory)

A low resolution PNG image for the difference between the current and predicted 2050 niches for Fucus vesiculosus

Added by **Biodiversity eLaboratory** ... about one hour ago (04/08/14 @ 11:35:41)

more

File: Fucus vesiculosus occurrences from GBIF (Biodiversity eLaboratory)

A spreadsheet indicating the source of occurrences of Fucus vesiculosus retrieved from GBIF.

Added by **Biodiversity eLaboratory** ... about one hour ago (04/08/14 @ 11:25:26)

more

Workflow: Bioclim workflow with interaction [2] (Renato De Giovanni)

Taverna 2 BioClim workflow developed as part of the BioVeL project. The workflow filters for environmentally unique points and generates a BioClim model to calculate the environmental range for each variable. The workflow uses the OpenModeller framework

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:05:40)

more

Workflow: BioVeL ESW DIFF Basic [2] (Robert Kulawik)

Taverna 2 ENM Statistical Workflow developed as part of the BioVeL pr... The workflow allows the spatial computation of changes in potential distribution maps by ...ring the differences between two raster layers using the R statistical environment

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 09:52:35)

more

Workflow: Ecological niche modelling workflow [20] (Renato De Giovanni)

Taverna 2 Ecological Niche Modelling Workflow developed as part of the BioVeL project. The

Local files

Workflows

Workflow Commons

10th Anniversary!

Added by **Biodiversity eLaboratory** ... 04/08/14 @ 10:23:33)

External: Idotea chelipes - (http://... ...p=taxdetails&id=119042)

A species of isopod that is prey fo... ...n the two species of Fucus algae.

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:22:40)

External: Idotea balthica - (http://www.marinespecies.org/aphia.php?p=taxdetails&id=11...

A species of isopod that is prey for Gasterosteus aculeatus and which feeds on the two species of Fucus algae.

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:21:43)

External: Fucus radicans - (http://www.marinespecies.org/aphia...

A species of brown algae that is eaten by Idotea

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:19...

External: Fucus vesiculosus - (http://www.marinespecies.org/ap...

A seaweed/algae that is eaten by Idotea

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:17:53)

External: OpenModeller ecological niche modeling experimentation framework - (http://openmodeller.sourceforge.net/)

Added by **Biodiversity eLaboratory** ... 2 hours ago (04/08/14 @ 10:00:38)

Pointer to 3rd Party Data Collection

Pointer to 3rd Party Code

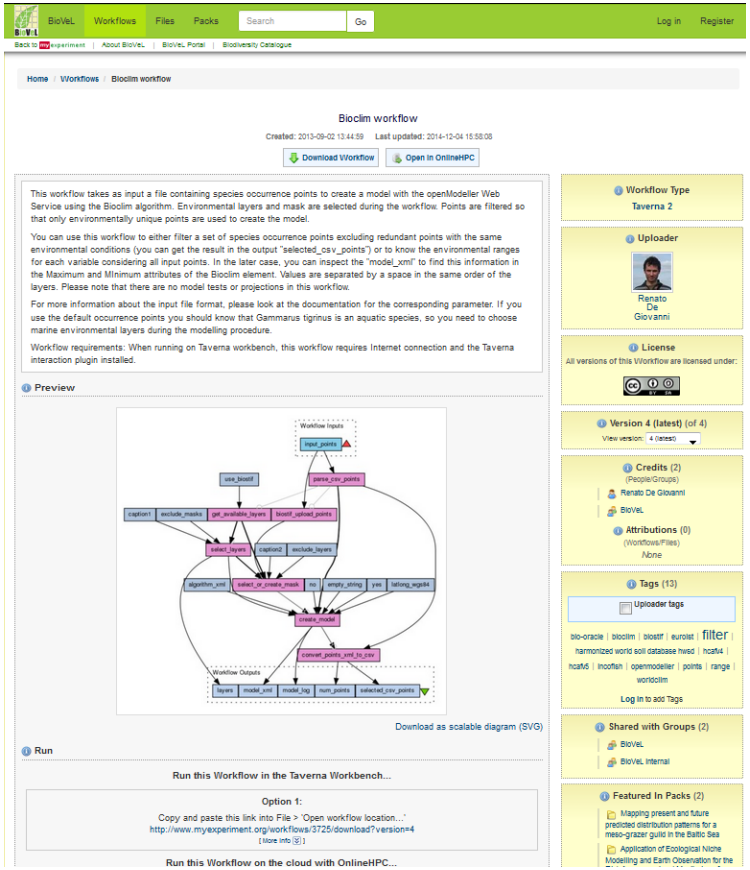
Relationships (5)

1. Fucus vesiculosus 2050 distribution is an output from Ecological ...
2. Fucus vesiculosus 2050 distribution is an input to BioVeL ESW D...
3. Fucus vesiculosus current distribution is an input to BioVeL ESW DIFF Basic.
4. Fucus vesiculosus current distribution is an output from Ecological niche modelling workflow.
5. Fucus vesiculosus current and 2050 difference is an output from BioVeL ESW DIFF Basic.

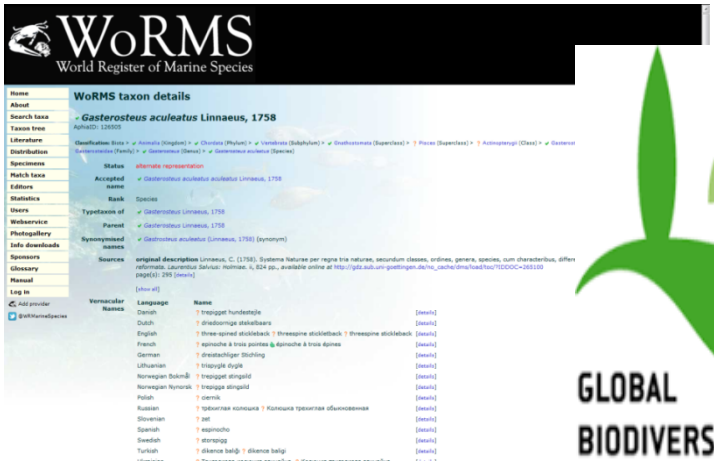
Download

Download Pack Items (ZIP archive)

Track?



Workflows of local or remotely executed codes



Collection in a Data Catalogue

GLOBAL BIODIVERSITY INFORMATION FACILITY

Third party remote web services or command line tools

open modeller

Systems and Synthetic Biology:

Research Components are Many and Various

the FEBS Journal

Construction and validation of a detailed kinetic model of glycolysis in *Plasmodium falciparum*

Gerald Penkler[1,2], Francois du Toit[1], Waldo Adams[1], Marina Rautenbach[1], Daniel C. Palm[1], David D. van Niekerk[1] and Jacky L. Snoep[1,2,3]

1 Department of Biochemistry, Stellenbosch University, Matieland, South Africa
2 Molecular Cell Physiology, Vrije Universiteit Amsterdam, The Netherlands
3 MIB, University of Manchester, UK

Keywords
enzyme kinetics; glucose metabolism; model workflow; mathematical model; systems biology

Correspondence
J. L. Snoep, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
Fax: +27 (0)2180 85863
Tel: +27 (0)2180 85844
E-mail: jls@sun.ac.za

(Received 19 August 2014, revised 7 February 2015, accepted 13 February 2015)

doi:10.1111/febs.13237

The enzymes in the Embden–Meyerhof–Parnas pathway of *Plasmodium falciparum* trophozoites were kinetically characterized and their integrated activities analyzed in a mathematical model. For validation of the model, we compared model predictions for steady-state fluxes and metabolite concentrations of the hexose phosphates with experimental values for intact parasites. The model, which is completely based on kinetic parameters that were measured for the individual enzymes, gives an accurate prediction of the steady-state fluxes and intermediate concentrations. This is the first detailed kinetic model for glucose metabolism in *P. falciparum*, one of the most prolific malaria-causing protozoa, and the high predictive power of the model makes it a strong tool for future drug target identification studies. The modelling workflow is transparent and reproducible, and completely documented in the SEEK platform, where all experimental data and model files are available for download.

Database
The mathematical models described in the present study have been submitted to the JWS Online Cellular Systems Modelling Database (http://jjs-bio.vu.nl/database/penkler). The investigation and complete experimental data set is available on SEEK (10.15490/seek.1.investigation.56).

Introduction

Despite several attempts at a complete eradication of the disease, malaria is still killing more than half a million people per year, mostly small children in sub-saharan Africa (World Health Organisation Malaria report 2013, http://www.who.int/malaria/publications/world_malaria_report_2013/en/). The disease is caused by parasitic protozoa of the *Plasmodium* genus, which have a complicated life cycle consisting of an insect vector and vertebrate host [1]. In the human host, parasite sporozoites first invade liver cells, but the malaria disease symptoms manifest only at a later stage during multiplication of the asexual stages of the parasite in red blood cells (RBCs). The blood life cycle consists of ring, trophozoite and schizont stages, and subsequent

Abbreviations
2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; ALD, fructose-bisphosphate aldolase; B13PG, 1,3-bisphosphoglycerate; DHAP, glycerone phosphate; ENO, phosphopyruvate hydratase; F16BP, fructose 1,6-bisphosphate; F6P, fructose 6-phosphate; G3P, glycerol 3-phosphate; G3PDH, glycerol 3-phosphate dehydrogenase; G6P, glucose 6-phosphate; GAP, D-glyceraldehyde 3-phosphate; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; GLC, glucose; GLY, glycerol; HK, hexokinase; LAC, lactate; LDH, lactate dehydrogenase; MCT, monocarboxylate transporter; ODE, ordinary differential equation; PEP, phosphoenolpyruvate; PFK, 6-phosphofructokinase; PGI, glucose 6-phosphate isomerase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; PK, pyruvate kinase; PYR, pyruvate; RBC, red blood cell; TCA, tricarboxylic acid; TPI, triose-phosphate isomerase.

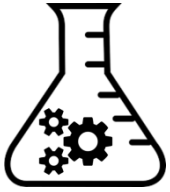
Penkler, G., du Toit, F., Adams, W., Rautenbach, M., Palm, D. C., van Niekerk, D. D. and Snoep, J. L. (2015), Construction and validation of a detailed kinetic model of glycolysis in *Plasmodium falciparum*. FEBS J, 282: 1481–1511. doi:10.1111/febs.13237

16 datafiles (kinetic, flux inhibition, runout)

19 models (kinetics, validation)

13 Standard Operating Procedures

3 studies (model analysis, construction, validation)



24 assays/analyses (simulations, model characterisations)

Systems & Synthetic Biology Commons for Projects

Browse

FAIRDOM

Yellow pages
- Programmes
- Projects
- Institutions
- People

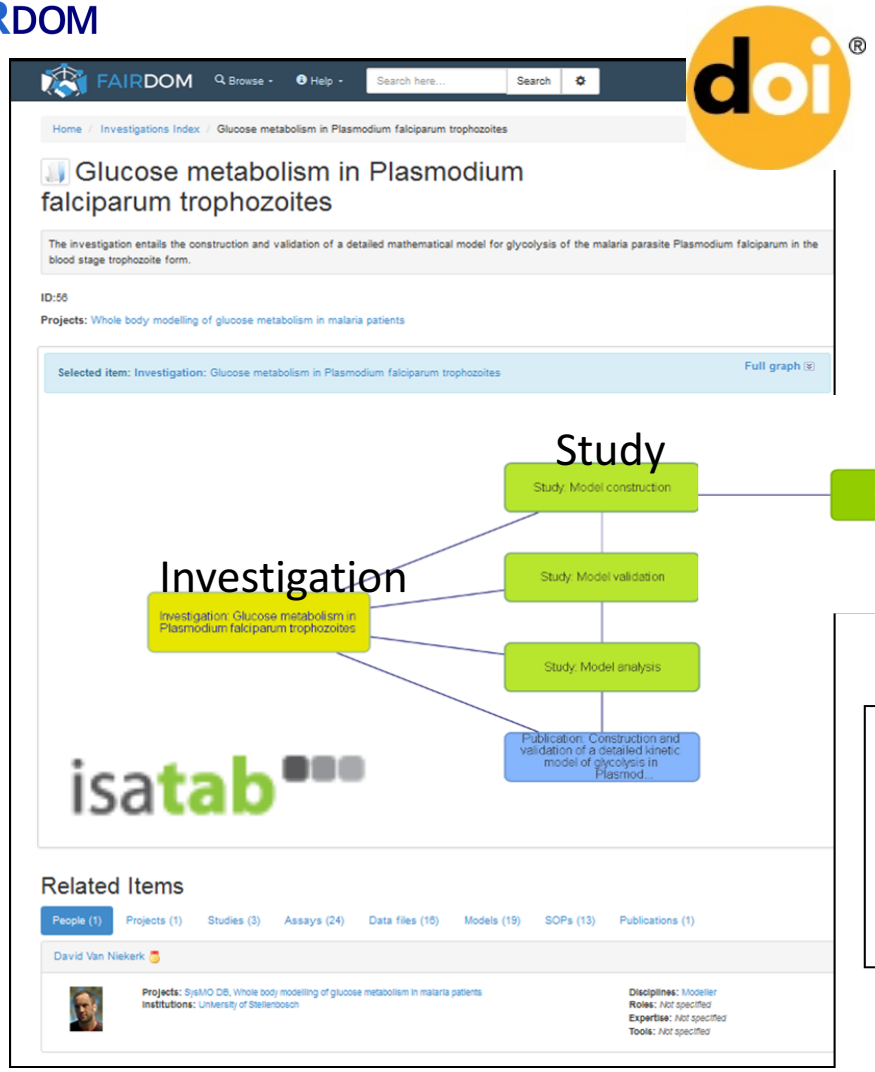
Experiments
- Investigations
- Studies
- Assays

Assets
- Data files
- Models
- SOPs
- Publications

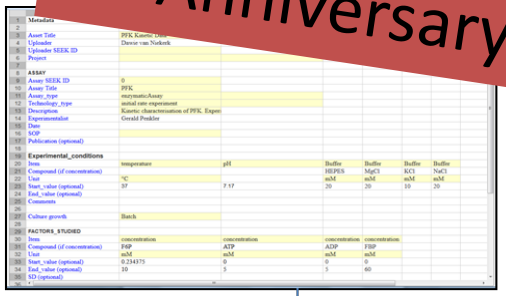
Activities
- Presentations
- Events

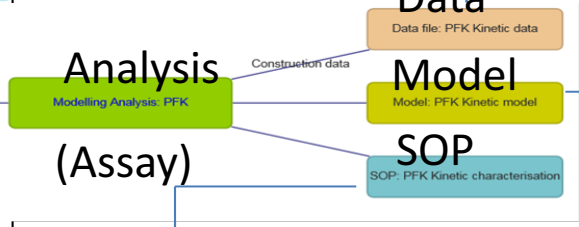
Samples
- Sample types
- Organisms

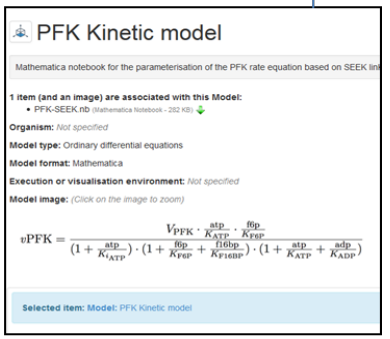
9th Anniversary!



doi®

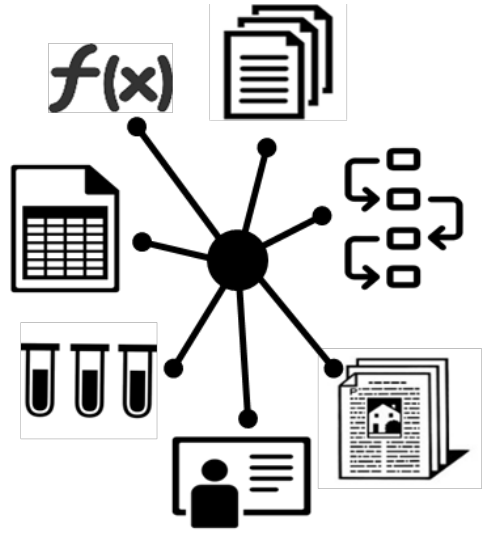




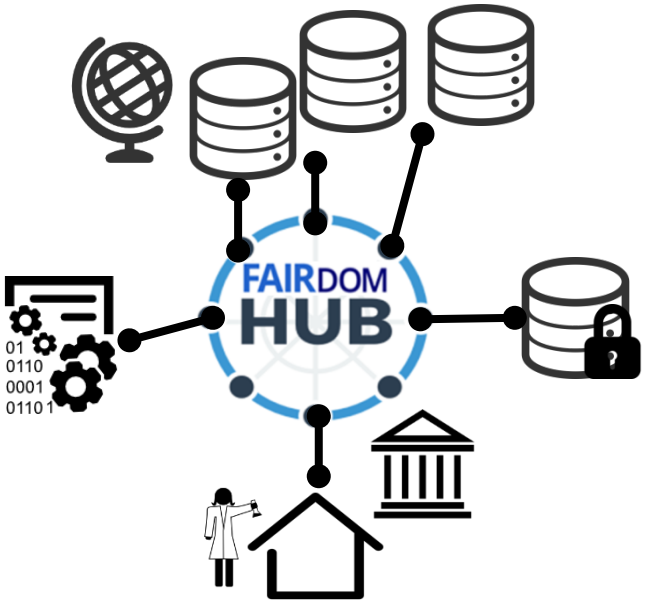


https://fairdomhub.org/investigations/56

Citation G. Penkler; F. du Toit; W. Adams; M. Rautenbach; D. C. Palm; D. D. van Niekerk; J. L. Snoep; (2014): Glucose metabolism in Plasmodium falciparum trophozoites; FAIRDOMHub. http://dx.doi.org/10.15490/seek.1.investigation.56



Multi-results & Versions
Data of many types…Primary, secondary, tertiary…
Methods, models, scripts …
Physical objects: samples, strains, specimens

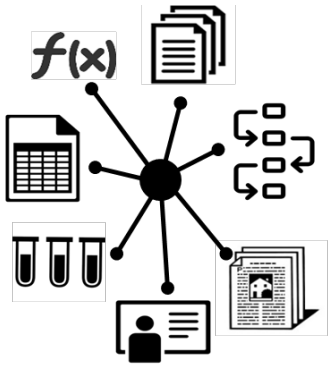


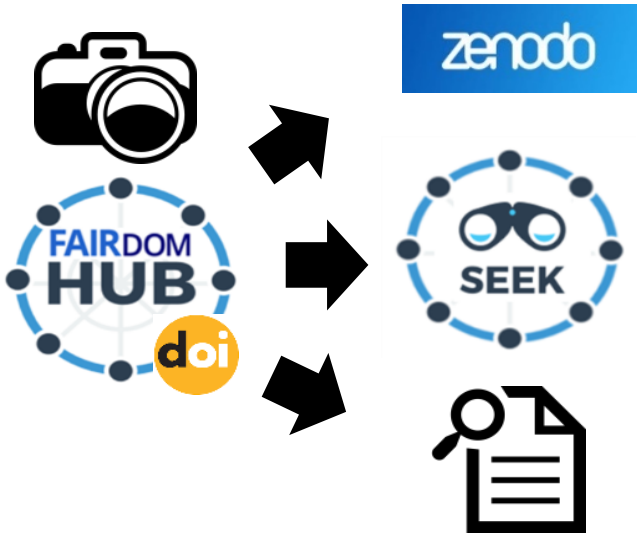
Distributed: Span repository silos, regardless of location and ownership
In house and External
Multi-site + multi-stewardship

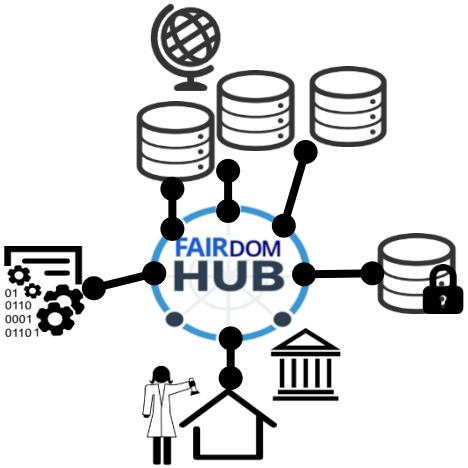
Structured organisation

Retaining context over fragmentation

Spanning across the Ecosystem







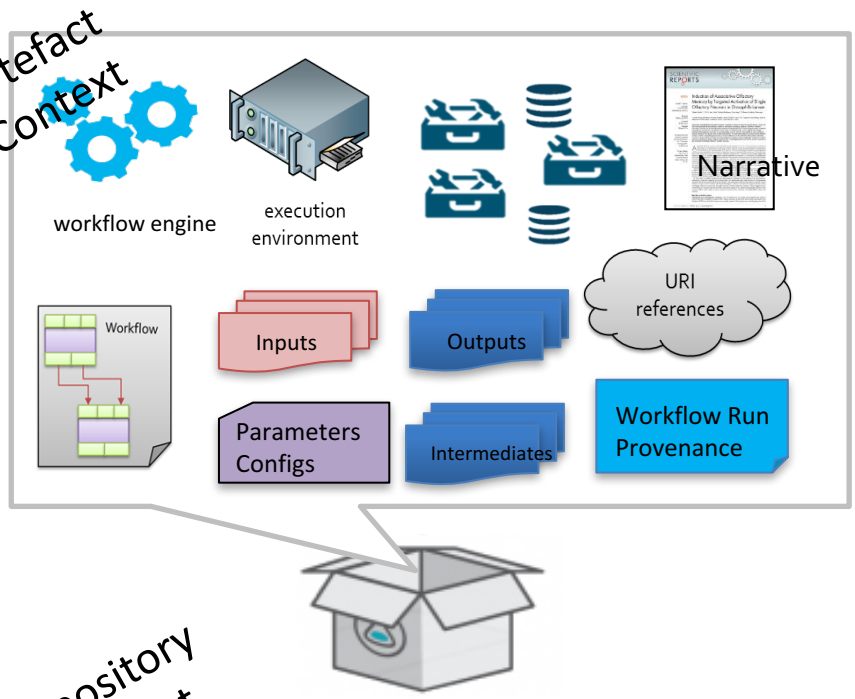
Context

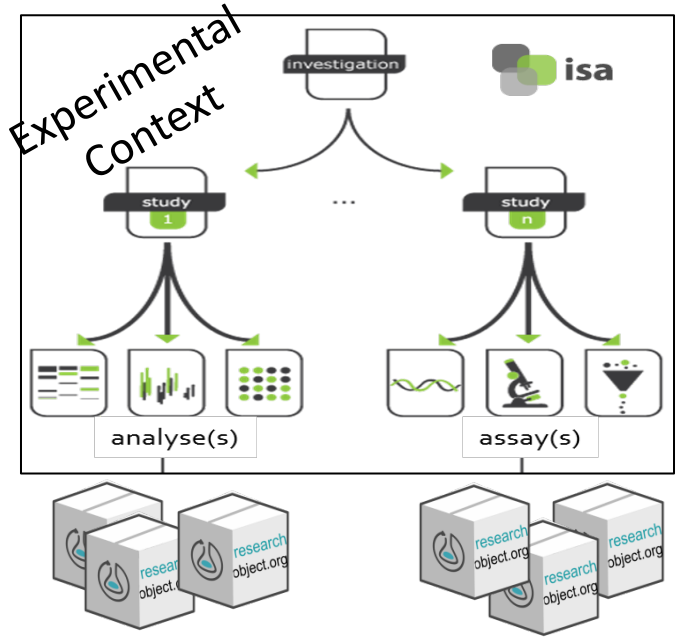
Publishing & Exporting

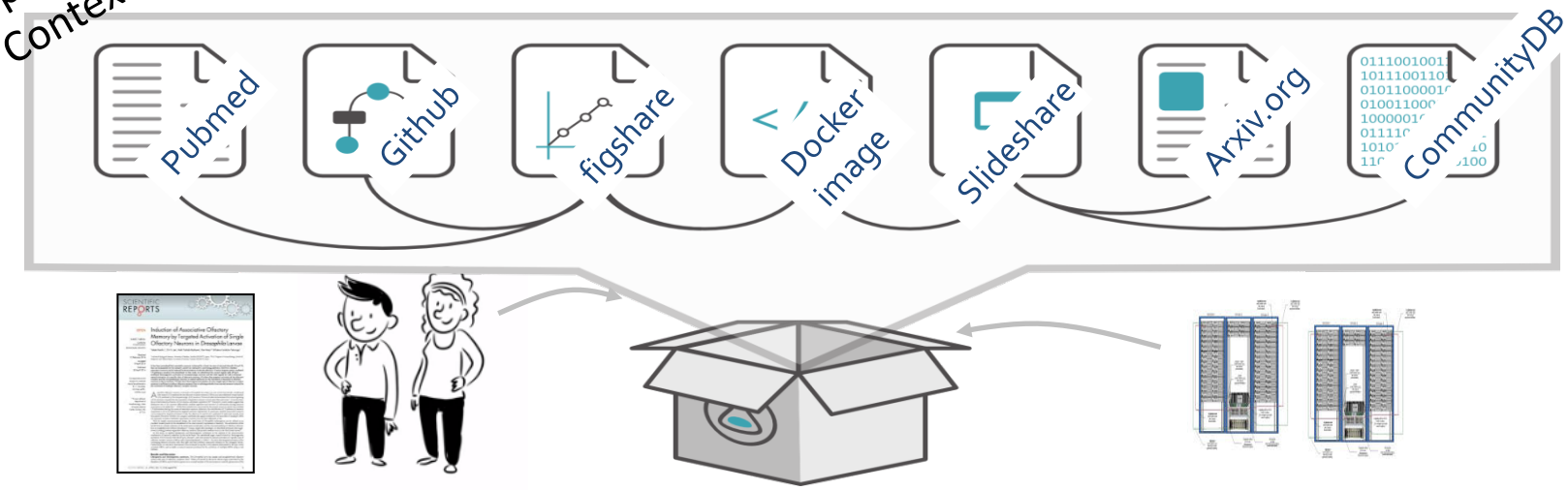
Snapshot versions and elements

DOI proliferation

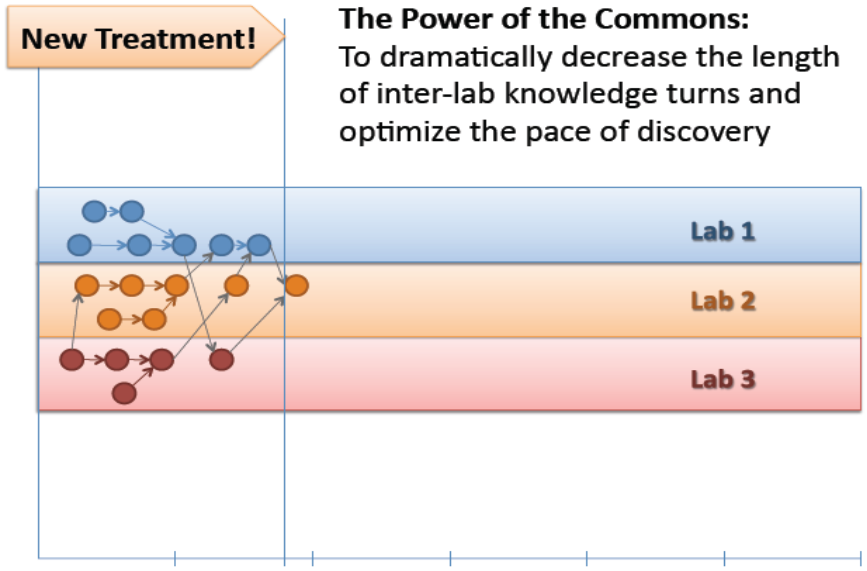
Digital Object Perspectives

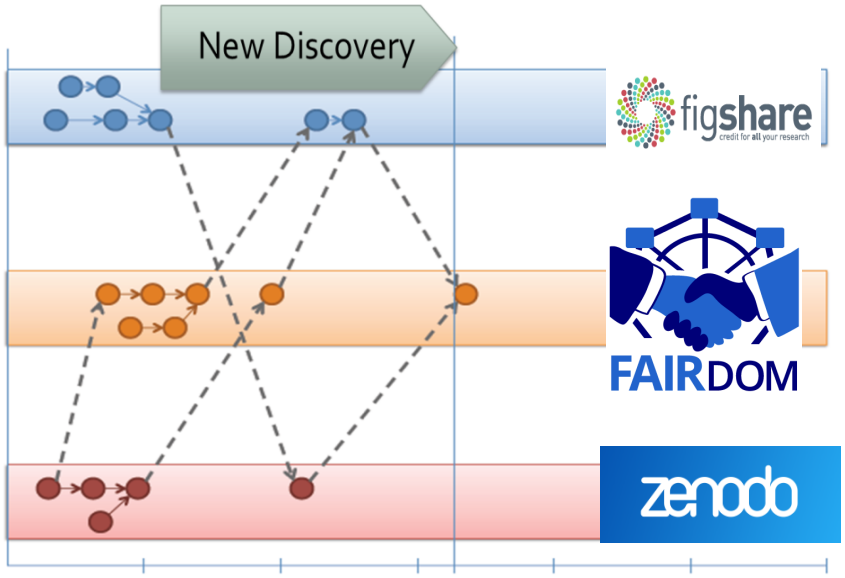






preserved, portable research products.
inter-platform exchange
multi-platform content & context dependencies

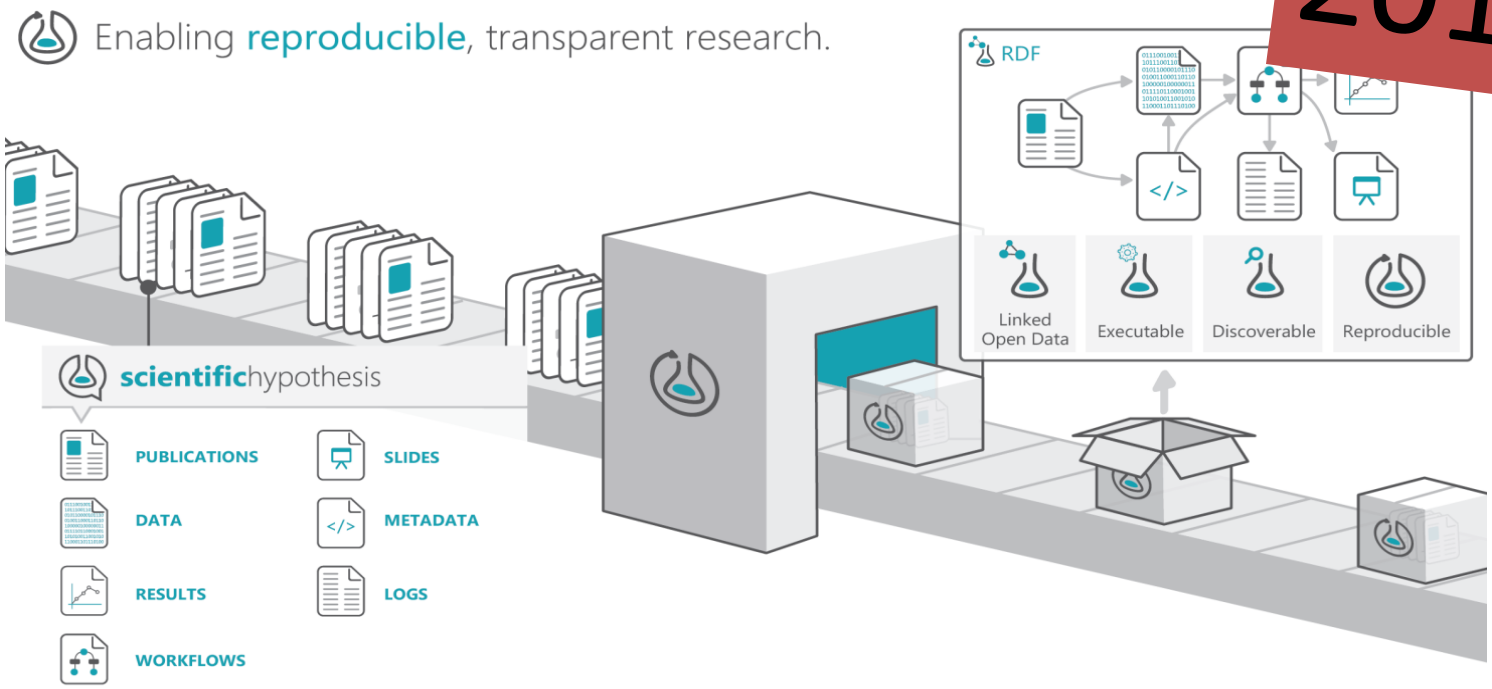




[Josh Sommers]

researchobject.org

2010



- **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, to improve understanding & interpretation

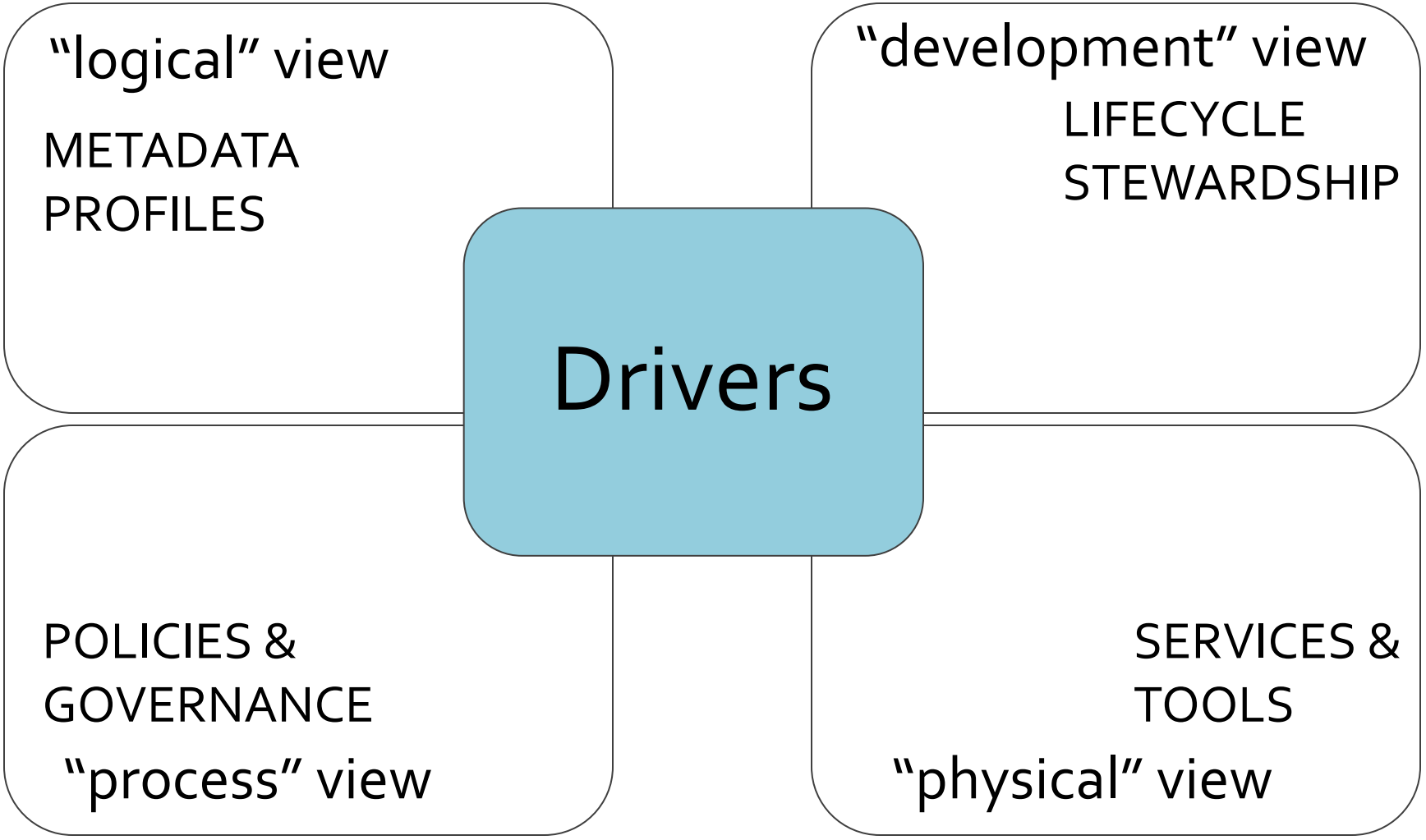
Bechhofer et al (2013) Why linked data is not enough for scientists https://doi.org/10.1016/j.future.2011.08.004
Bechhofer et al (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge, https://eprints.soton.ac.uk/268555/

Workflow 4Ever

Atomicity,
Granularity,
Aggregation
Composition
Fragmentation
Versioning
Forking
Cloning
Portability
Dependency
management

Research Objects
Analogous to
Software
artefacts and practices
rather than
Data or Articles

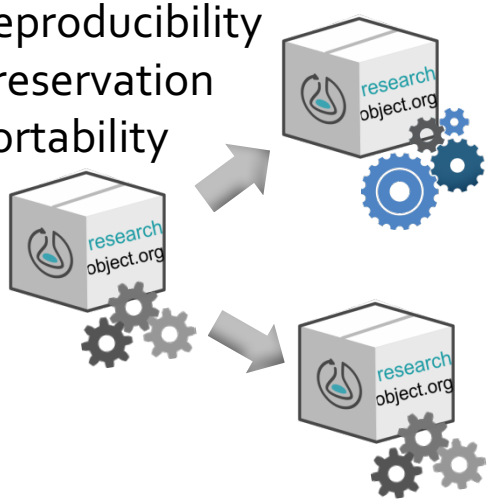
Research Objects
4+1 Architecture Framework



Drivers within the Ecosystem

Commons & Catalogues
Publishing,
Exchange between
people and platforms
Sharing,
Training

Replication
Reproducibility
Preservation
Portability



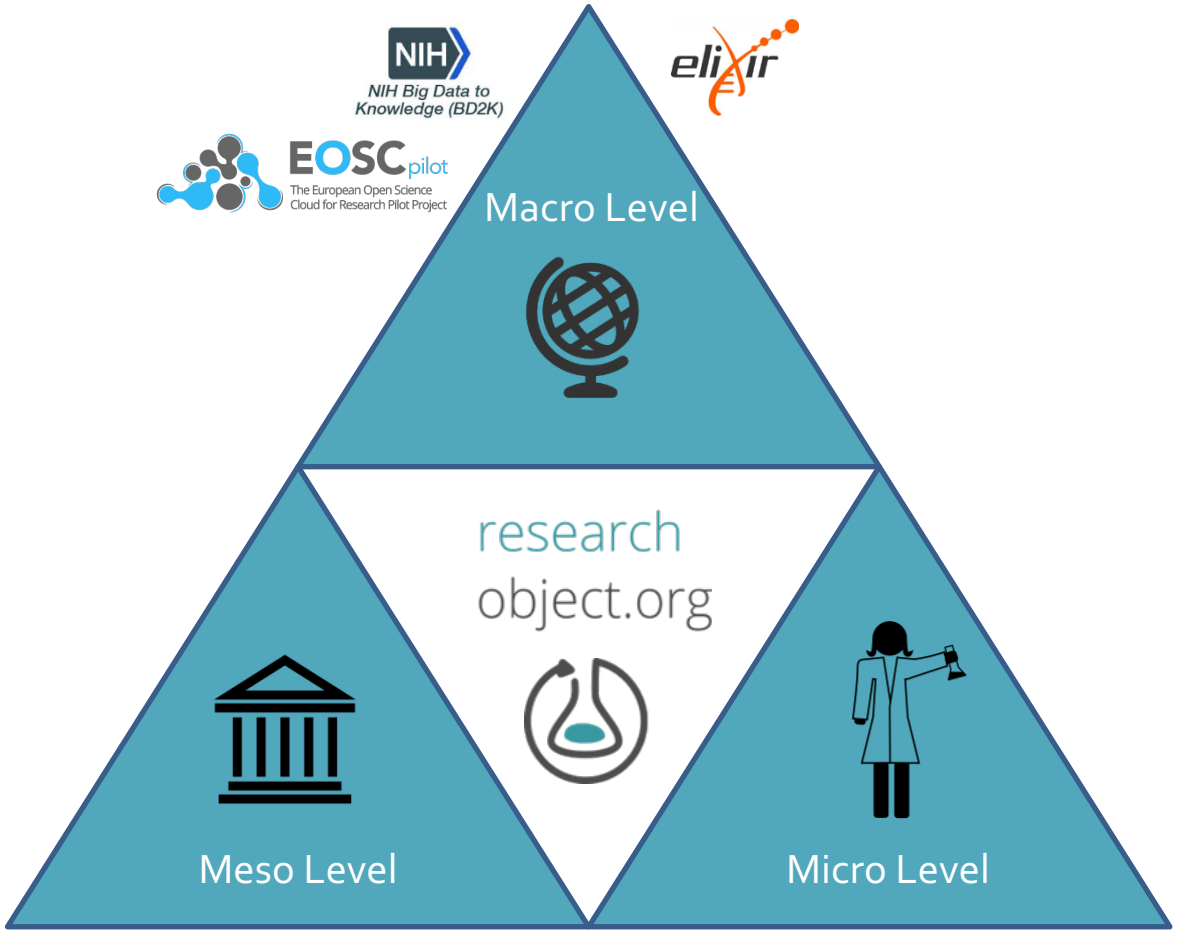
Active Research Release
Evolution & Snapshots
Remixing, Comparison, Review
Automated processing

Conservation
Repair
Archive

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

ROs working across the ecosystem

(Inter)National Commons
Public Community Archives
Publishers



Boundary Objects

Institutional repositories

Personal Electronic Lab Notebooks,
Project / Group Commons
Research Context

Knowledge Exchange Report: http://www.knowledge-exchange.info/event/ke-approach-open-scholarship
The 'last mile' challenge for European research e-infrastructures https://doi.org/10.3897/rio.2.e9933

FAIR ROs Desiderata

Use Open Standards.

Incremental.

Machine-processable.

Graceful degradation.

Technology Independent.

Multi-platform.

Commodity tooling.

The least possible.

The simplest feasible. Low tech.

Low user overhead and thin client

Manifests of Metadata

Manifest Construction

Identification
to locate and cite
Aggregates
to enumerate
& link together
ROs and elements
Annotations
about RO, elements &
their relationships

Manifest Profile Description

Type Checklists
what should be there
Provenance
where it came from
Versioning
its evolution
Dependencies
what else is needed

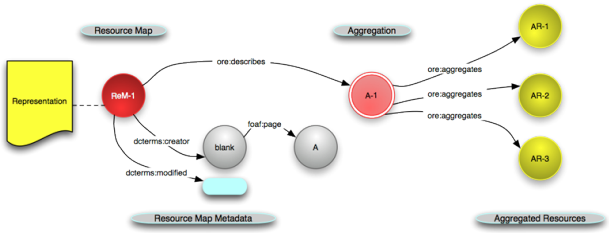
Container

Standards & COTS Platforms

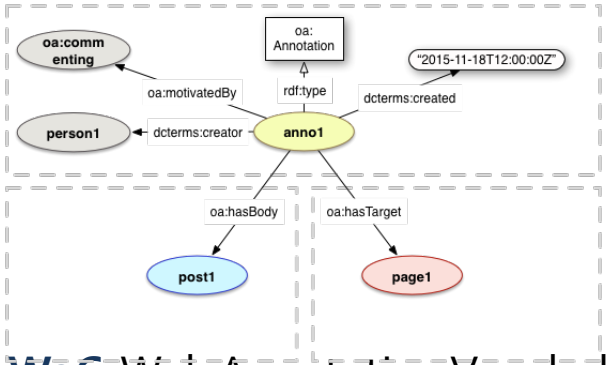
Identifiers: URI, RRI, DOI, ORCID

Identifiers.org

Construction



Open Archives Initiative
Object Exchange and Reuse
Aggregation



W3C Web Annotation Vocabulary
Annotation

Container

BagIt

docker

VAGRANT

CONDA®

bioboxes

BIOCONDA®

Linking across ROs and into the Linked Open Data Cloud

- Recording & linking together the components of an experiment
- Linking across experiments.
- Linked ROs

- Semantic Web + Digital Objects

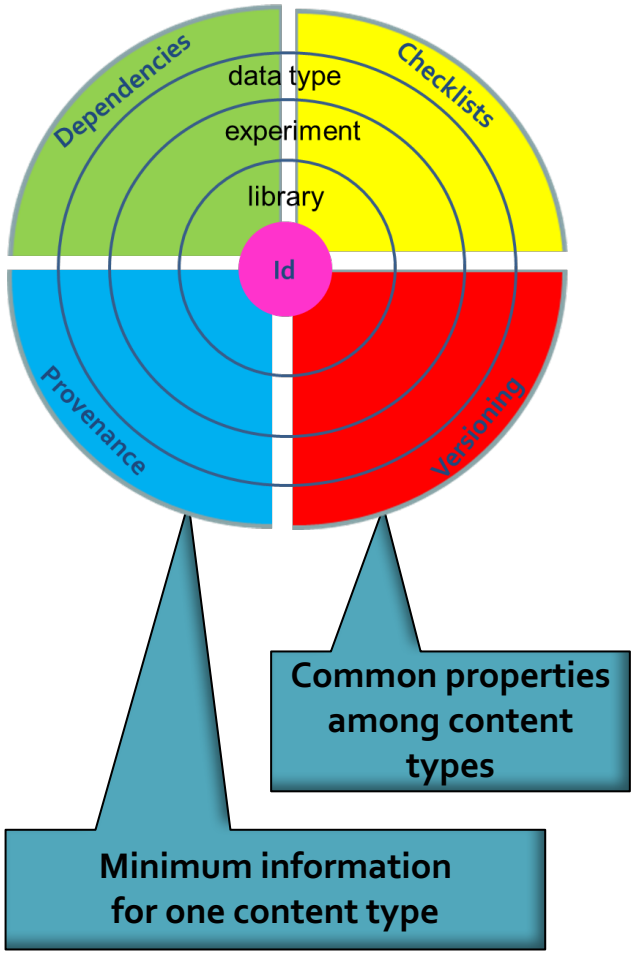
Goldilocks Profiles & Progression Levels to define and interpret content

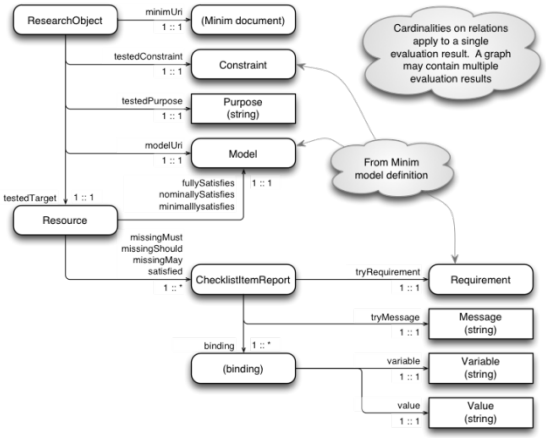
its evolution	where it came from	what else is needed	what should be there for types	
Versions	**Provenance**	**Dependencies**	**Checklists**	Project / Lab Specific
Fedora, DSPACE, MENDELEY DATA, DIGITAL COMMONS, openBIS	Apache Taverna, KNIME, Galaxy PROJECT	puppet, CHEF, Scientific Linux, debian med	FAIRDOM, Galaxy	
SEEK, GitHub	wfprov, ProvONE, CWL	elixir Tools and Data Services Registry, BIOCONDA, BIOCONTAINERS, bioboxes	combine Archive, EDAM, mibbi, BioCompute Objects, COMMON WORKFLOW LANGUAGE, isatab, Bioschemas.org, <DATS>	Community-based Types, Context
doi, zenodo, PAV, figshare, ROHUB	ReproZip Package, W3C PROV, VoID	docker, vagrant	DC1 Data Citation Principles, <JATS>, schema.org, DataCite	All

Gamble, Zhao, Klyne, Goble. "MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data", IEEE eScience 2012 Chicago, USA October, 2012), http://dx.doi.org/10.1109/eScience.2012.6404489

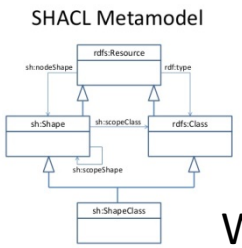
Profiles & Progression Levels

simplify the solution space
but still encode data types





Minim model for defining checklists

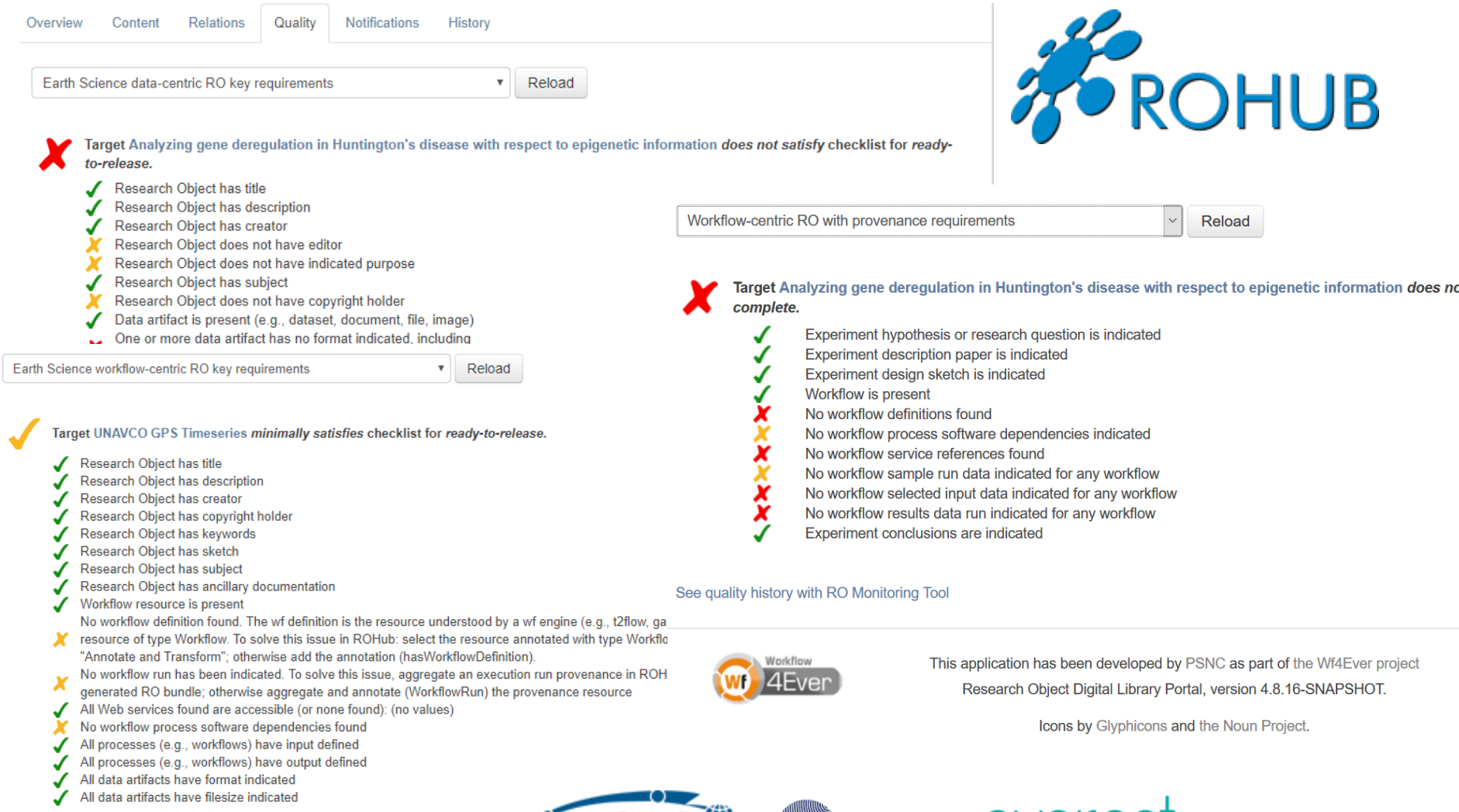


http://purl.org/minim/description

W3C Shape Specs

Gamble, Zhao, Klyne, Goble. "MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data", IEEE eScience 2012 Chicago, USA October, 2012), http://dx.doi.org/10.1109/eScience.2012.6404489

Validation and Monitoring Tools interpret the content

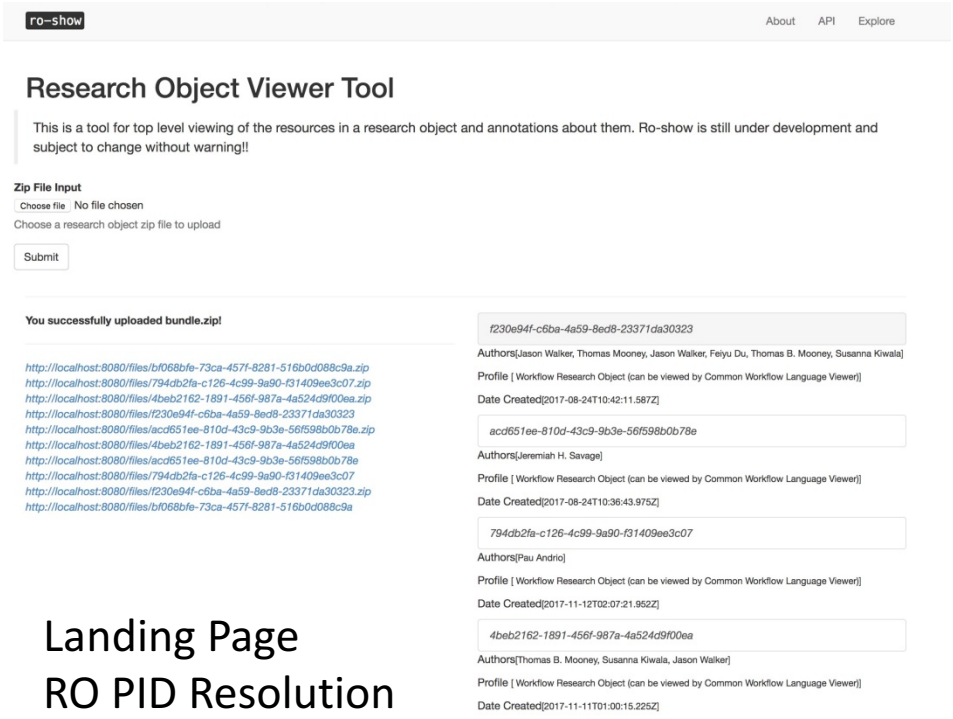


http://www.rohub.org/

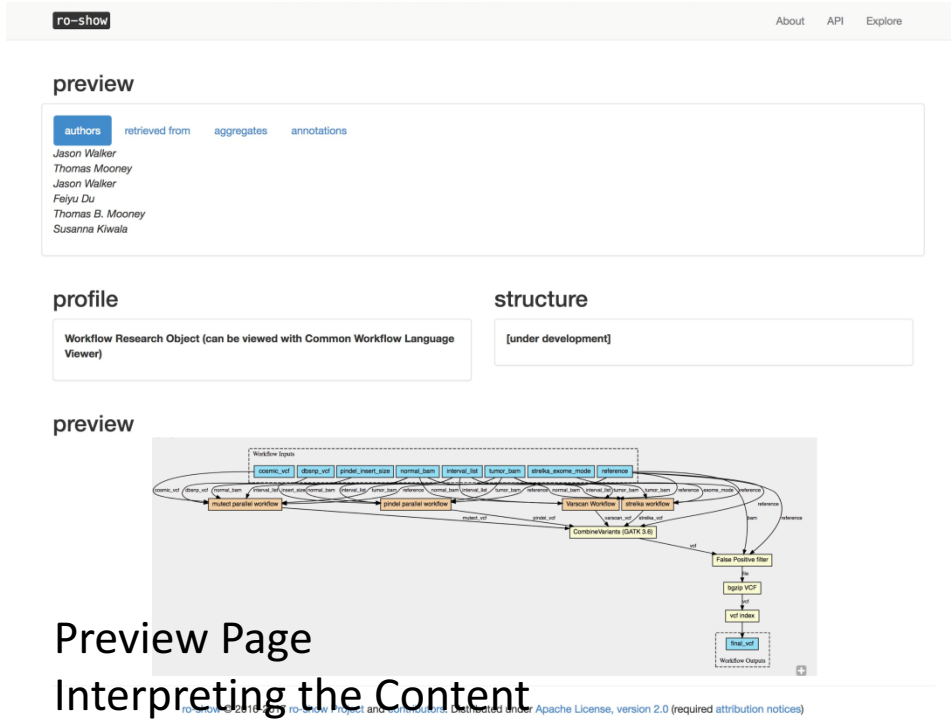
PSNC esa everest

[Raul Palma]

Generic Viewing Tooling interpret the content



Landing Page
RO PID Resolution

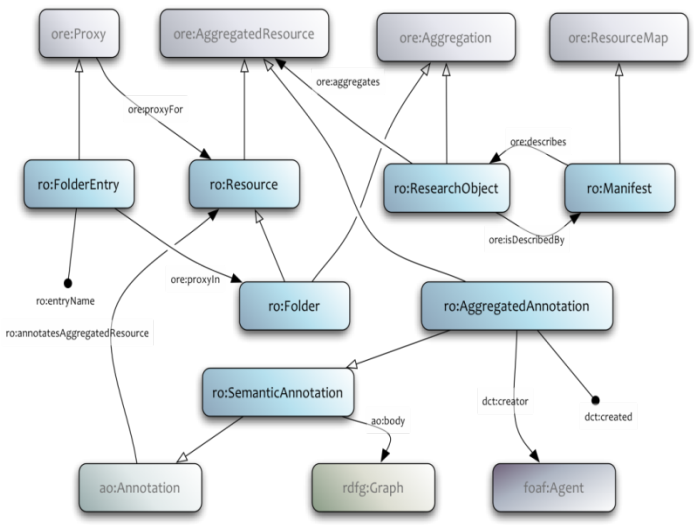


Preview Page
Interpreting the Content

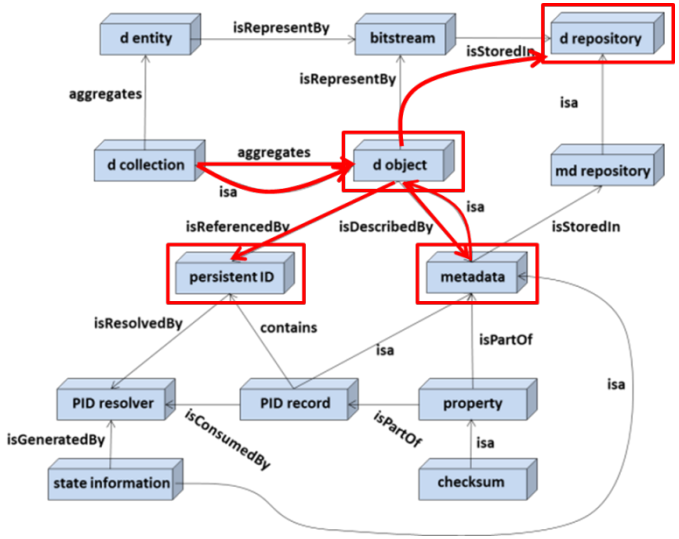
Making use of these various objects will depend on available infrastructure & tools etc.

[Lilian Gorea, Oluwatomide Fasugba]

Use the “conformsTo” Property



Research Object Manifest Model
http://www.researchobject.org/specifications/



RDA Data Foundation and Terminology WG Core model.

Bioschemas.org

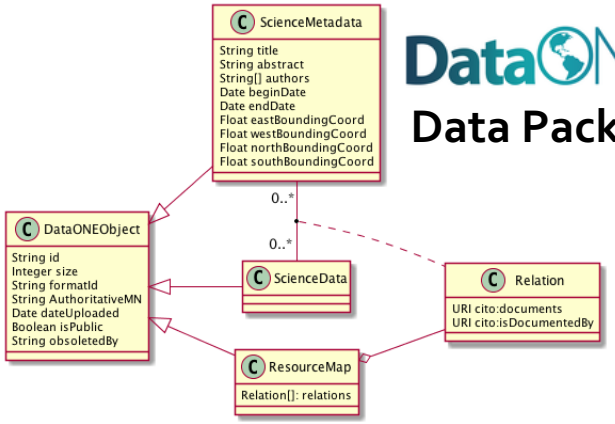
schema.org

<DATS>

FAIR DATA POINT

WICUS ONTOLOGIES

Dockeronto



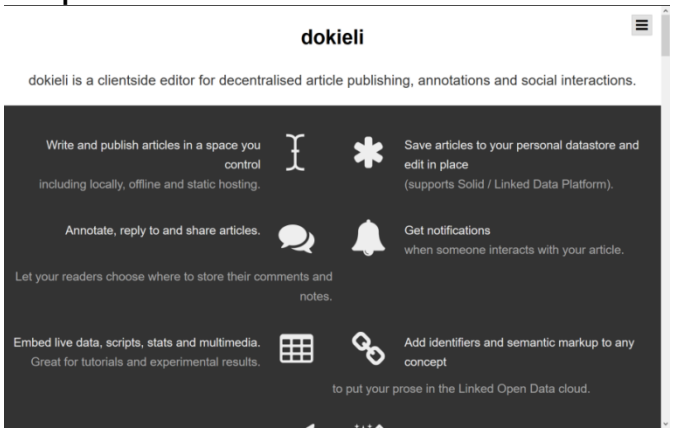
DataONE

Data Package

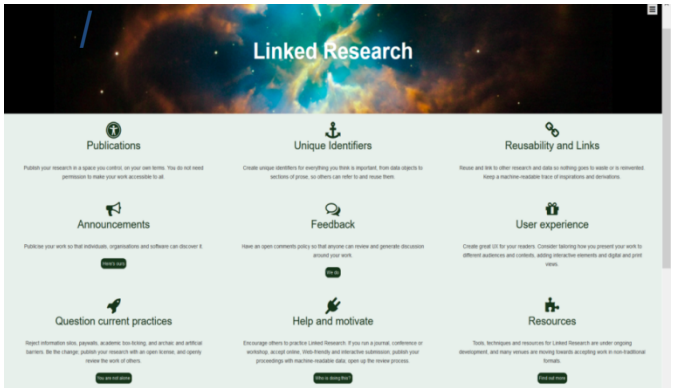
Lots of roads from A to B

A publishing trend…. JSON(-LD) + schema.org

https://dokie.li/



https://linkedresearch.org



Manifest: Schema.org, JSON-LD, RDF
Archive: .tar.gz

Reproducible Document Stack project
eLife, Substance and Stencila

DataCrate Specification

BagIT data profile + schema.org JSON-LD annotations

Bioschemas.org

schema.org tailored to the Biosciences

*simple **structured metadata markup** on web pages & sitemaps*

SCALE!

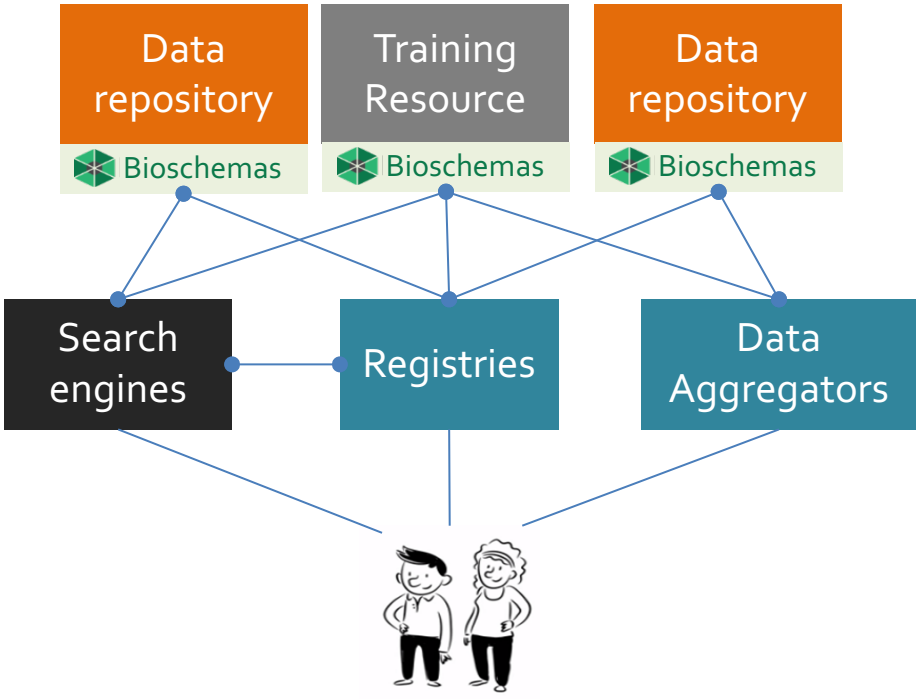
Finding, Citing, Metadata Exchange

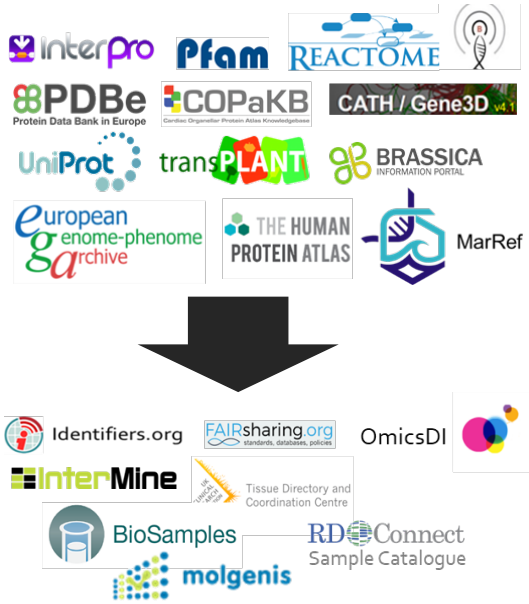
Bioschemas **Profiles**

Schema.org **Types**

Standardised metadata mark-up

Metadata published & harvested without APIs or special feeds





FORCE11
The Future of Research Communications and e-Scholarship

NIH Big Data to Knowledge (BD2K)

elixir

Bioschemas.org

schema.org tailored to the Biosciences

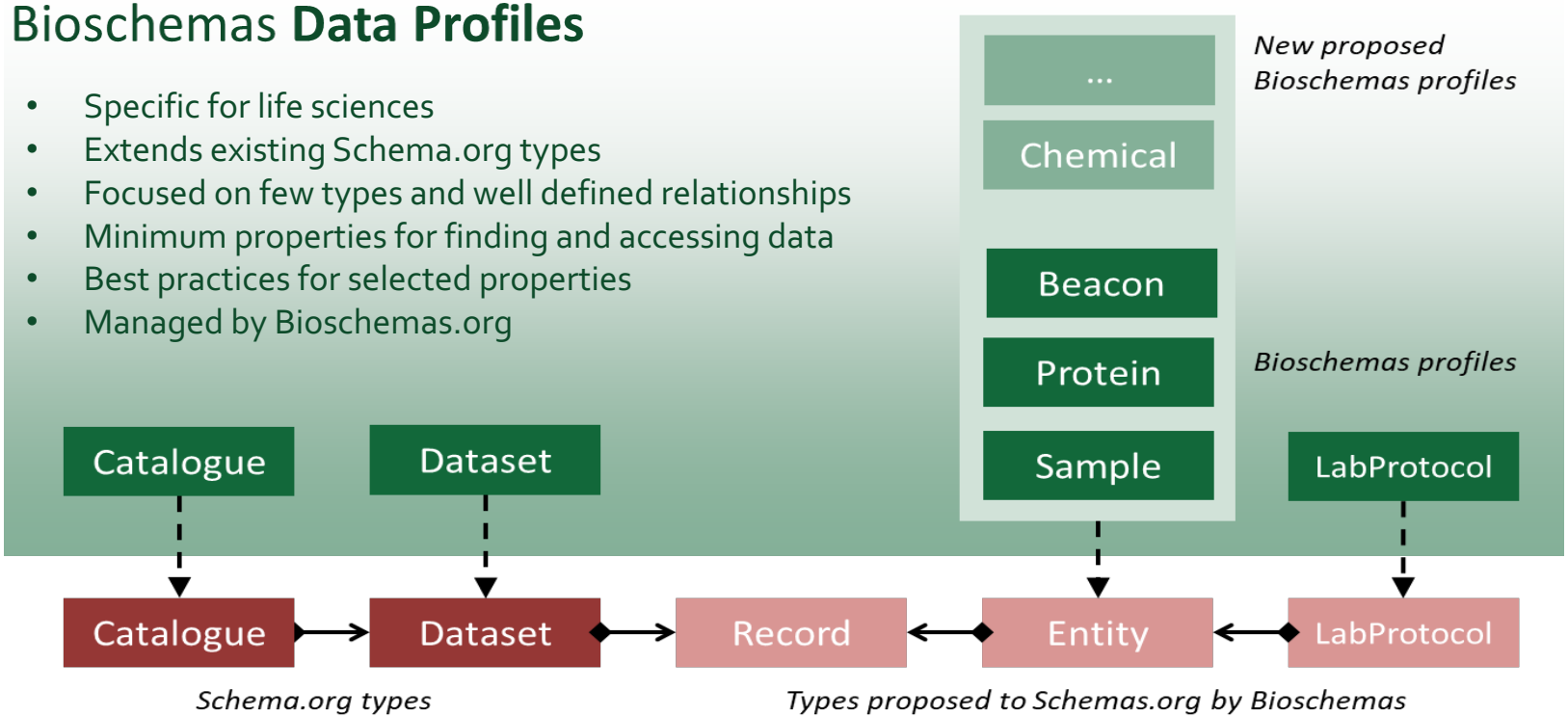
*simple **structured metadata markup** on web pages & sitemaps*

don't register – harvest & index

SCALE!

Bioschemas **Data Profiles**

- Specific for life sciences
- Extends existing Schema.org types
- Focused on few types and well defined relationships
- Minimum properties for finding and accessing data
- Best practices for selected properties
- Managed by Bioschemas.org



Schema.org **Types**

- Generic data model
- Generous list of properties to describe data types
- Managed by Schema.org

Research schemas

Proposal

Common Research **Types**

Common Research **Profiles**

Specific Research **Profiles**

- Bioschemas
- Agroschemas
- Astroschemas
- Earthschemas
- Biodiversityschemas…

EOSC pilot
The European Open Science Cloud for Research Pilot Project

Bioschemas.org

Encyclopedia of Life

AGU American Geophysical Union

Tissue Directory and Coordination Centre

Maintain common profiles across scientific domains focused on finding and accessing data and exchanging metadata in catalogues.
Serving Cloud Services & Supporting Boundary Objects

BioCompute Objects

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Precision medicine
NGS pipelines regulation

FDA U.S. FOOD & DRUG ADMINISTRATION

NIH NIH Big Data to Knowledge (BD2K)

BDDS

ISI, USC

Metagenomics pipelines and LARGE datasets

STELAR

Asthma Research e-Lab sharing and computing statistical cohort studies

Farr The Farr Institute of Health Informatics Research

Public Health Learning Systems

FAIRDOM

ISA based packaging, snapshotting, exporting and publishing for Systems Biology models

the combine computational modeling in biology network

Export RO from WebCAT

U Rostock

Standardised packaging of Systems Biology models

esa ROHUB

European Space Agency RO Library
Everest Project

COMMON WORKFLOW LANGUAGE

Manifest description of CWL workflows

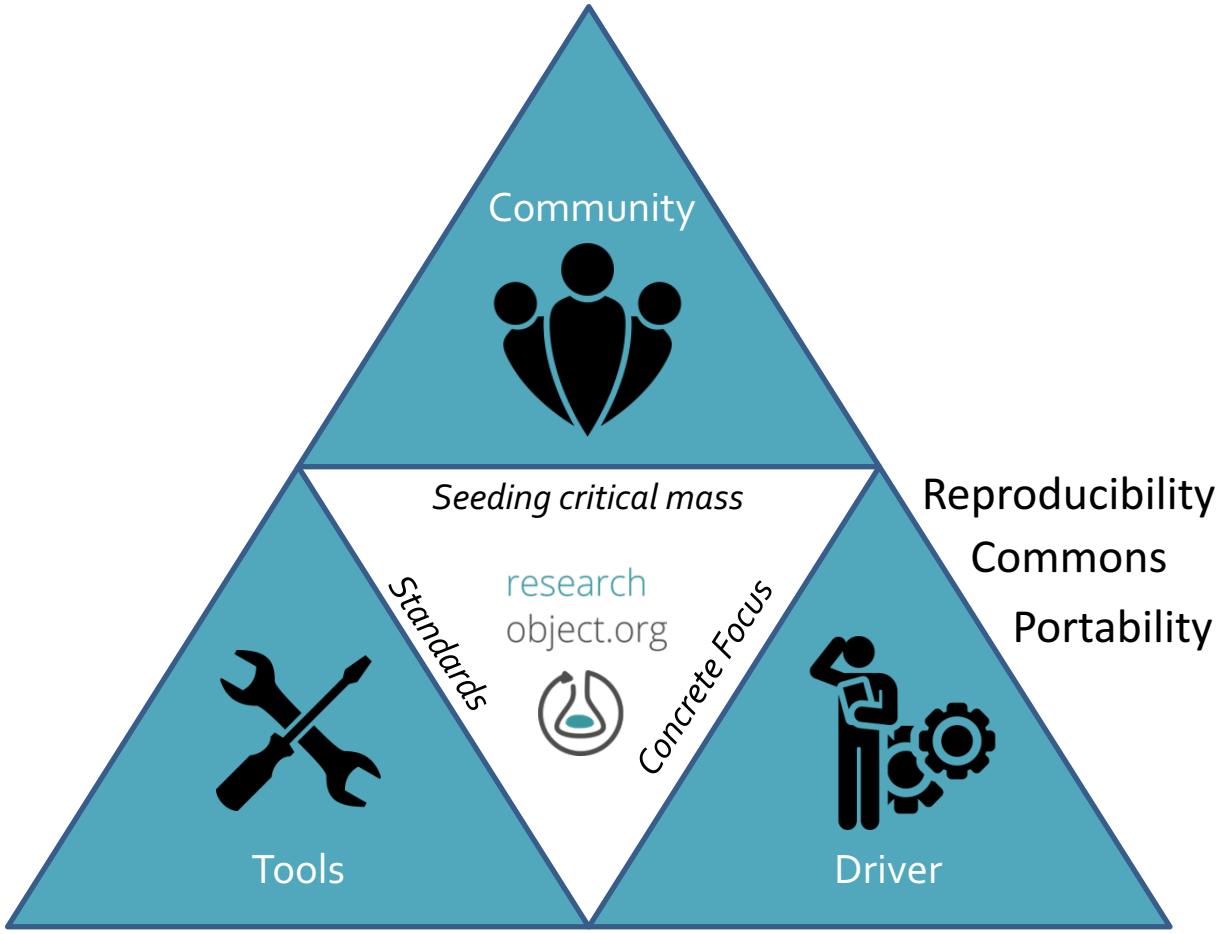
Research Object Bundles for Data Releases
Dataset "build" tool

Open PHACTS
Open Pharmacological Space

ROs acceptance in the ecosystem

Platform & user buy-in from the get-go

Passionate, dedicated leadership



Reproducibility

Commons

Portability

Easy to make

Hard to consume

Generic vs Specific

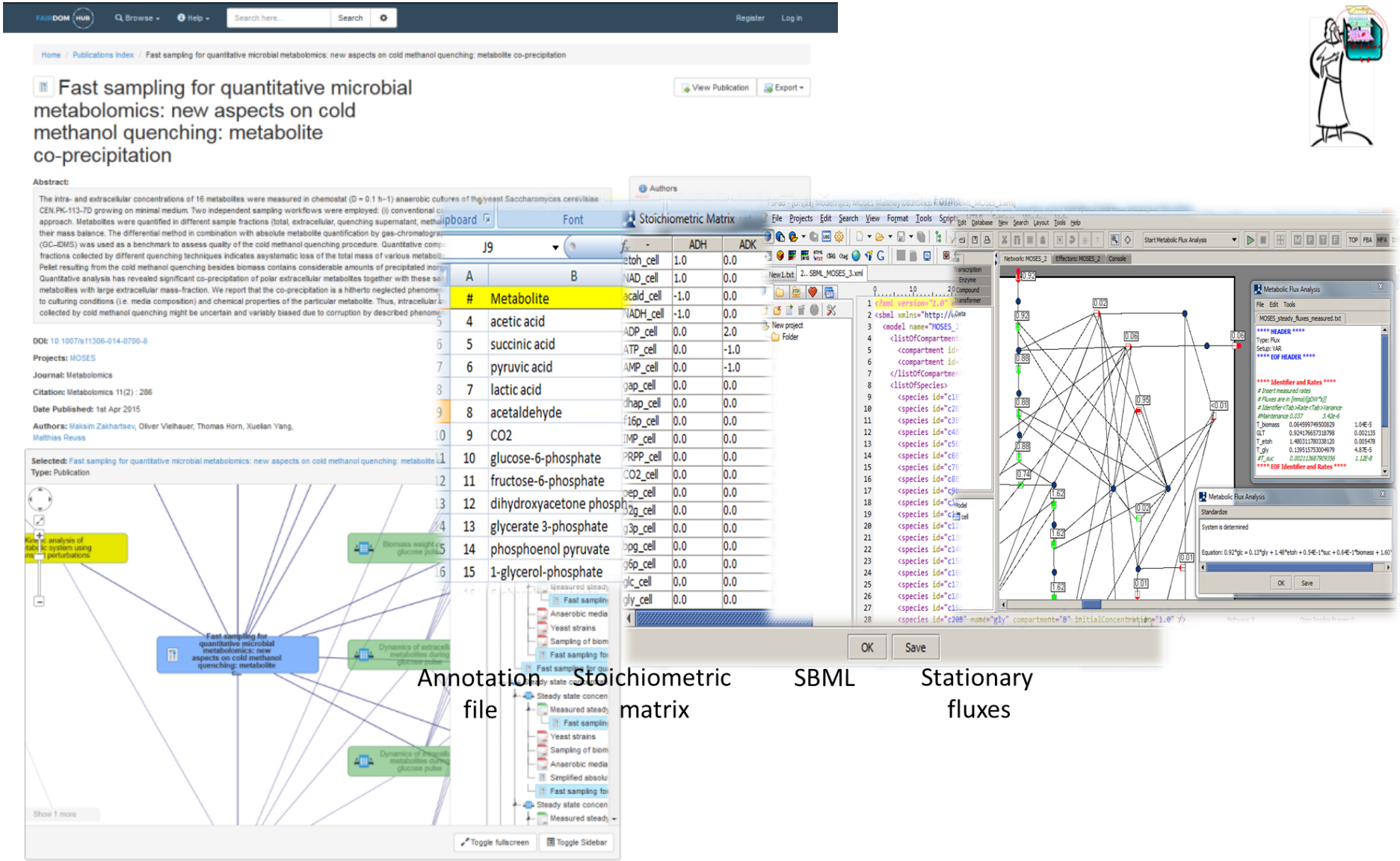
Don't be too flexible!

Complex Objects types

Multi-artefact Objects

Computational Workflows & Pipelines

Multi-disciplinary investigations



Annotation file

Stoichiometric matrix

SBML

Stationary fluxes

FAIRDOM

[Maxim Zakhartsev]

Computational Workflow Research Objects

COMMON WORKFLOW LANGUAGE

researchobject.org

Community led standard way of expressing and running workflows and the command line tools they orchestrate, supporting containers for portability.

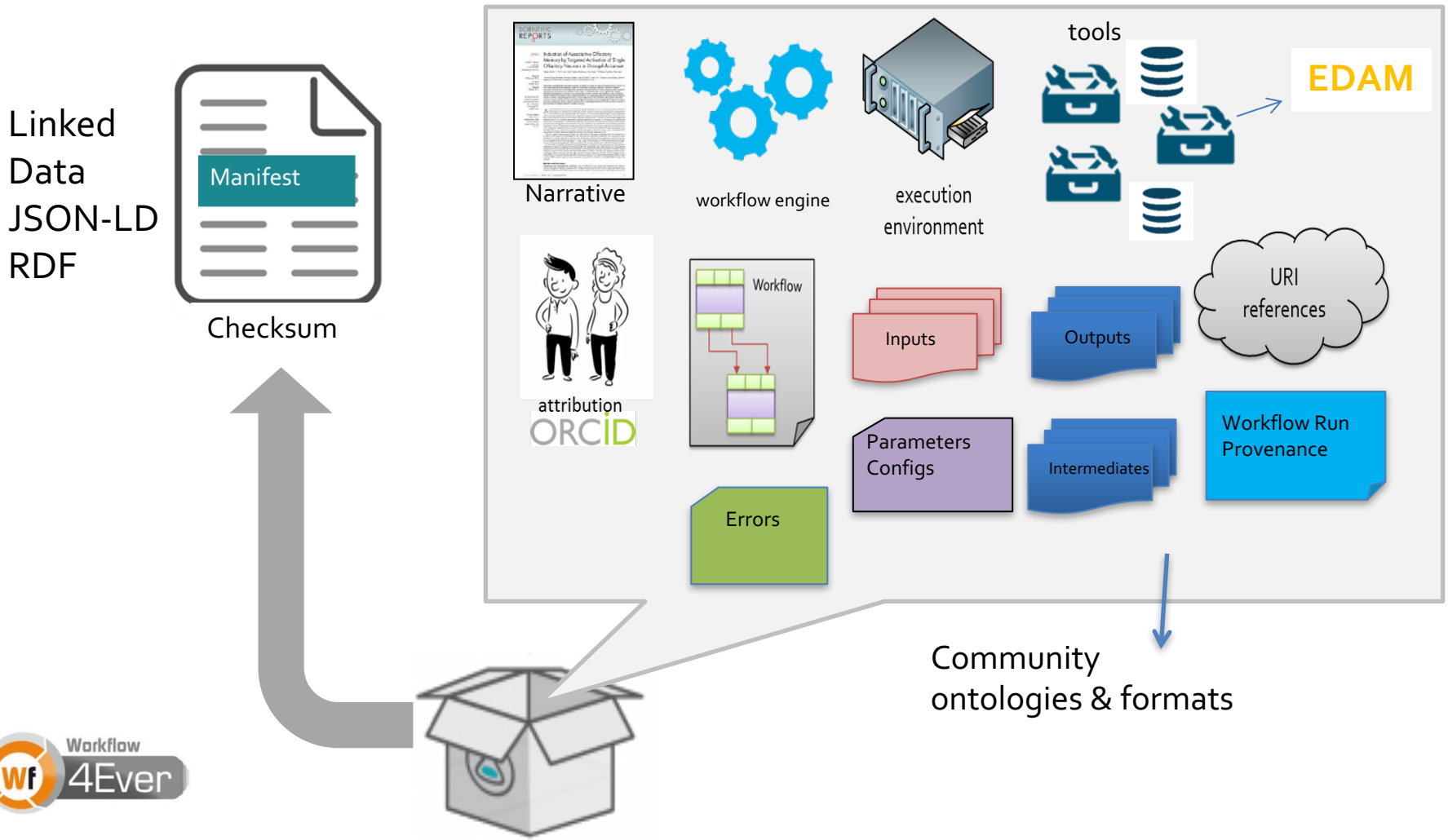
Gathers CWL workflow descriptions together with rich context and provenance using multi-tiered descriptions

Snapshots the workflow.

Relates to other objects.

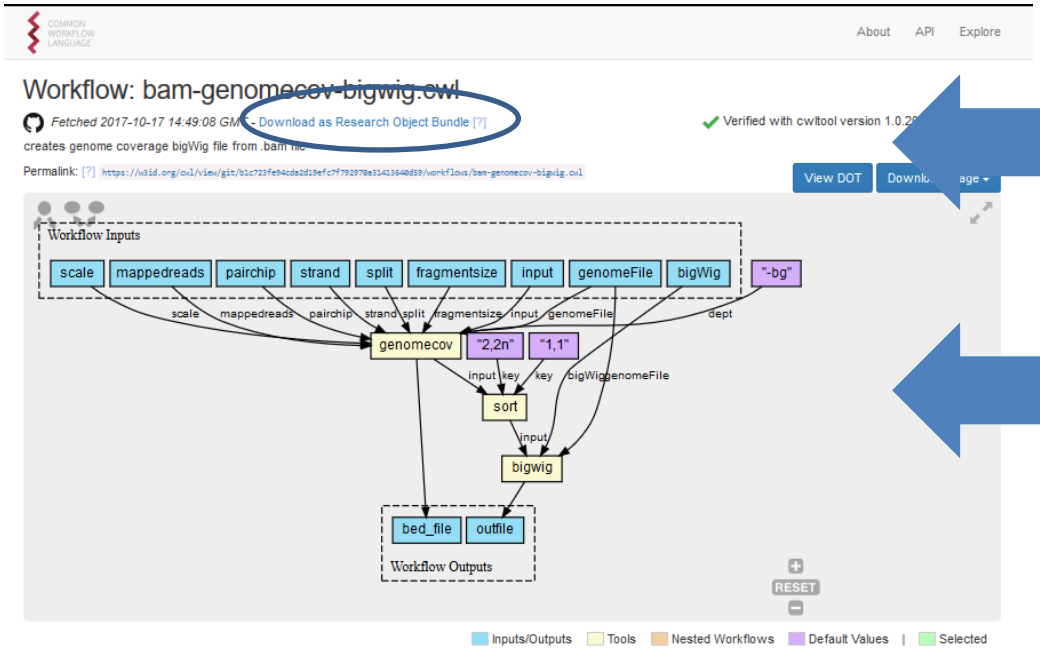
schema.org

Reports + Machines : Workflow Research Objects



Belhajjame et al (2015) Using a suite of ontologies for preserving workflow-centric research objects, J Web Semantics doi:10.1016/j.websem.2015.01.003

Hettne KM, et al (2014), Structuring research methods and data with the research object model: genomics workflows as a case study. J. Biomedical Semantics 5: 41



Download as a Research Object Bundle

Permalink to snapshot the GitHub entry and RO identifier

Inputs

ID	Type	Label	Doc
genomeFile	File		
input	File		
mappedreads	Double (Optional)		
strand	String (Optional)		
split	Boolean (Optional)		
pairchip	Boolean (Optional)		
bigWig	String (Optional)		
scale	Float (Optional)		
fragmentsize	Integer (Optional)		

Steps

ID	Runs	Label	Doc
genomecov	../tools/bedtools-genomecov.cwl (CommandLineTool)		Tool is used to calculate
bigwig	../tools/ucsc-bedgraphtobigwig.cwl (CommandLineTool)		Tool is used to convert
sort	../tools/linux-sort.cwl (CommandLineTool)		Tool is used to run sort

Over an **active GitHub** entry for an **actively developing** workflow

Special Tooling: Common Workflow Language Viewer

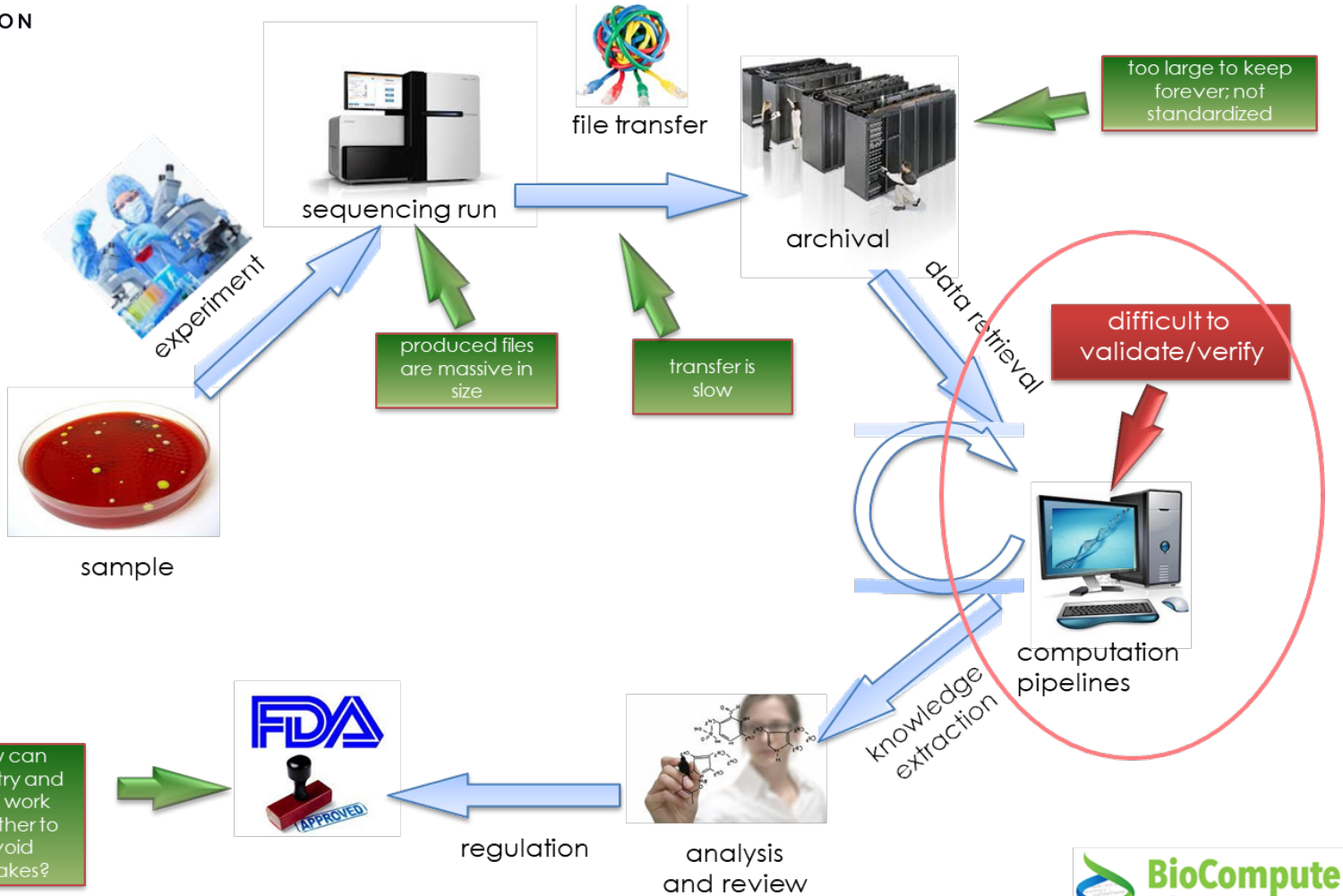
https://view.commonwl.org/

bioexcel
Center of Excellence for Computational Biomolecular Research
elixir

BioCompute Objects

U.S. FOOD & DRUG ADMINISTRATION

Precision Medicine
High Throughput Sequencing, from a biological sample to biomedical research and regulation



[Vahan Simonyan]

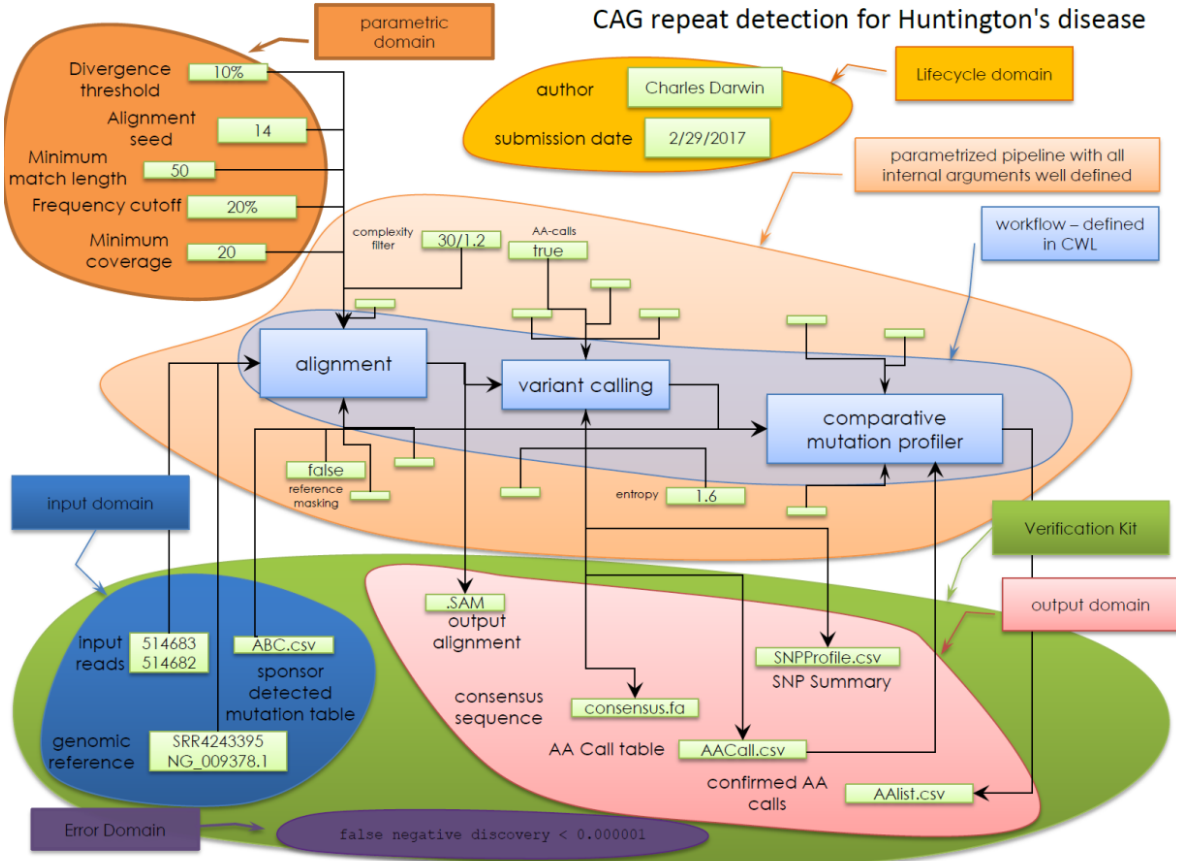
FDA BAA contract HHSF223201510129C (PI: Raja Mazumder)

BioCompute Objects

Precision Medicine High Throughput Sequencing, from a biological sample to biomedical research and regulation

U.S. FOOD & DRUG ADMINISTRATION

Emphasis on parametric domain and robust, safe reuse.



Linked Data, JSON-LD, Ontologies (EDAM, SWO)

+

researchobject.org

+

COMMON WORKFLOW LANGUAGE

+

FHIR® Global Alliance for Genomics & Health

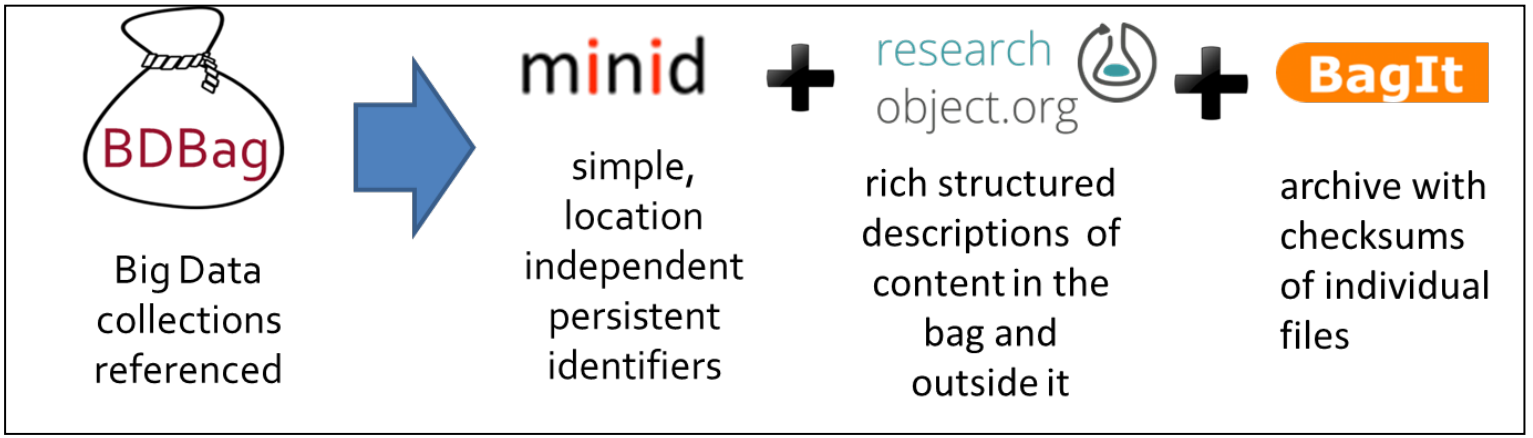
data formats, elements and APIs for EHR & genomics

Alterovitz, Dean II, Goble, Crusoe, Soiland-Reyes et al Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results, biorxiv.org, 2017, https://doi.org/10.1101/191783

Ecosystem of tools and services for big data analysis and sharing in an ecosystem

Assemble, share, and analyze large and complex multi-element datasets to integrate into biomedical HTS analytic pipelines

Secure large scale moving of patient data



1000s of images and genome sequences assembled from diverse repositories,

data *distributed* across multiple locations, *referenced* because big and persisted, *efficiently moved* by Grid technologies

Chard et al I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets, https://doi.org/10.1109/BigData.2016.7840618

BDDS

NIH Big Data to Knowledge (BD2K)

minid

Identifier:
ark:/57799/b9j01d

Created:
2016-08-06 15:02:12.180428

Creator:
BDBag Service (None)

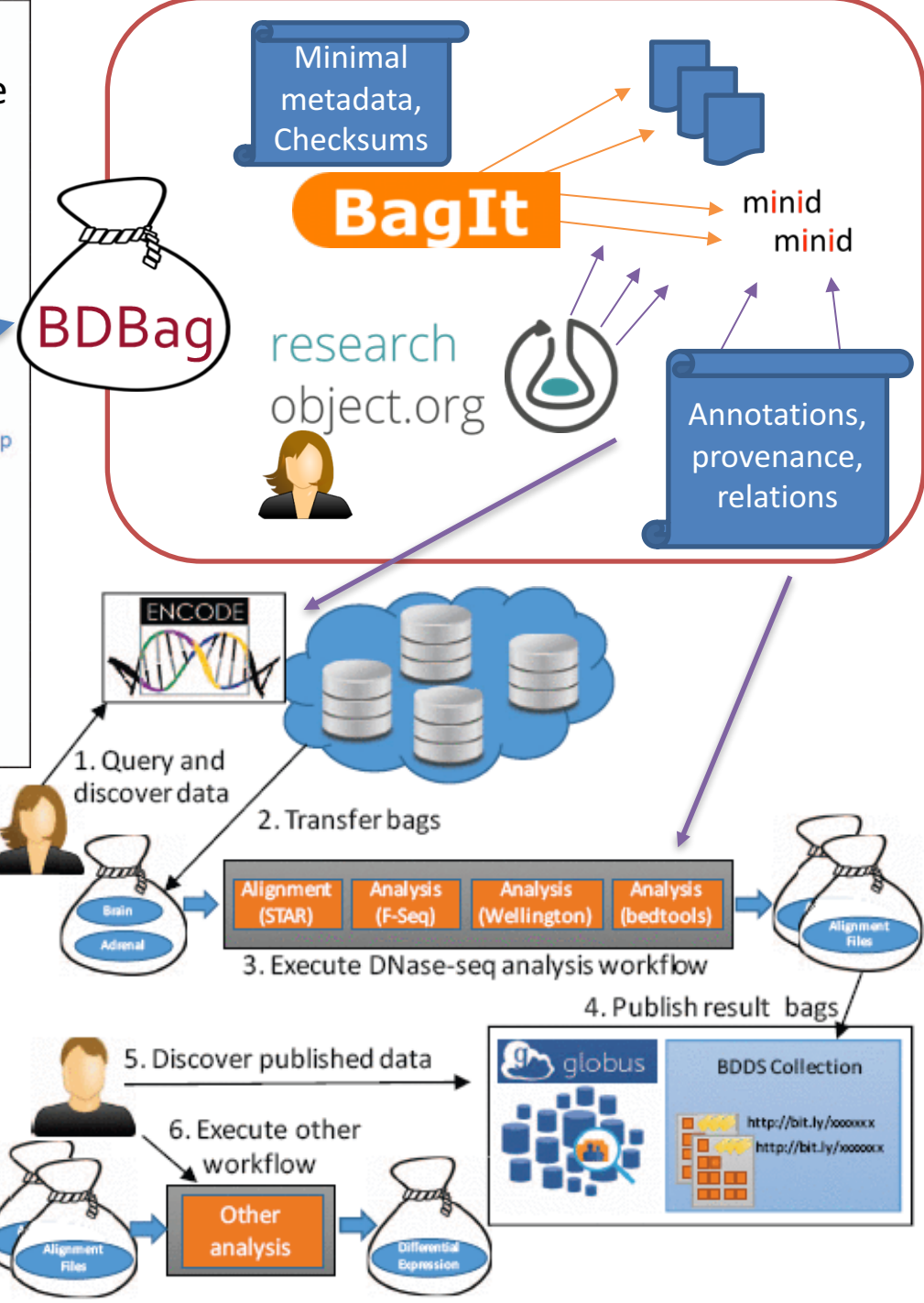
Checksum:
8bdc4180dd08ff5bf21cd8a20096e0c44ca69d38e95337e1a72edb0c0c817c15

Status:
ACTIVE

Locations:
https://s3.amazonaws.com/encode-bags/c5c7084b-bf91-4b5a-b673-07de38d8388a.zip

Titles:
ENCODE BDBag

Archival Resource Key



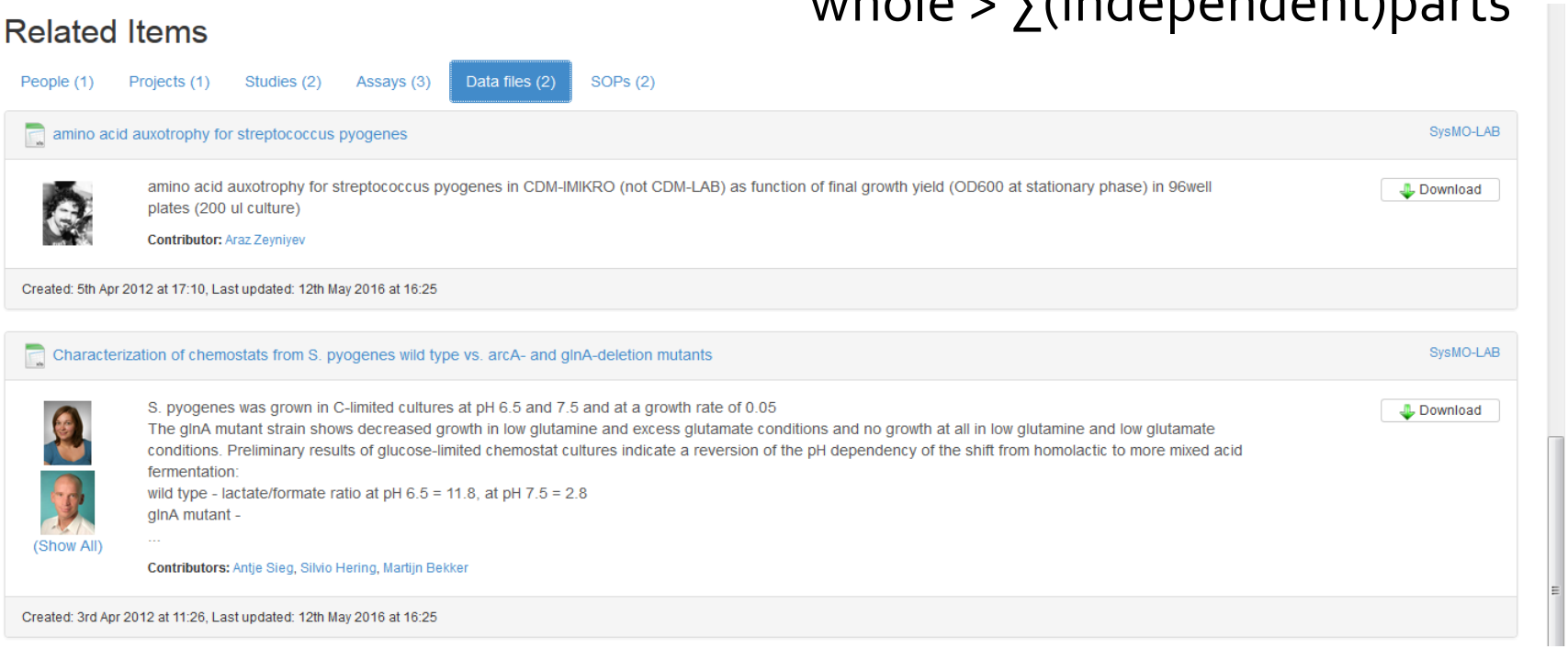
BDDS

NIH Big Data to Knowledge (BD2K)

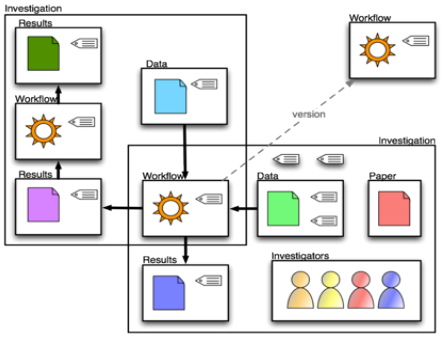
Chard et al: **I'll take that to go:** Big data bags and minimal identifiers for exchange of large, complex datasets, https://doi.org/10.1109/BigData.2016.7840618

Management in the Ecosystem

whole > ∑(independent)parts



Inspirations:
software development &
old skool data integration



Tragedy of the Commons

Manifests of Metadata

- profile making
- template making
- template elements
- auto manufacture
- spreadsheet tooling

https://metadatacenter.org

CEDAR

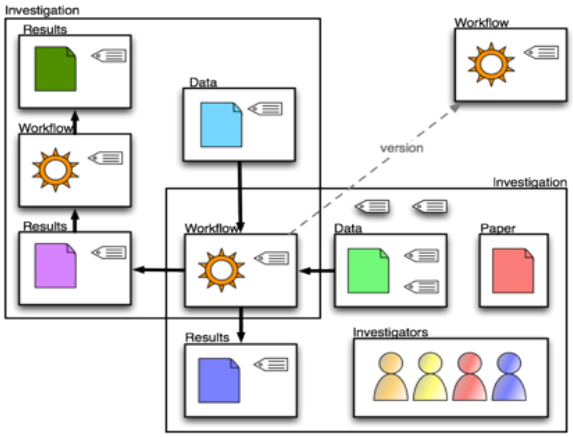
"The challenge for all the data-commons initiatives — is that many online datasets are annotated with metadata that are simply terrible…. Creating good metadata takes considerable work ….

When investigators act in their own self-interest, taking short cuts to generate metadata as quickly as possible, we should expect that the overall utility of the resource will decline.

The creation of a data commons requires the ability to deal with extremely varied — and often unanticipated — metadata patterns and data types …. a need for easy-to-use solutions that are generic to provide guidance over the entire life cycle of metadata — streamlining metadata creation, discovery, and access, as well as supporting metadata publication to third-party repositories"

https://ncip.nci.nih.gov/blog/face-new-tragedy-commons-remedy-better-metadata/

Stewardship in a multi-component, evolving ecosystem



Dependencies & Responsibilities with multi-stewardship at different granularities

Who manages the RO and who manages and governs the parts?

Who maintains the manifests?

Delegation and trust!

Expect component rot

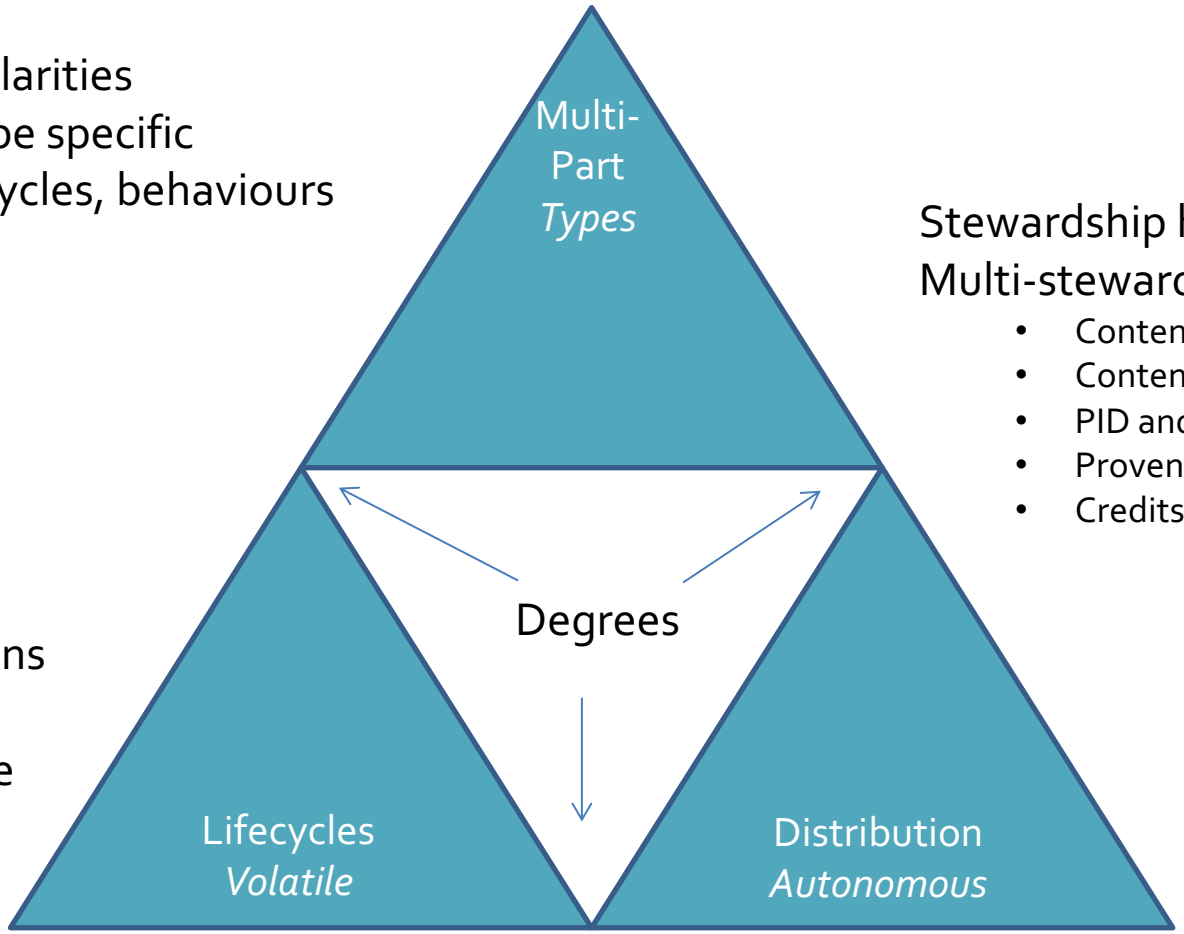
A new career?

research object.org

Multi-Stewardship of ROs and Elements and the stewardship of manifests.....

Different granularities
Domain and type specific standards, lifecycles, behaviours

Atomicity
Composition
Dependency



Stewardship hand-offs
Multi-stewardship guarantees

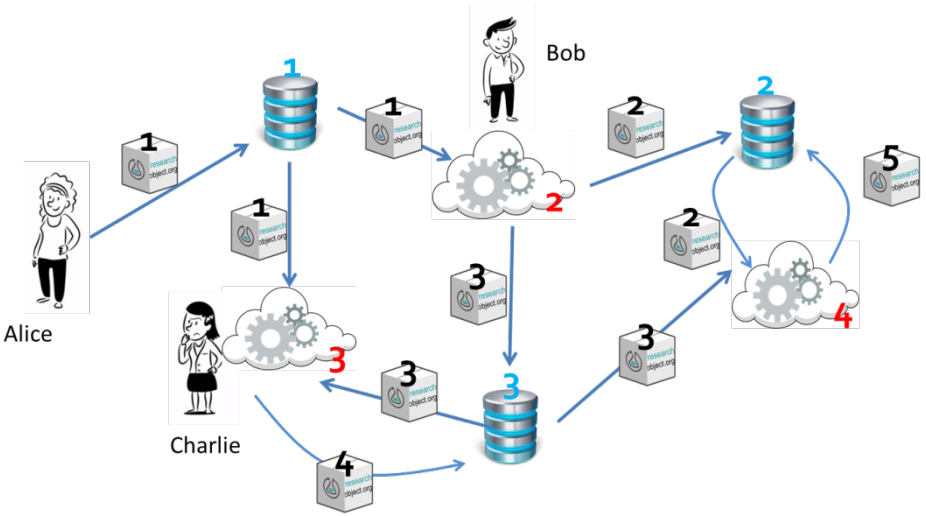
- Content change
- Content decay
- PID and Resolution services
- Provenance attribution
- Credits

Fixity, verifying intended versions of contents, element change detection, snapshots

Who is responsible? Spectrum of governance? Delegation and degradation.

research object.org

Creation, Credit, Curation



Missier, Data Trajectories: tracking reuse of published data for transitive credit attribution, IDCC 2016

Authenticity, Tamper-proofing

- Hashing & Checksums
- Secure signature & probity services
- Block chain & Ethereum

DOI proliferation

- Channelling for Counting
- Landing Pages

Katz and Smith "Contriponents"

- Micro-credit and citation aggregation
- Tracking RO usage & indirect contributions
- Awarding fractional weighted credit to contributors
- Networked Credit maps*

** D. S. Katz, "Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products," Journal of Open Research Software, v.2(1): e20, pp. 1-4, 2014. DOI: 10.5334/jors.be

Trend - bottom up initiatives sheltered by big umbrellas

- Grassroots community activities
- Fostered by Infrastructure Initiatives
- Don’t swash the start up!
- Open standards and lightweight
- Practical engineering
- Keeping it simple and real
- Ramps rather than Revolution

Barend Mons
Sean Bechhofer
Matthew Gamble
Raul Palma
Jun Zhao
Josh Sommer
Matthias Obst
Jacky Snoep
David Gavaghan
Stuart Owen
Finn Bacall
Paolo Missier
Phil Crouch
Oscar Corcho
Dan Katz
Arfon Smith

Mark Robinson
Alan Williams
Jo McEntyre
Norman Morrison
Stian Soiland-Reyes
Paul Groth
Tim Clark
Alejandra Gonzalez-Beltran
Philippe Rocca-Serra
Ian Cottam
Susanna Sansone
Kristian Garza
Daniel Garijo
Catarina Martins
Alasdair Gray
Rafael Jimenez
Iain Buchan
Caroline Jay
Michael Crusoe
Katy Wolstencroft

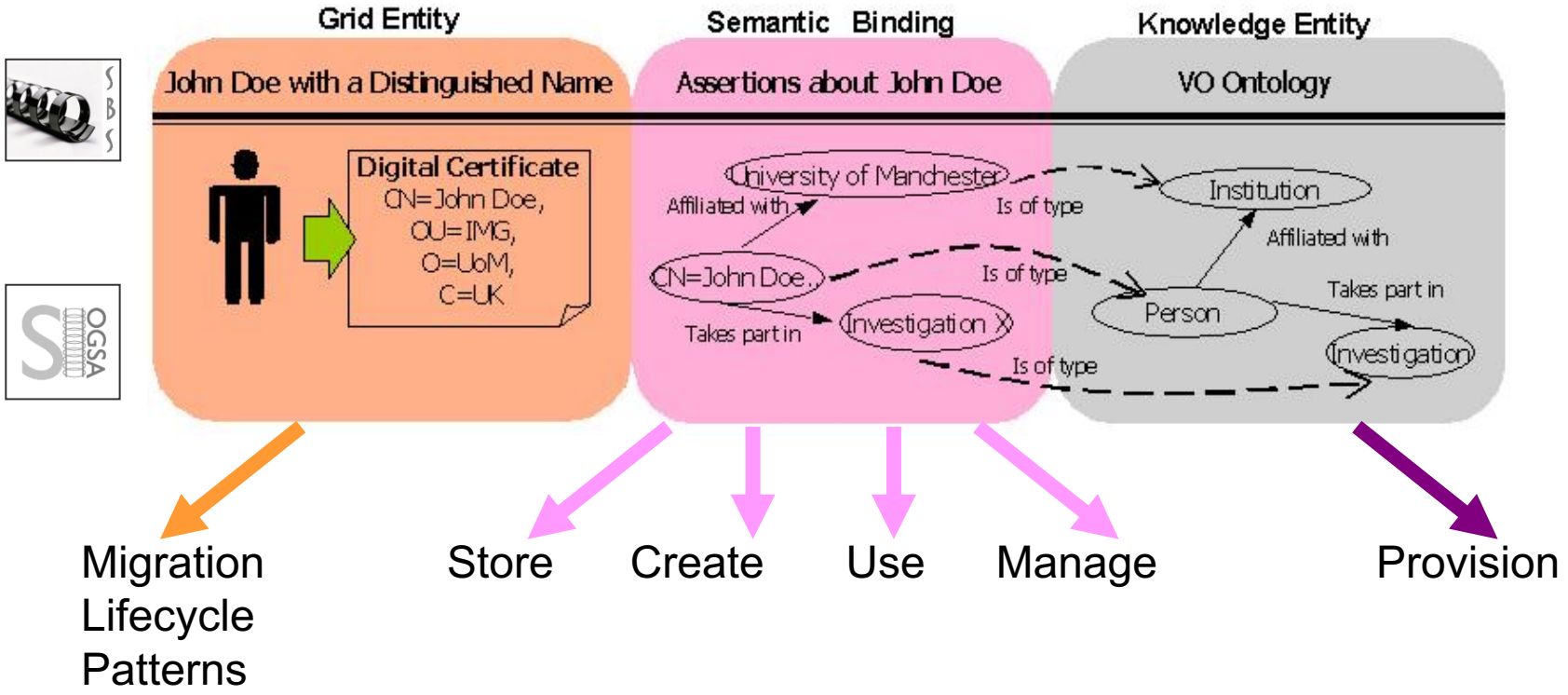
http://www.researchobject.org
http://wf4ever.org
http://www.fair-dom.org
http://seek4science.org
http://rightfield.org.uk
http://www.bioschemas.org
http://www.commonwl.org
http://www.bioexcel.eu

All the members of the Wf4Ever team
Colleagues in Manchester's Information Management Group, ELIXIR-UK, Bioschemas

researchobject.org

Semantic Bindings for non-embedded metadata

Bind Grid Entities and Knowledge Entities



An overview of S-OGSA: A Reference Semantic Grid Architecture
Oscar Corcho, Pinar Alper, Ioannis Kotsiopoulos, Paolo Missier, Sean Bechhofer, Carole Goble (2006) https://doi.org/10.1016/j.websem.2006.03.001

Beware the Bucket

Kathleen Fitzpatrick
@kfitz

Follow

One of the mistakes that's been made repeatedly in open scholarly communication projects has been the attempt to create the bucket of everything.

5:20 AM - 10 Nov 2017

4 Retweets **34** Likes

5 4 34

https://twitter.com/kfitz/status/928975762081898496

RO based ecosystems

Beyond islands with the odd special ferry-services (e.g. GitHub software releases archived in Zenodo to get DOI)

Bridges to connect anything-to-anything.
ROs carried between or collected in depots

Ecosystem evolution

- *Mitigate*
- *Adapt*
- *Migrate*
- *Die*

Ecosystem incentivisation

- *Love*
- *Money*
- *Fame*
- *By Side effect*