MANCHESTER
1824
The University of Manchester

The swings and roundabouts of a decade of fun and games with Research Objects

Carole Goble

The University of Manchester

Researchobject.org

carole.goble@manchester.ac.uk

DaMaLOS workshop 02 Nov 2020 at ISWC 2020

Special Acknowledgement

Stian Soiland-Reyes
The University of Manchester, UK

Our RO start up – what, why and how…

ROs In the Large – The Vision
A new form of Scholarly Communication.
RDM support throughout the research cycle.

ROs in the Small – The Implementation
Packaging digital components.
Referencing physical components.

Our World of FAIR Thematic Research Infrastructures (aka Cyberinfrastructure) – Biology, Biodiversity

EOSC-*Life*

elixir

bioexcel

IBISBA

SYNTHESYS+

DiSSCo

RSE

Research
Software
Engineers

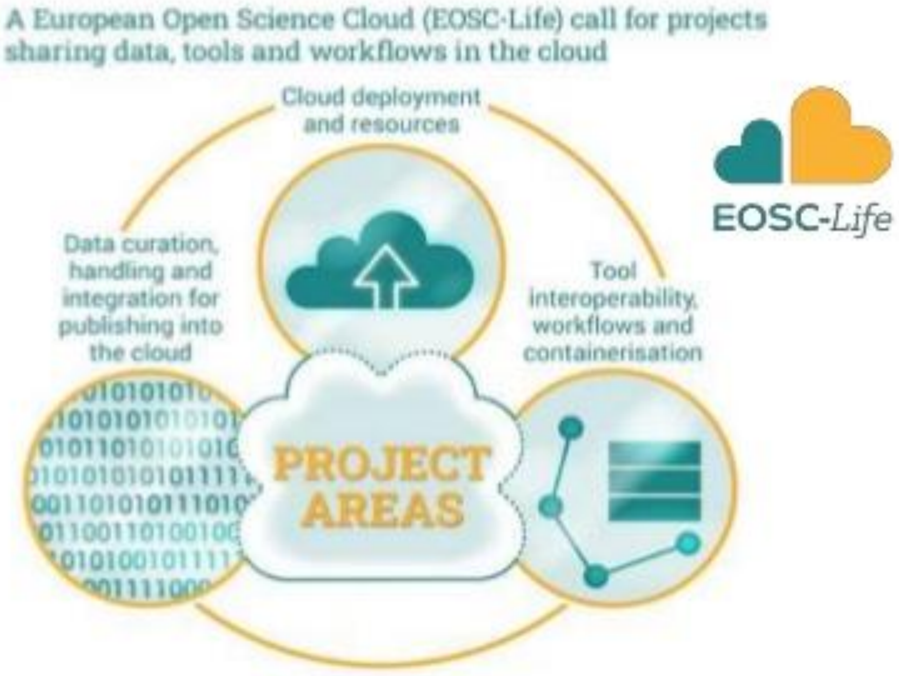
"facilities that provide resources and services for research communities to conduct research and foster innovation….they may be single-sited, distributed, or virtual.

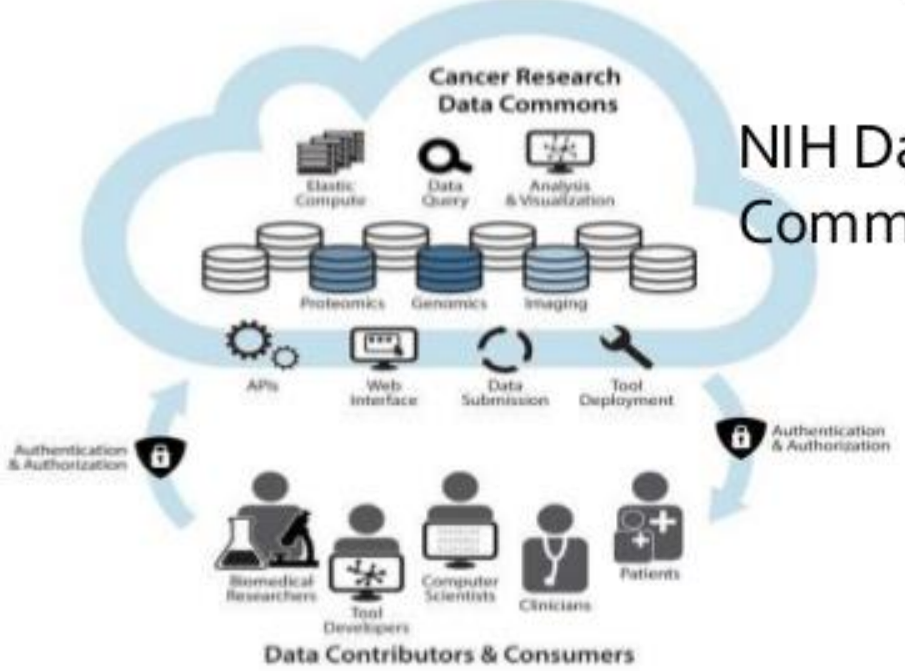
- major scientific equipment or sets of instruments
- collections, archives or scientific data
- computing systems and communication networks
- any other research and innovation infrastructure of a unique nature which is open to external users"

European Commission

https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures_en

FAIR Data Commons


A European Open Science Cloud (EOSC-Life) call for projects sharing data, tools and workflows in the cloud
Cloud deployment and resources
Data curation, handling and integration for publishing into the cloud
Tool interoperability, workflows and containerisation
PROJECT AREAS
EOSC-Life



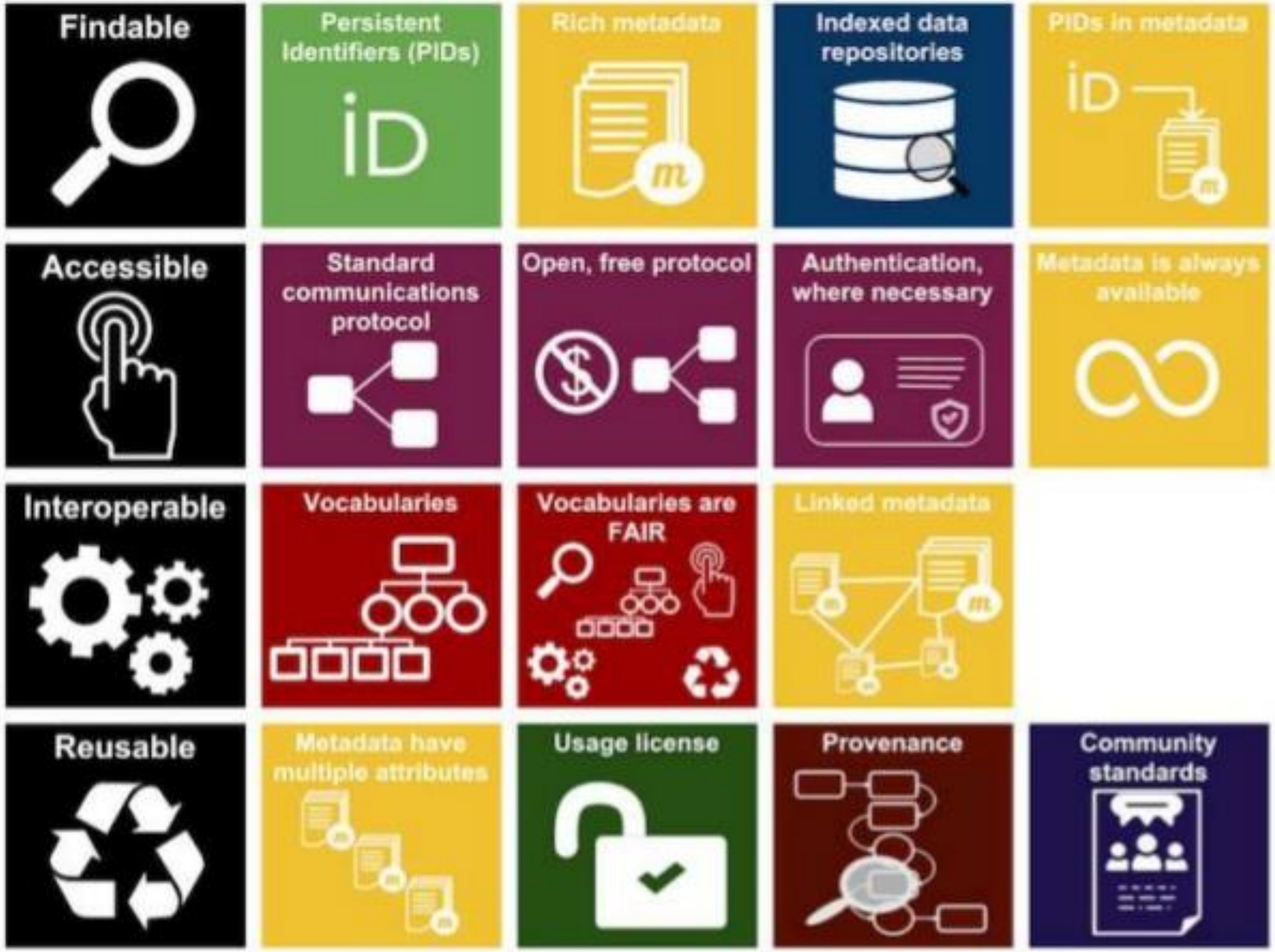
NIH Data Commons

Reproduce, port, share, and execute **analytics & pipelines**

Assemble and share **large scale, multi-element datasets.** Secure referencing and moving of sensitive data. Zoo of catalogues & resources.
Across 13 Research Infrastructures.

Diverse Research Objects – models, data, pipelines, lab protocols and SOPs, provenance… citable, exchangeable, publishable, preserved, executable objects and collections of objects.

FAIR Digital Objects

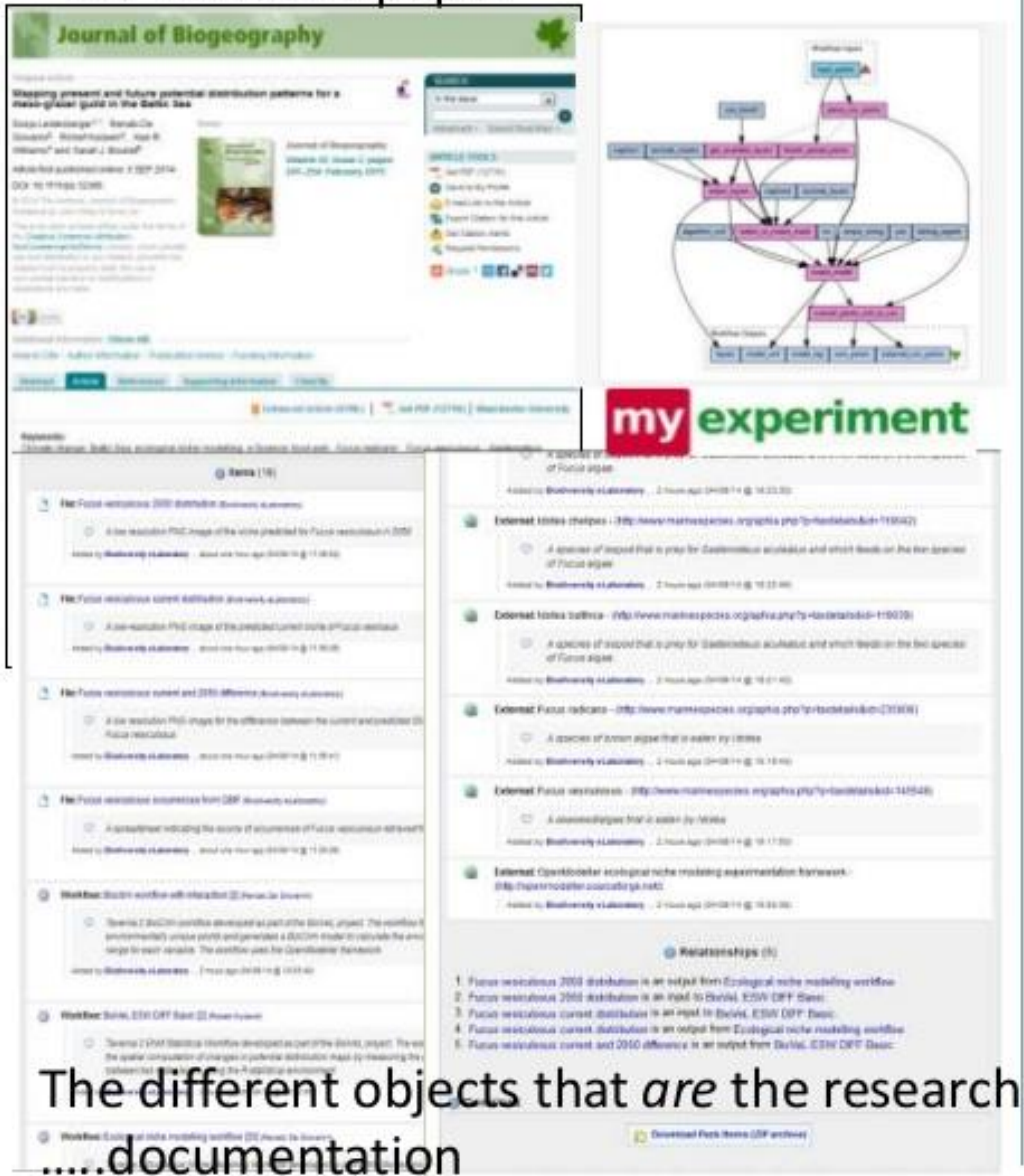


PIDs + Metadata

Sounds like Linked Data!

The FAIR Guiding Principles for Data Stewardship and Management
Scientific Data **3**, 160018 (2016)
doi:10.1038/sdata.2016.18

The narrative paper



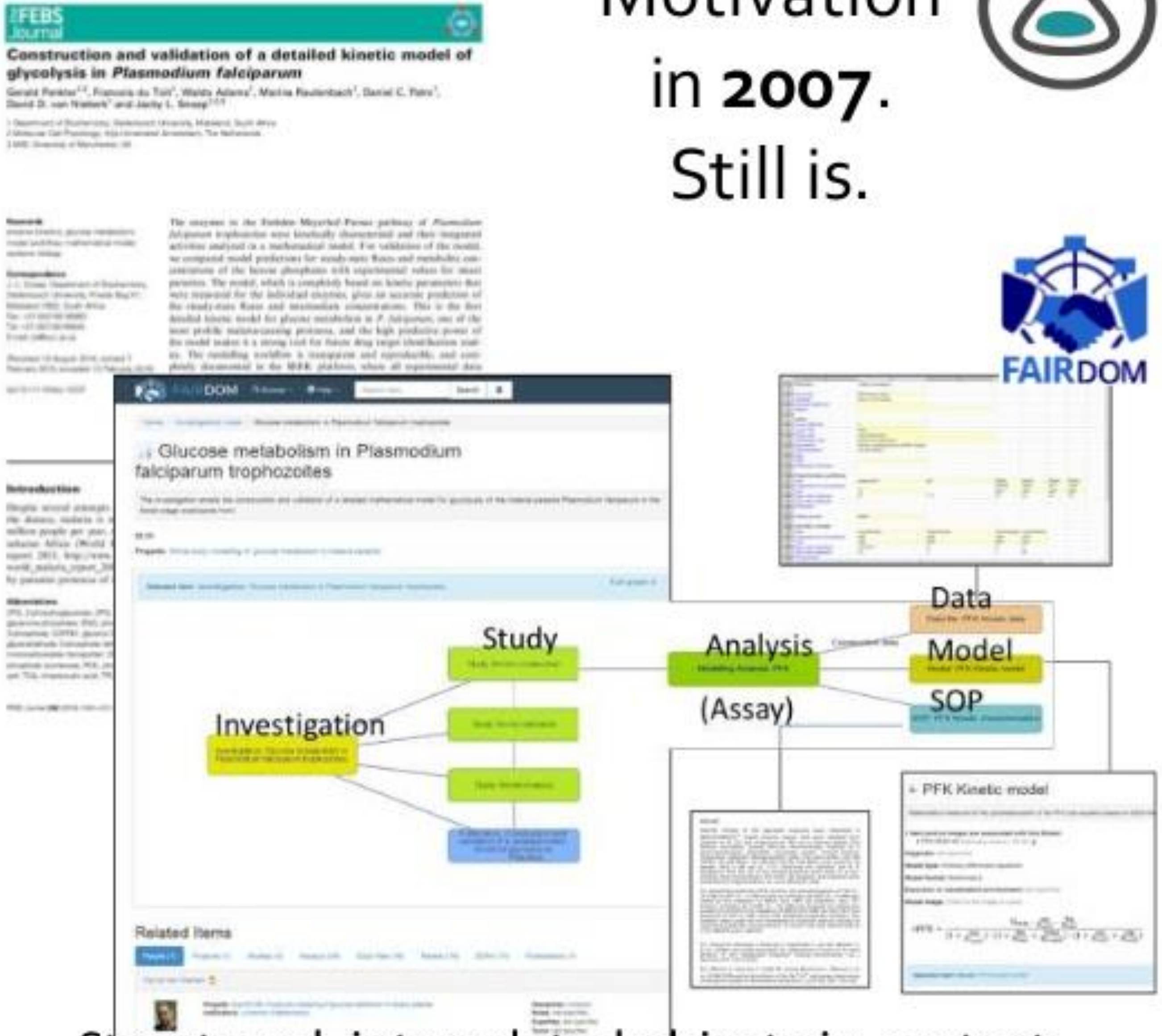
Journal of Biogeography

The different objects that *are* the research

.....documentation

The narrative paper

Motivation in **2007**. Still is.



FEBS Journal

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

Glucose metabolism in Plasmodium falciparum trophozoites

Related Items

PFK Kinetic model

Structured, interrelated objects in *context*

....documentation

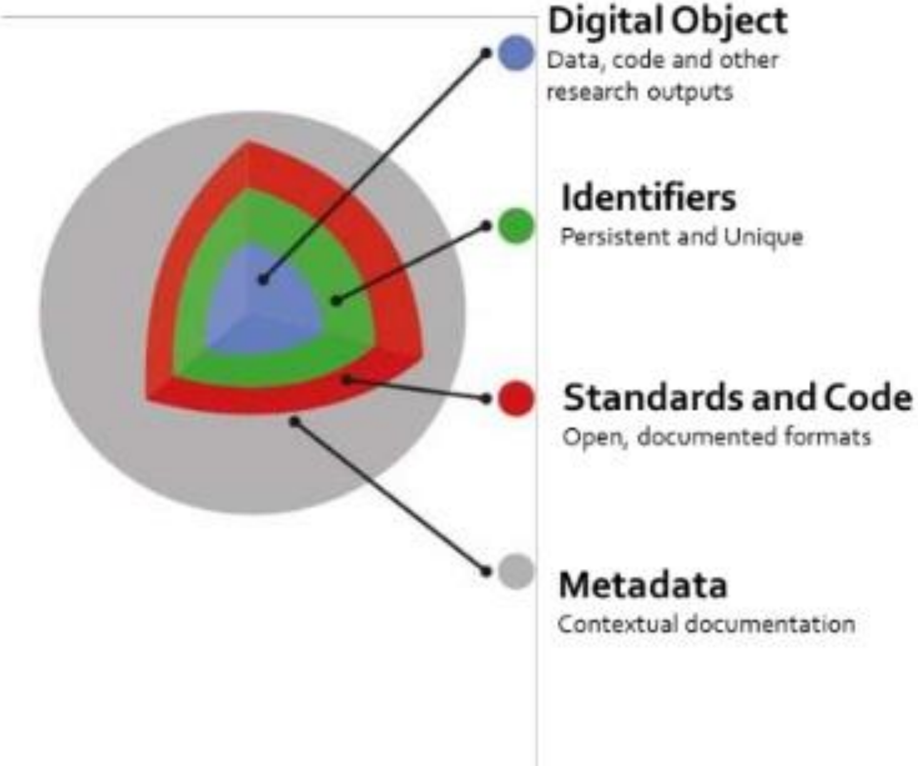
From Manuscripts to Research Objects

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the **complete software development environment**, [the complete data] and the complete set of instructions which generated the figures." David Donoho, "Wavelab and Reproducible Research," 1995

research outcomes more than just publications

data software, models, workflows, SOPs, lab protocols are ***first class citizens of scholarship***

added information required to make research ***FAIR*** *and* ***Reproducible*** (FAIR+R) …

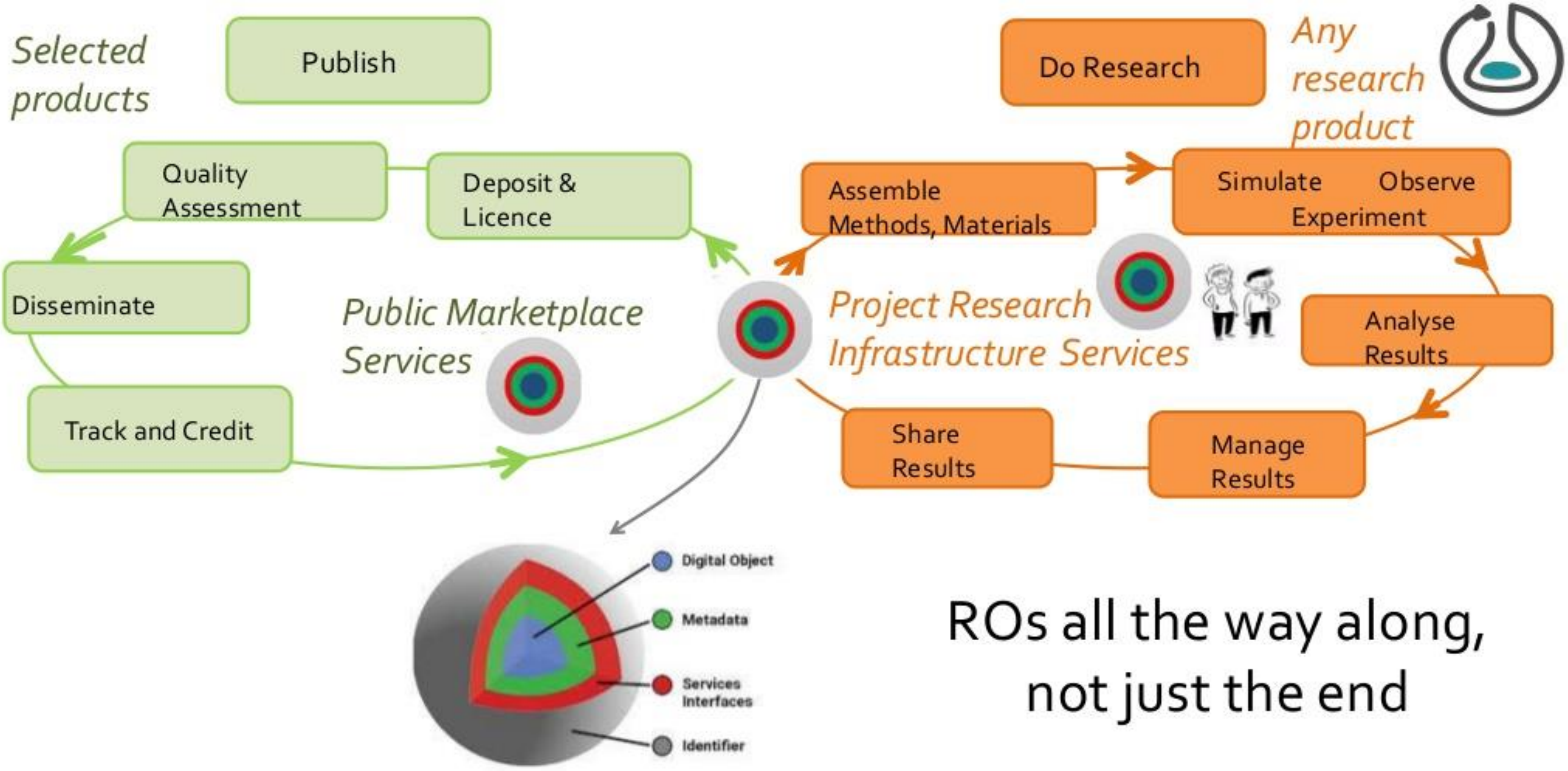


Digital Object
Data, code and other research outputs

Identifiers
Persistent and Unique

Standards and Code
Open, documented formats

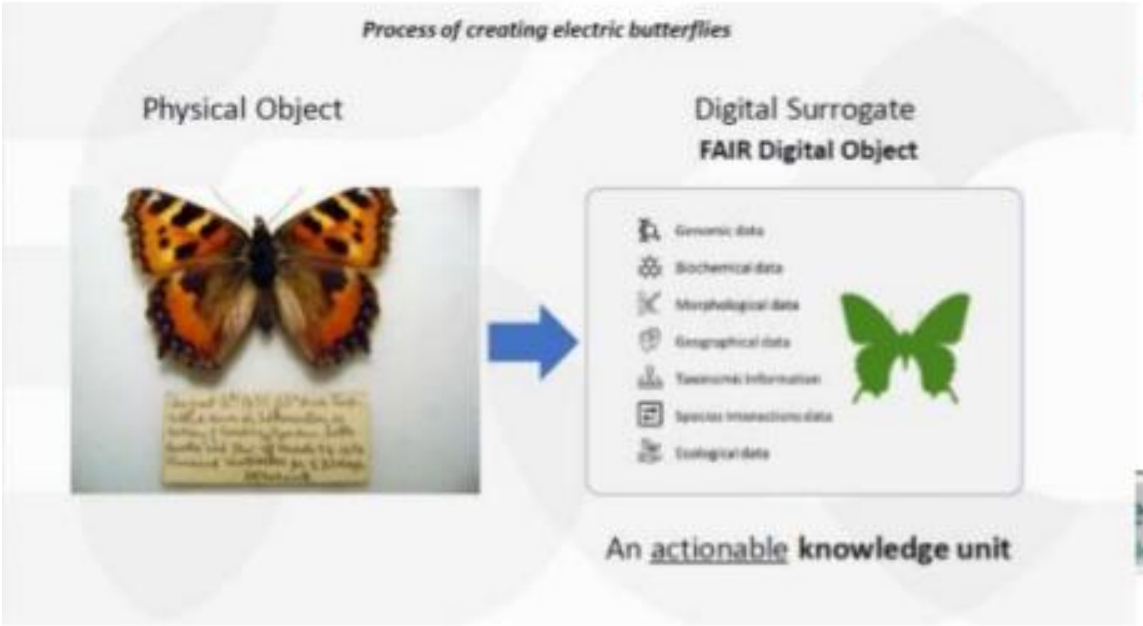
Metadata
Contextual documentation



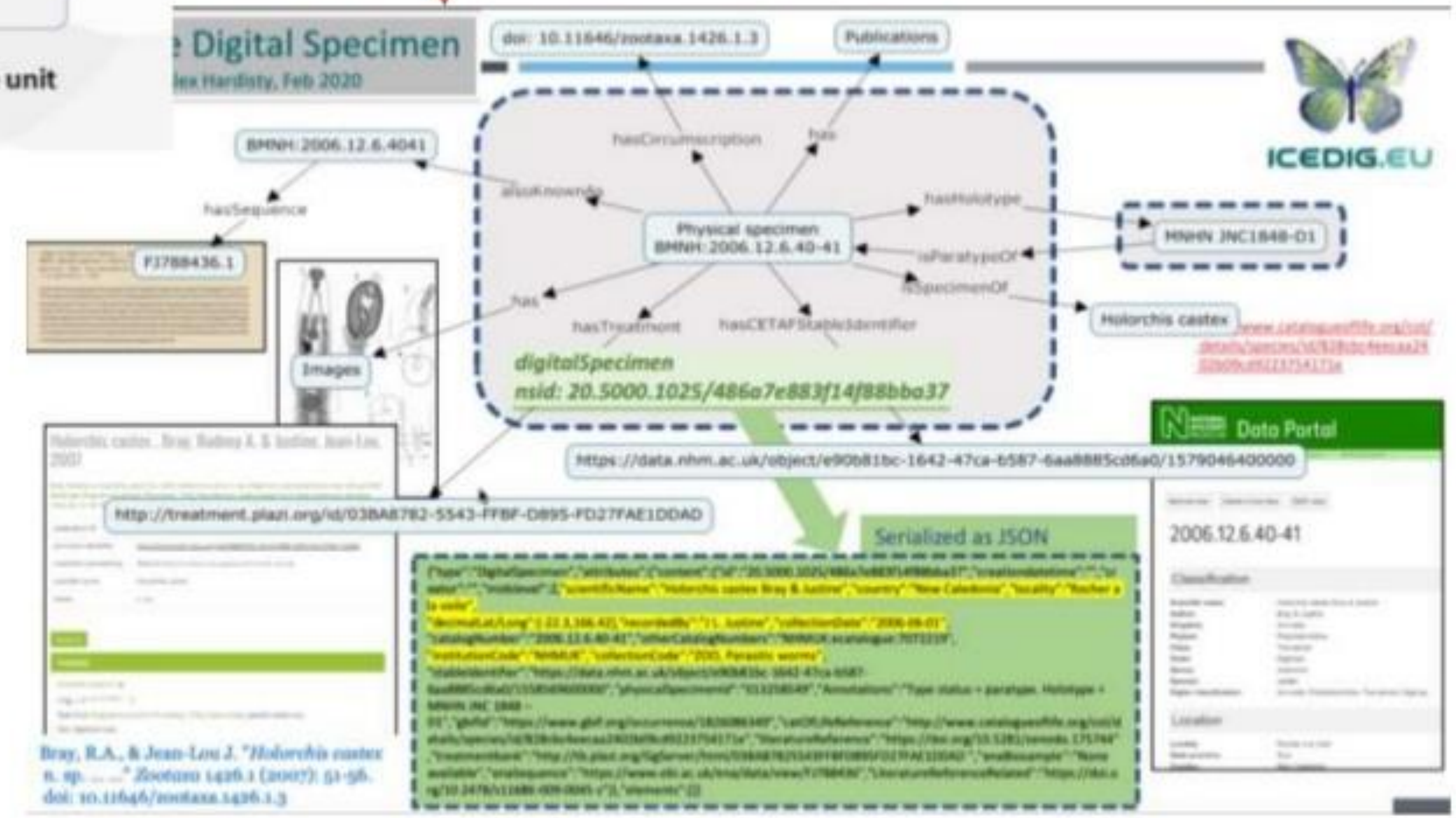
ROs all the way along, not just the end

Science 2.0 Repositories: Time for a Change in Scholarly Communication Assante, Candela, Castelli, Manghi, Pagano, D-Lib 2015
Ainsworth and Buchan: e-Labs and Work Objects: Towards Digital Health Economies, EuropeComm 2009, pp 205-216

Digital Objects electric butterflies



An actionable **knowledge unit**



{"type":"DigitalSpecimen","attributes":{"content":{"id":"20.5000.1025/486a7e883f14f88bba37","creationdatetime":"", [illegible]

Bray, R.A., & Jean-Lou J. "*Holorchis castex* n. sp. ...." Zootaxa 1426.1 (2007): 51-56. doi: 10.11646/zootaxa.1426.1.3

Digital twins

Actionable knowledge units

FAIR Digital Objects

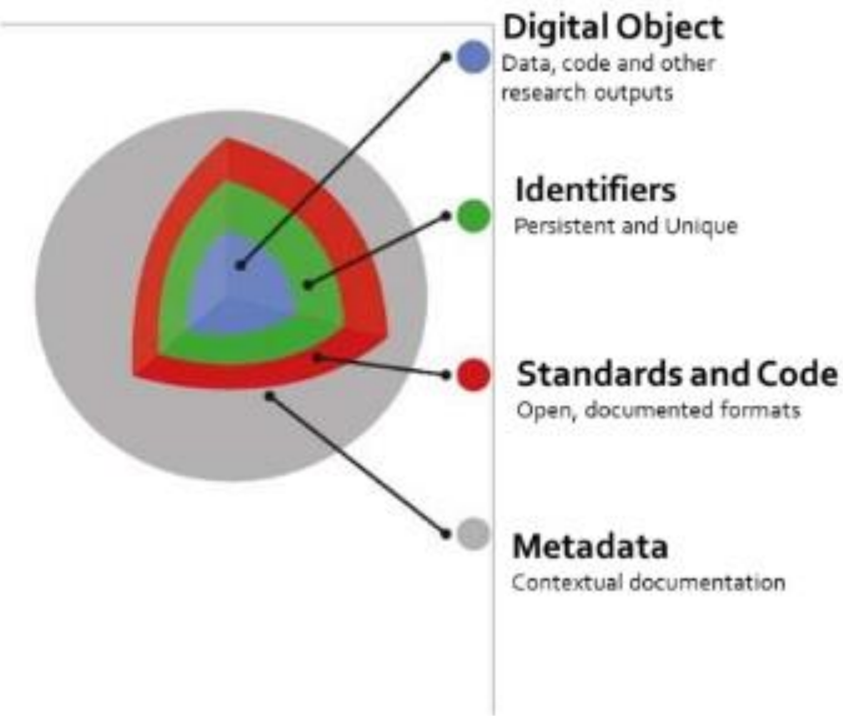
courtesy Dimitris Koureas
Coordinator DiSSCo EU Research Infrastructure

Specimen object image courtesy of Alex Hardisty

Digital Objects as First Class Entities

FAIR Digital Object Framework

A Knowledge Graph of FDOs



Digital Object
Data, code and other research outputs

Identifiers
Persistent and Unique

Standards and Code
Open, documented formats

Metadata
Contextual documentation

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

TURNING FAIR INTO REALITY

2018

GEDE RDA
Group of European Data Experts in RDA

EOSC FAIR
Executive Board Working Group

EOSC Architecture
Executive Board Working Group

DiSSCo

DONA

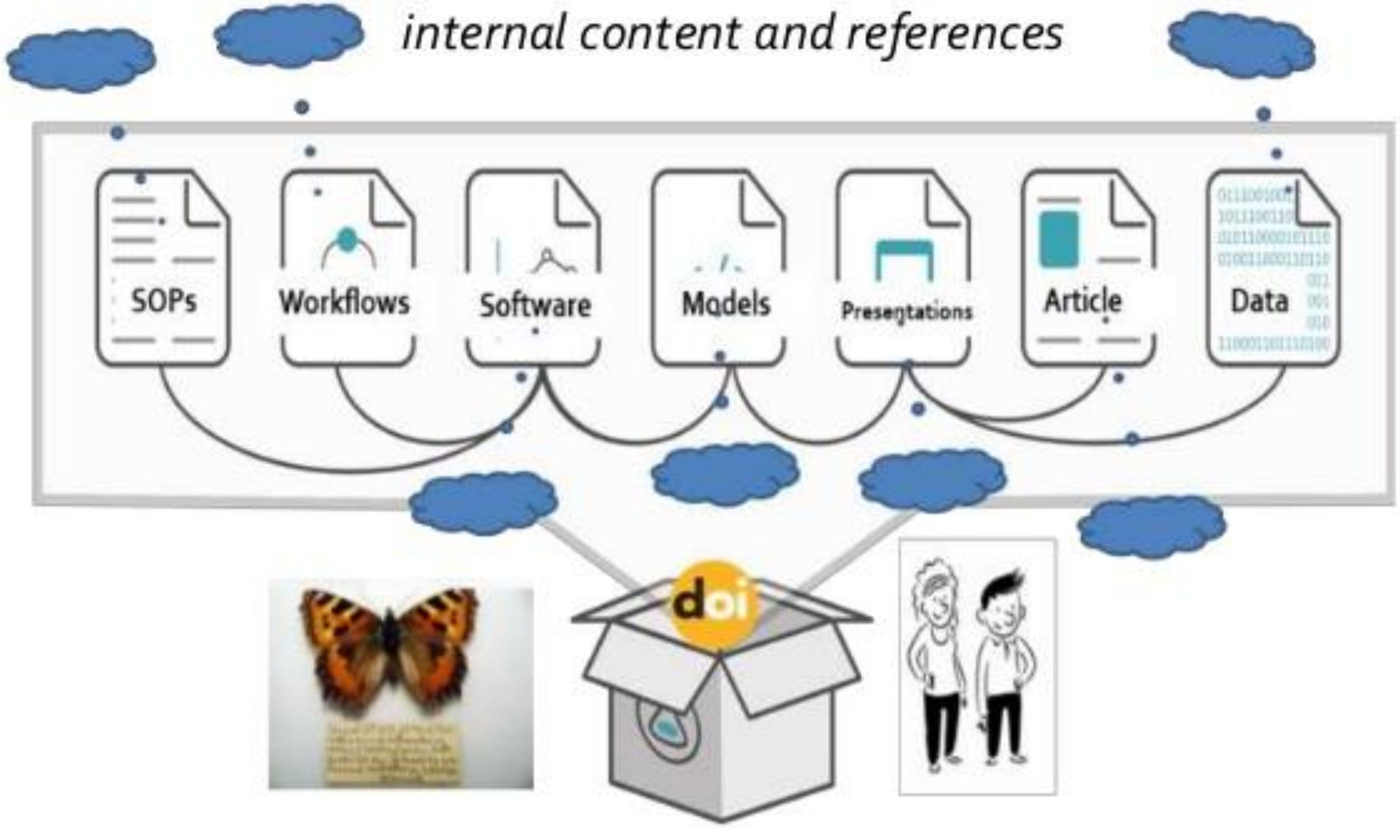
- Schwardmann (2020), Digital Objects – FAIR Digital Objects: Which Services Are Required? Data Science Journal
- EOSC Interoperability Framework Draft (2020)
- Hardisty A, et al (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure RIO 6: e54280.
- DONA Digital Object Architecture Digital Object Interface Protocol (2018)
- https://fairdigitalobjectframework.org/

From Manuscripts to FAIR+R Research Objects

research objects **related** and **bundled** together …
one shareable, cite-able, exchangable resource that can be versioned and snapshot …

metadata describing **context** and content of objects
dependencies, versions, relationships, provenance …

enough to be **reproducible**

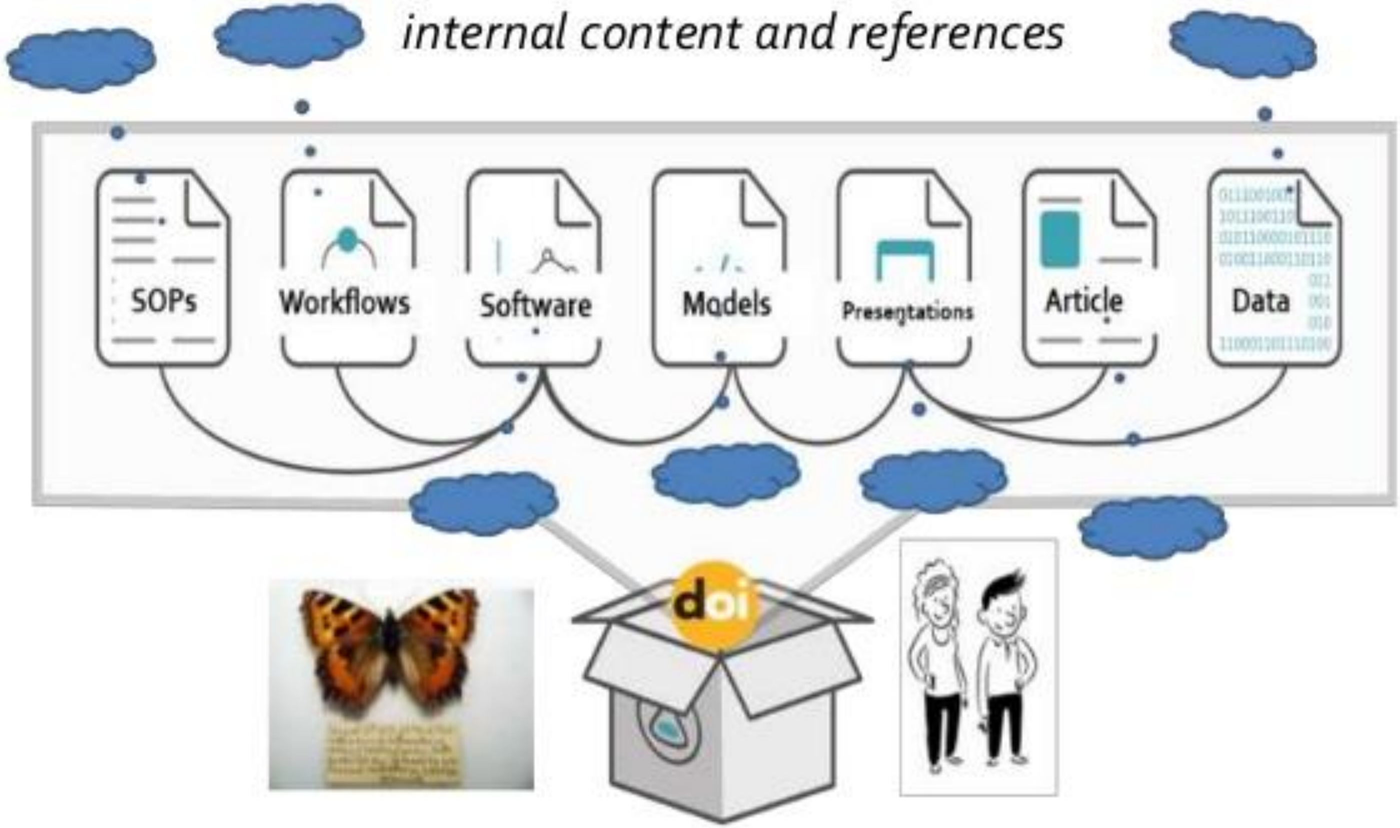


internal content and references

virtual objects , links to physical objects (people, specimens, equipment)
integrated view over fragmented & scattered specialised repositories

From Manuscripts to FAIR+R Research Objects

Bigger on the inside than the outside

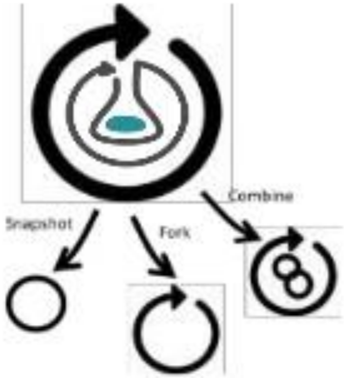


Packaging

RDM role

Commons
Currency
Credit

Archival preservation
Reproducibility
Portability
Virtual Witnessing



Releasing
Living Objects

Findable

Accessible

Interoperable

Reusable

FAIR ROs

Analogous to software
FAIR Enough

Structure: *Composite*
Dynamic: *Versioning*
Executable: *Portability*
Virtual: *References*
Maintenance: *Decay*

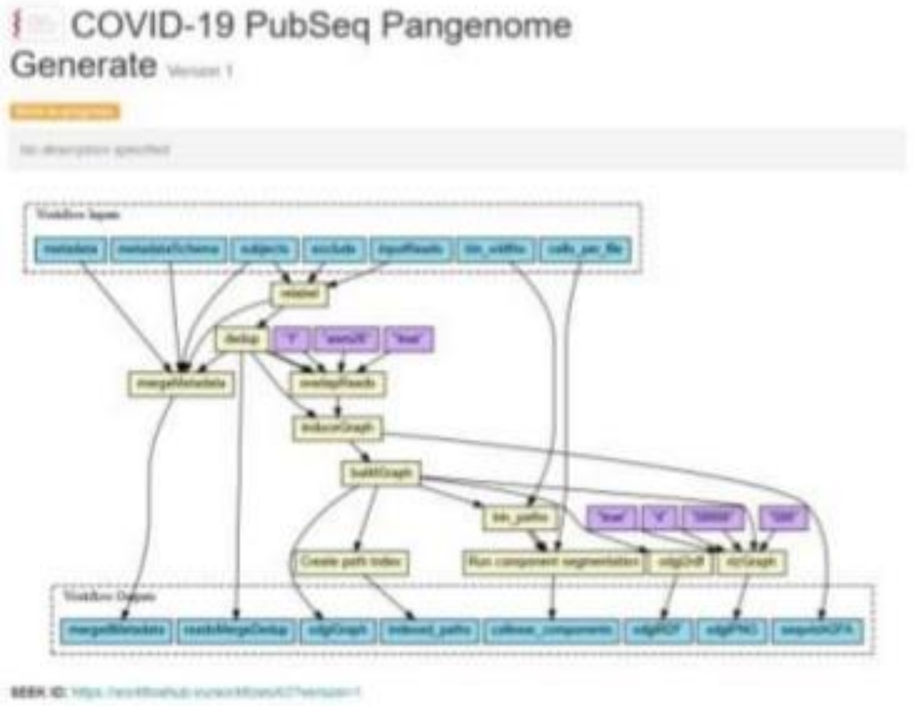
PID resolution?
Metadata?
Access?
Licences?

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

2010

Packaging of Digital Objects

Driver: Computational Workflows



COVID-19 PubSeq Pangenome Generate Version 1

Workflow 4Ever

BioVeL

http://wf4ever.org/

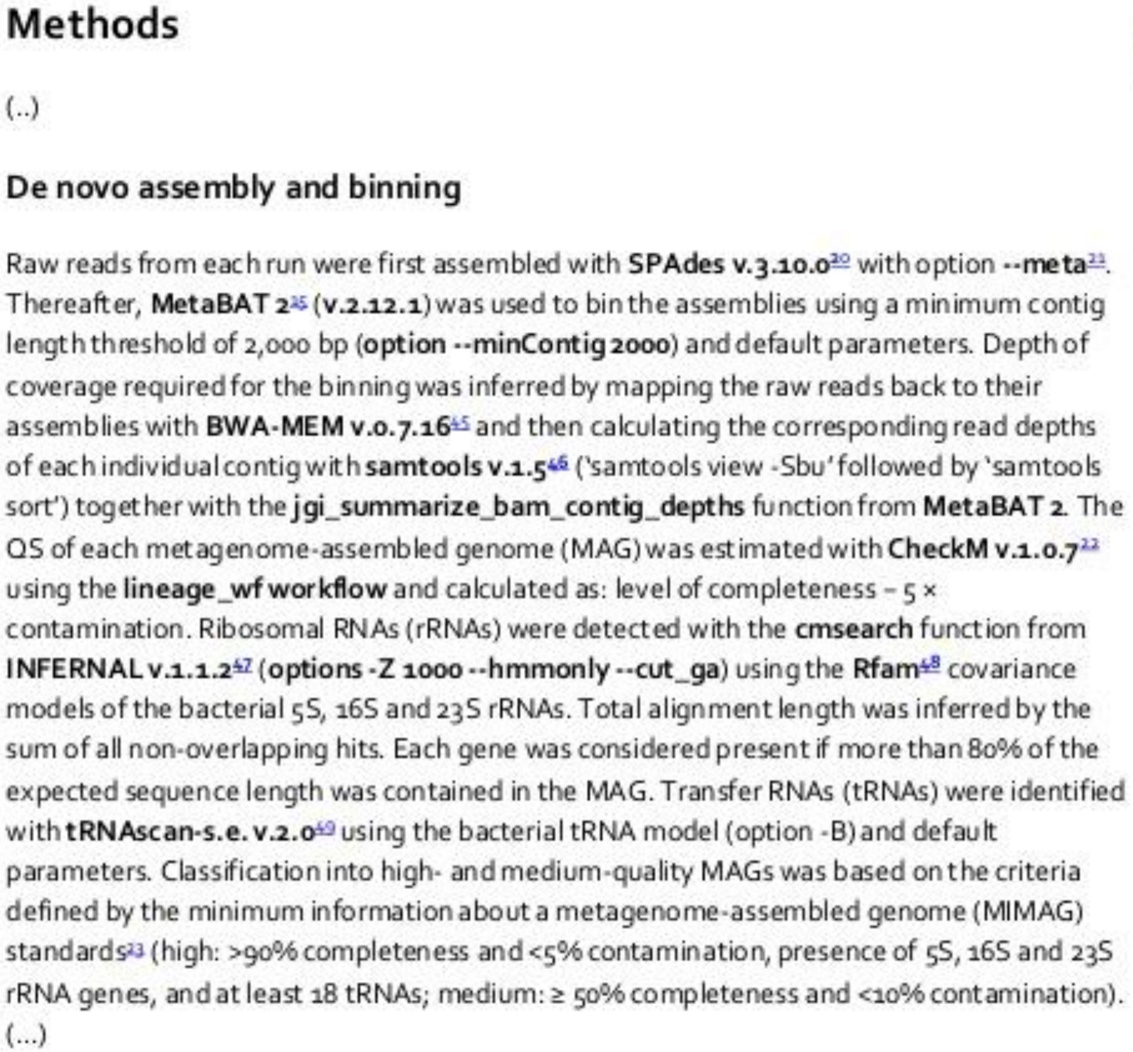
Preservation of computational workflows in data-intensive science

- **Workflow-centric Research Objects**
- Computational workflows, provenance of executions, interconnections between workflows and related resources (e.g., datasets, publications, etc.), social aspects in the experiments.

- Wf-centric RO creation & management **best practices**

- analysis and management of **decay** in workflows.

Data pipeline & analysis reporting & reproducibility



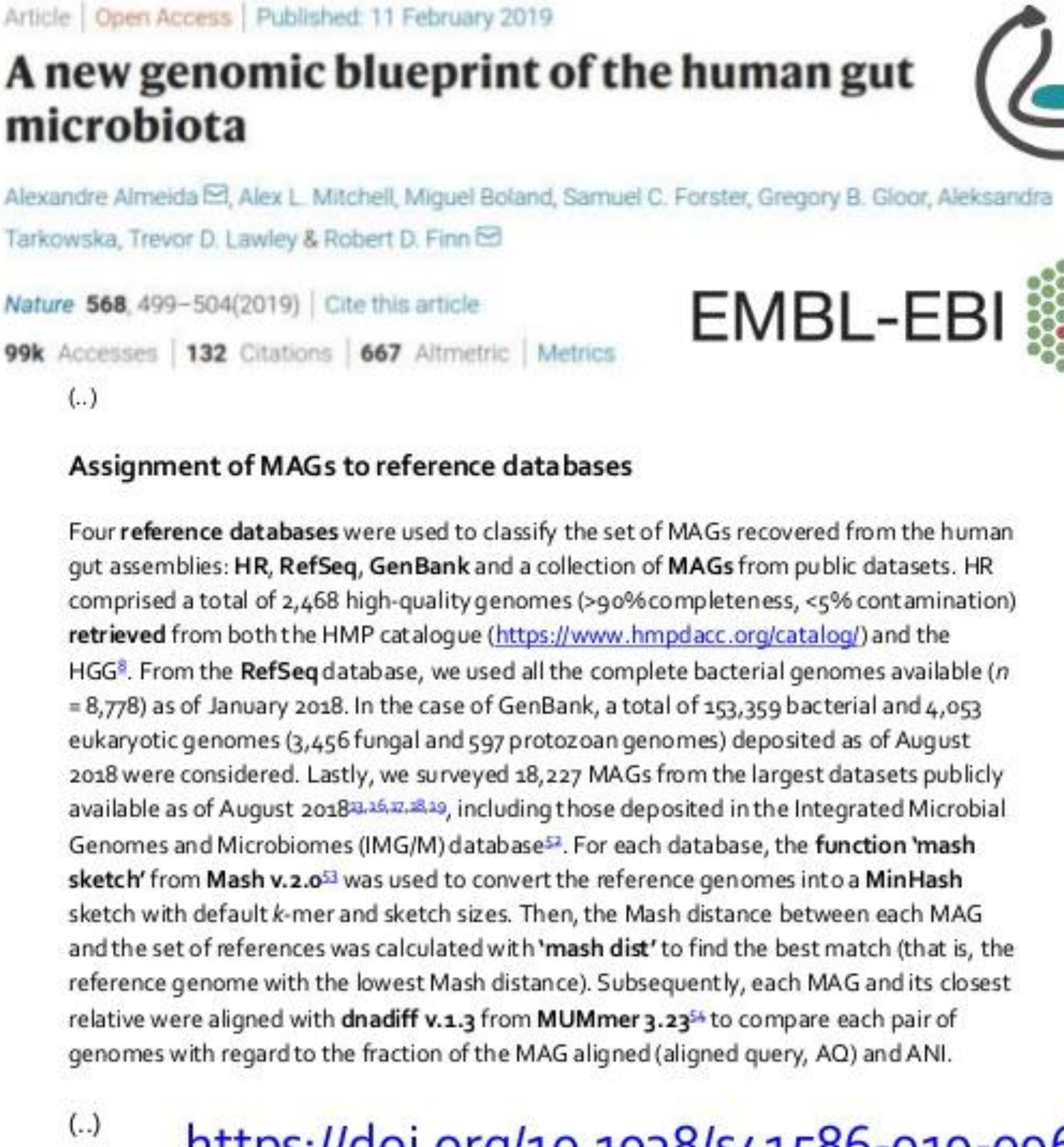
Methods

(..)

De novo assembly and binning

Raw reads from each run were first assembled with **SPAdes v.3.10.0**[20] with option **--meta**[21]. Thereafter, **MetaBAT 2**[25] (**v.2.12.1**) was used to bin the assemblies using a minimum contig length threshold of 2,000 bp (**option --minContig 2000**) and default parameters. Depth of coverage required for the binning was inferred by mapping the raw reads back to their assemblies with **BWA-MEM v.0.7.16**[45] and then calculating the corresponding read depths of each individual contig with **samtools v.1.5**[46] ('samtools view -Sbu' followed by 'samtools sort') together with the **jgi_summarize_bam_contig_depths** function from **MetaBAT 2**. The QS of each metagenome-assembled genome (MAG) was estimated with **CheckM v.1.0.7**[22] using the **lineage_wf workflow** and calculated as: level of completeness − 5 × contamination. Ribosomal RNAs (rRNAs) were detected with the **cmsearch** function from **INFERNAL v.1.1.2**[47] (**options -Z 1000 --hmmonly --cut_ga**) using the **Rfam**[48] covariance models of the bacterial 5S, 16S and 23S rRNAs. Total alignment length was inferred by the sum of all non-overlapping hits. Each gene was considered present if more than 80% of the expected sequence length was contained in the MAG. Transfer RNAs (tRNAs) were identified with **tRNAscan-s.e. v.2.0**[49] using the bacterial tRNA model (option -B) and default parameters. Classification into high- and medium-quality MAGs was based on the criteria defined by the minimum information about a metagenome-assembled genome (MIMAG) standards[23] (high: >90% completeness and <5% contamination, presence of 5S, 16S and 23S rRNA genes, and at least 18 tRNAs; medium: ≥ 50% completeness and <10% contamination).

(...)



Article | Open Access | Published: 11 February 2019

A new genomic blueprint of the human gut microbiota

Alexandre Almeida ✉, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn ✉

Nature **568**, 499–504(2019) | Cite this article

99k Accesses | **132** Citations | **667** Altmetric | Metrics

(..)

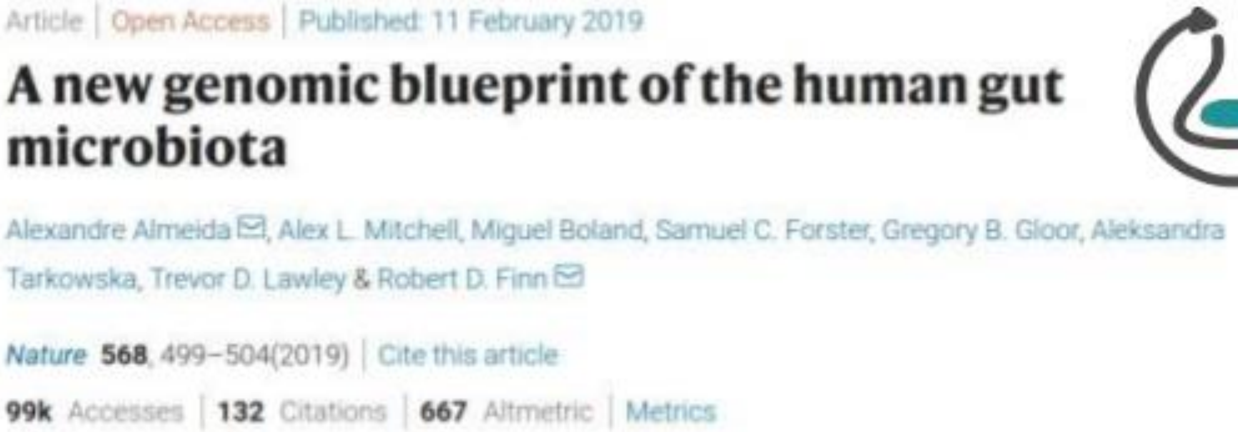
Assignment of MAGs to reference databases

Four **reference databases** were used to classify the set of MAGs recovered from the human gut assemblies: **HR**, **RefSeq**, **GenBank** and a collection of **MAGs** from public datasets. HR comprised a total of 2,468 high-quality genomes (>90% completeness, <5% contamination) **retrieved** from both the HMP catalogue (https://www.hmpdacc.org/catalog/) and the HGG[8]. From the **RefSeq** database, we used all the complete bacterial genomes available (*n* = 8,778) as of January 2018. In the case of GenBank, a total of 153,359 bacterial and 4,053 eukaryotic genomes (3,456 fungal and 597 protozoan genomes) deposited as of August 2018 were considered. Lastly, we surveyed 18,227 MAGs from the largest datasets publicly available as of August 2018[23,26,27,28,29], including those deposited in the Integrated Microbial Genomes and Microbiomes (IMG/M) database[52]. For each database, the **function 'mash sketch'** from **Mash v.2.0**[53] was used to convert the reference genomes into a **MinHash** sketch with default *k*-mer and sketch sizes. Then, the Mash distance between each MAG and the set of references was calculated with **'mash dist'** to find the best match (that is, the reference genome with the lowest Mash distance). Subsequently, each MAG and its closest relative were aligned with **dnadiff v.1.3** from **MUMmer 3.23**[54] to compare each pair of genomes with regard to the fraction of the MAG aligned (aligned query, AQ) and ANI.

(..)

https://doi.org/10.1038/s41586-019-0965-1

Data pipeline & analysis reporting & reproducibility



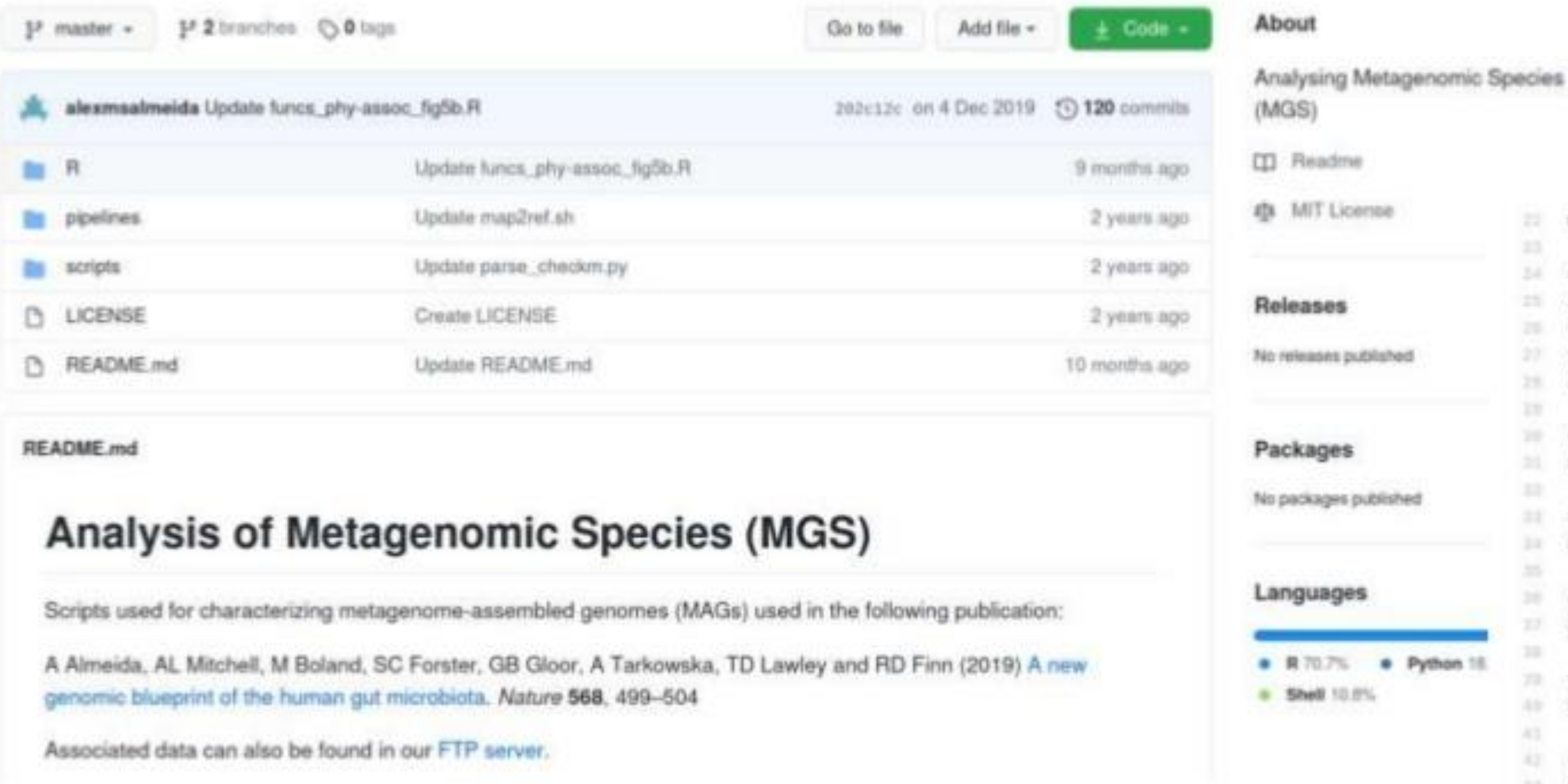
Article | Open Access | Published: 11 February 2019

A new genomic blueprint of the human gut microbiota

Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn

Nature **568**, 499–504(2019) | Cite this article

99k Accesses | **132** Citations | **667** Altmetric | Metrics



alexmsalmeida Update funcs_phy-assoc_fig5b.R 202c12c on 4 Dec 2019 120 commits

R	Update funcs_phy-assoc_fig5b.R	9 months ago
pipelines	Update map2ref.sh	2 years ago
scripts	Update parse_checkm.py	2 years ago
LICENSE	Create LICENSE	2 years ago
README.md	Update README.md	10 months ago

README.md

Analysis of Metagenomic Species (MGS)

Scripts used for characterizing metagenome-assembled genomes (MAGs) used in the following publication:

A Almeida, AL Mitchell, M Boland, SC Forster, GB Gloor, A Tarkowska, TD Lawley and RD Finn (2019) A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504

Associated data can also be found in our FTP server.

About

Analysing Metagenomic Species (MGS)

Readme

MIT License

Releases

No releases published

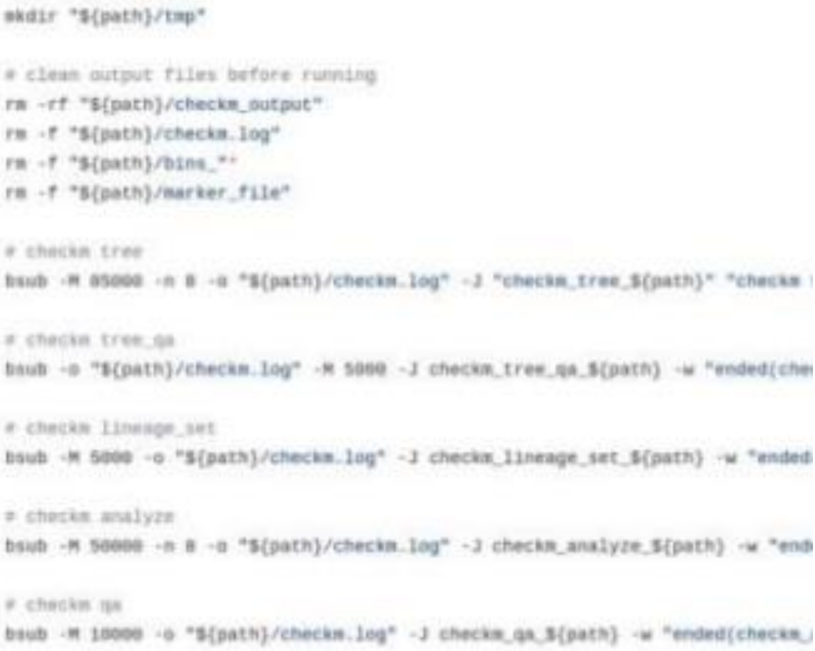
Packages

No packages published

Languages

● R 70.7% ● Python 18 ● Shell 10.8%

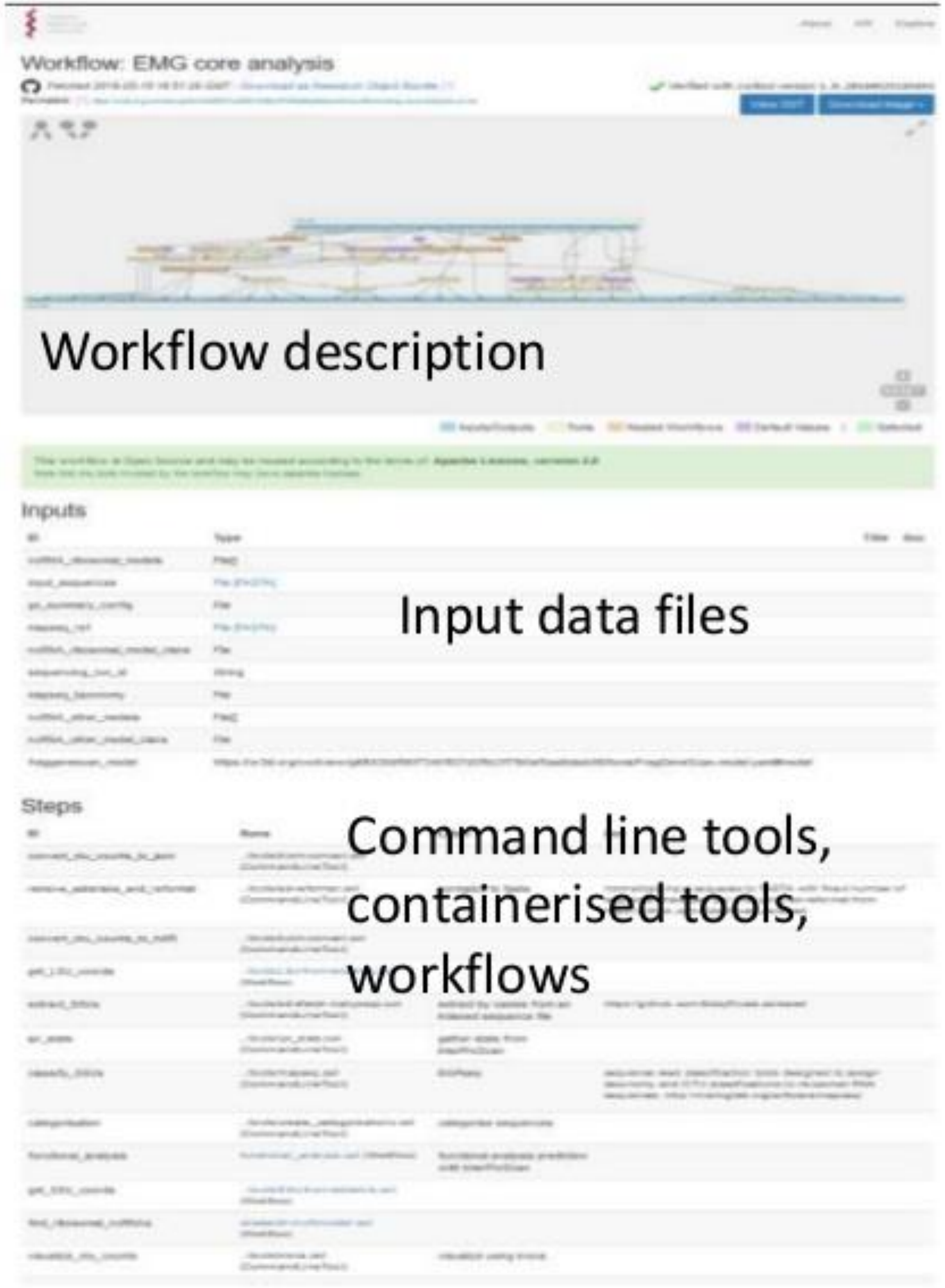
Lots of scripts & datasets



```
mkdir "${path}/tmp"

# clean output files before running
rm -rf "${path}/checkm_output"
rm -f "${path}/checkm.log"
rm -f "${path}/bins_"*
rm -f "${path}/marker_file"

# checkm tree
bsub -M 85000 -n 8 -o "${path}/checkm.log" -J "checkm_tree_${path}" "checkm

# checkm tree_qa
bsub -o "${path}/checkm.log" -M 5000 -J checkm_tree_qa_${path} -w "ended(che

# checkm lineage_set
bsub -M 5000 -o "${path}/checkm.log" -J checkm_lineage_set_${path} -w "ended

# checkm analyze
bsub -M 50000 -n 8 -o "${path}/checkm.log" -J checkm_analyze_${path} -w "end

# checkm qa
bsub -M 10000 -o "${path}/checkm.log" -J checkm_qa_${path} -w "ended(checkm_
```

https://doi.org/10.1038/s41586-019-0965-1

EBI's MGnify metagenomics workflows as workflows using a WfMS



Workflow: EMG core analysis

Workflow description

Inputs

Input data files

Steps

Command line tools, containerised tools, workflows

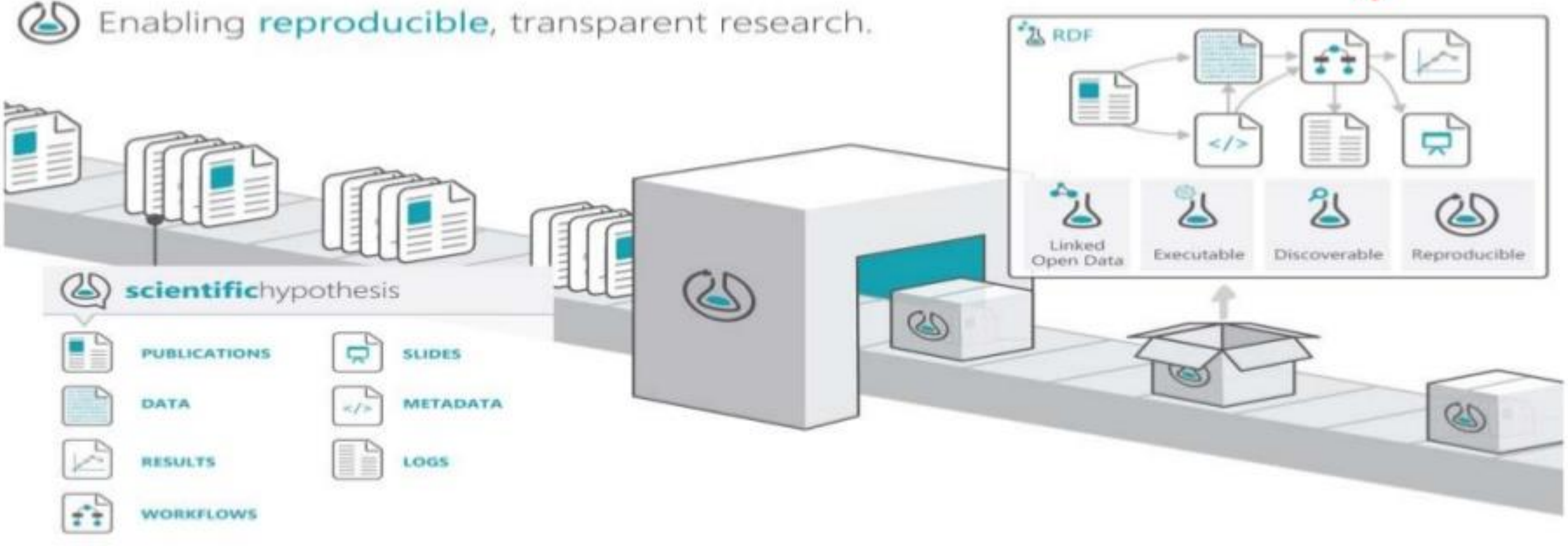
Outputs

Output data files

COMMON WORKFLOW LANGUAGE

researchobject.org

Linked Data!!



Standards-based metadata framework for bundling resources (physically and logically) with context into citable reproducible packages.

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/

A Research Object **bundles** and **relates** digital resources of a scientific experiment/investigation + context

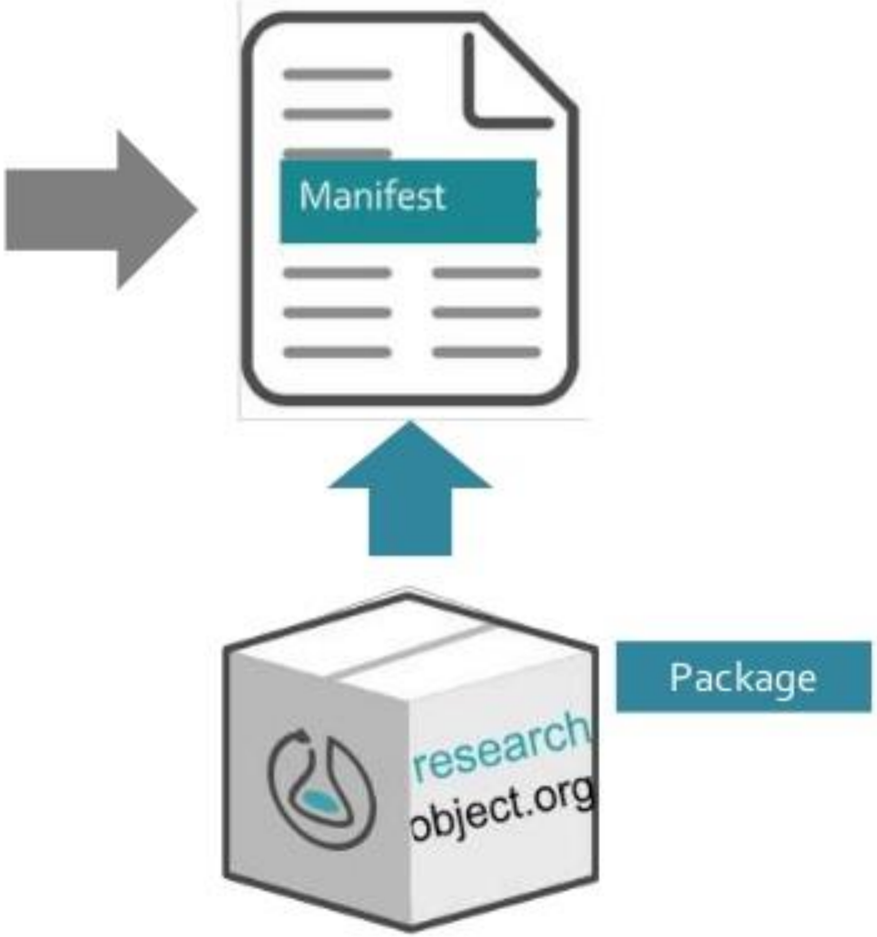
Data used and results produced in experimental study

Methods employed to produce and analyse that data

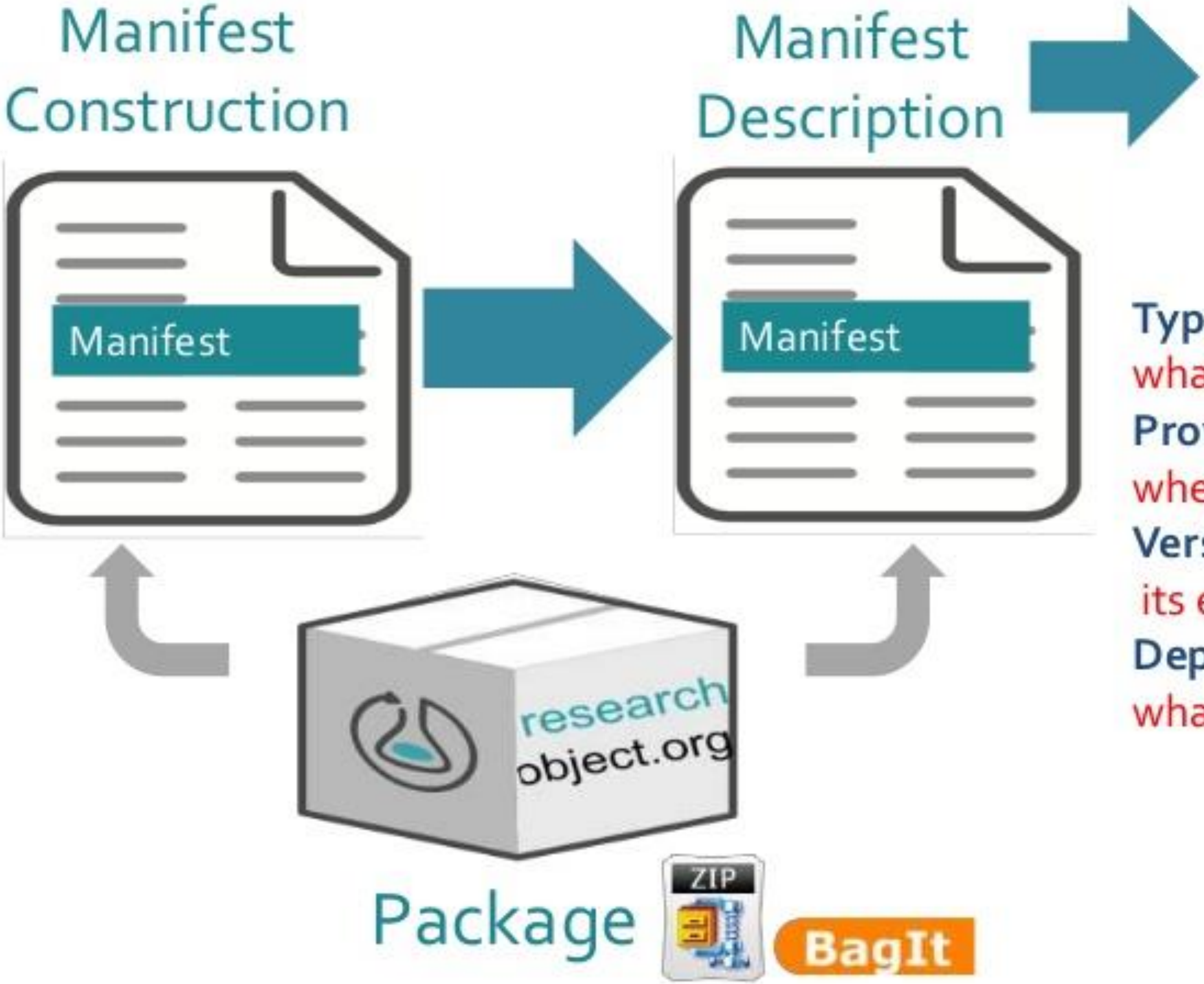
Provenance and settings for the experiments

People, specimens, equipment etc involved in the investigation

Annotations about these resources, to improve understanding and interpretation



Research Objects => Metadata Objects

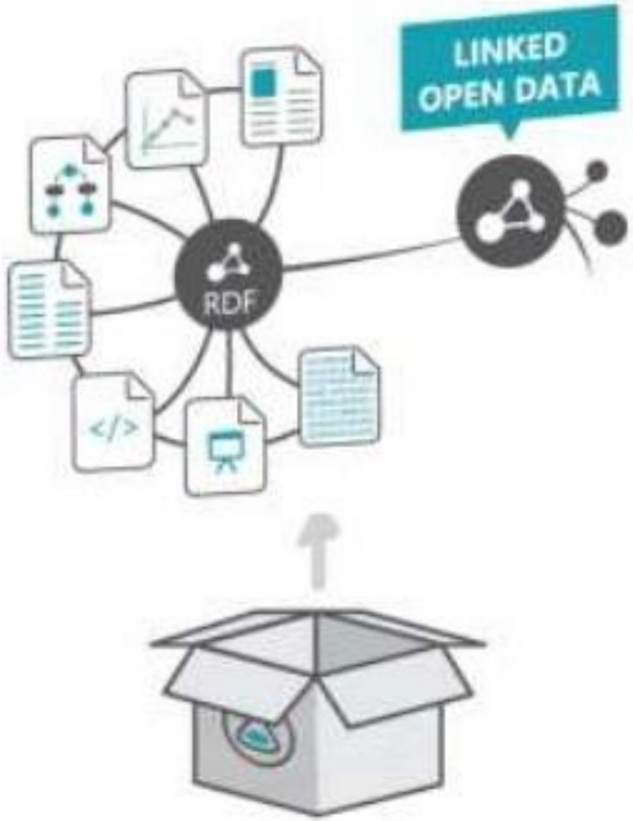


Base Profile, Profiles for RO types

Identification
to locate things
Aggregates
to link things together
Annotations
about things & their relationships

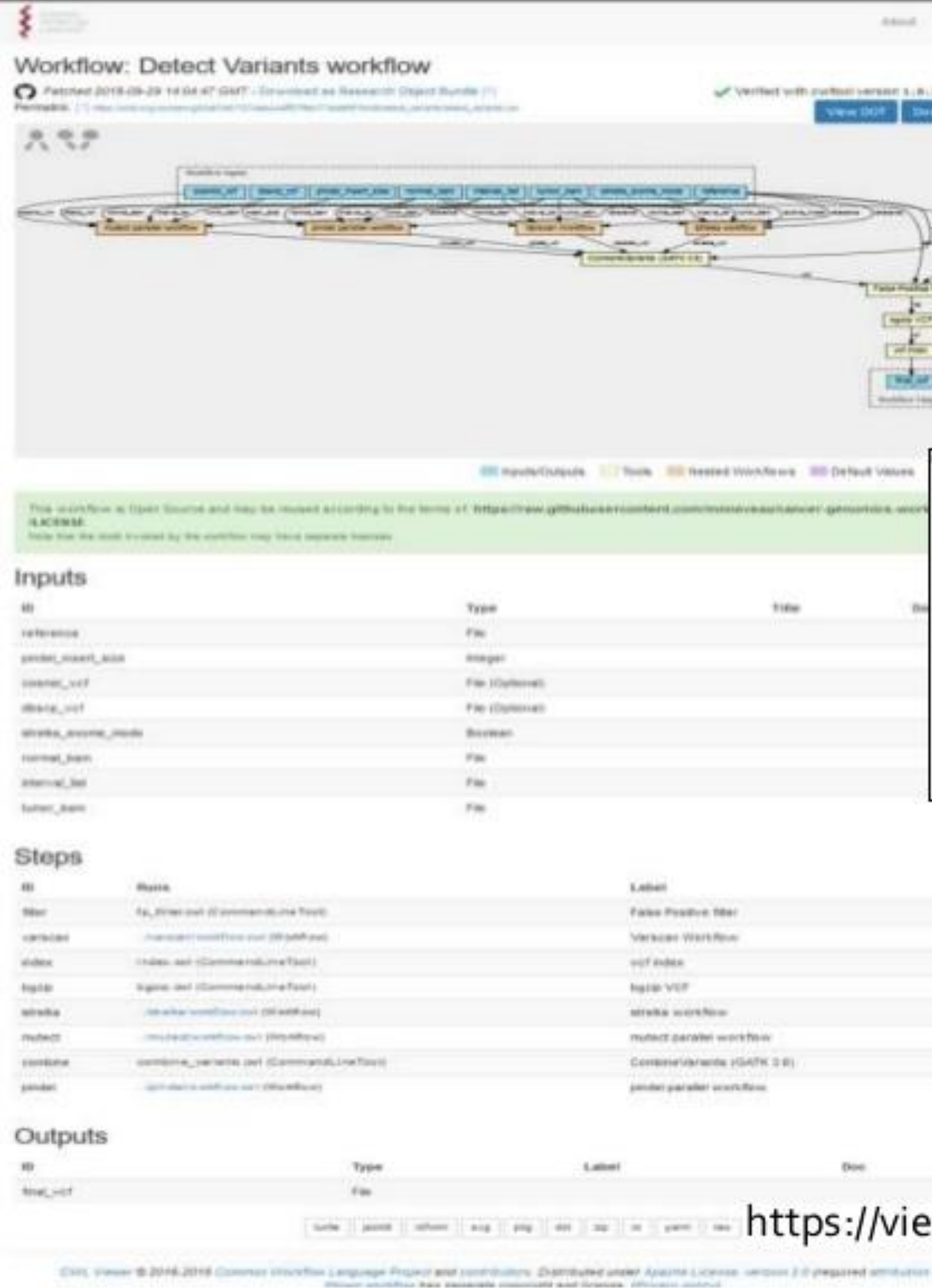
Type Checklists
what should be there
Provenance
where it came from
Versioning
its evolution
Dependencies
what else is needed

We are in a Semantic Web Conference….
Linked Data Middleware



- **Manifests described using Linked Data**
  - Identifiers to resources, including people (orcid)
  - OWL / RDF / SPARQL / JSON-LD
- **Mismash of specialized ontologies**
  - Construct the manifest itself
    - *W3C Web Annotation Vocabulary*
    - *OAI Object Exchange and Reuse*
  - Describe manifest content
    - *Wf4Ever RO ontology, Wf4Ever ROEvo …*
    - *Dublin Core, FOAF, SIOC, Creative Commons, PROV, PAV…*
- **RDF shapes (SHACL, ShEx)**
  - Capture requirements, expectations and validate profiles
  - Hard to express checklists

Evidence!



Workflow: Detect Variants workflow

View DOT | Download Image

Inputs

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

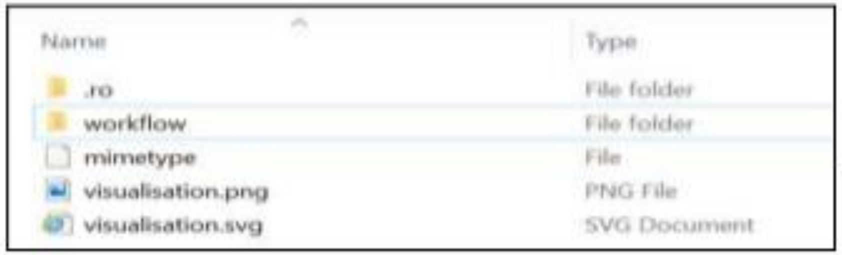
Steps

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

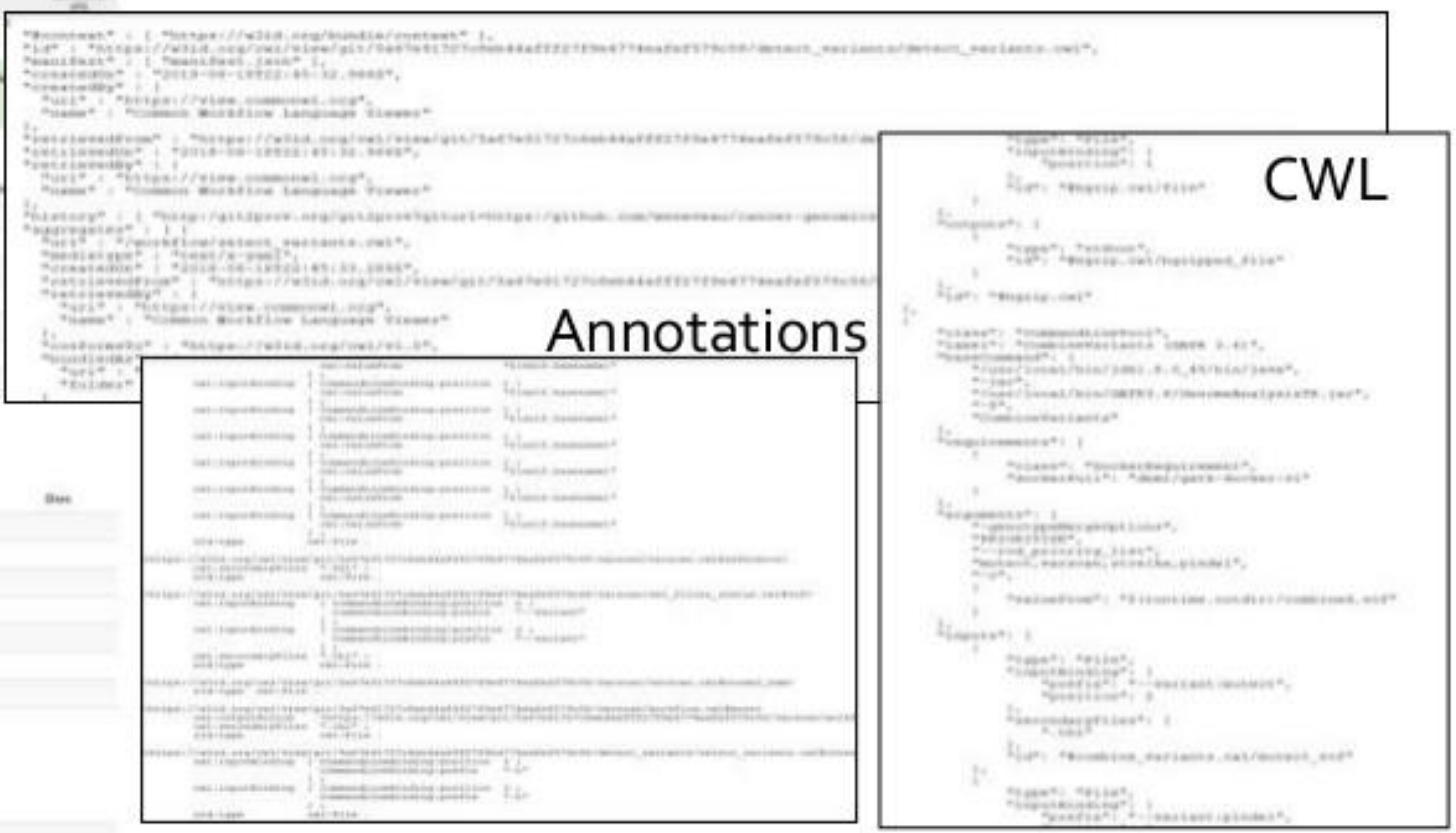
ID	Type	Label	Doc
final_vcf	File		

Manifest



Name	Type
.ro	File folder
workflow	File folder
mimetype	File
visualisation.png	PNG File
visualisation.svg	SVG Document

Name
annotations
manifest.json



Annotations

CWL

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

Influence? Publishers...

Int J Digit Libr (2017) 18:99–111
DOI 10.1007/s00799-016-0174-6

Experiences in integrated data and research object publishing using GigaDB

Scott C Edmunds[1] · Peter Li[1] · Christopher I Hunter[1] · Si Zhe Xiao[1] · Robert L Davidson[1,2] · Nicole Nogoy[1] · Laurie Goodman[1]

Received: 29 June 2015 / Revised: 30 April 2016 / Accepted: 10 May 2016 / Published online: 27 May 20

Abstract In the era of computation and data-driven research, traditional methods of disseminating research are no longer fit-for-purpose. New approaches for disseminating data, methods and results are required to maximize knowledge discovery. The "long tail" of small, unstructured datasets is well catered for by a number of general-purpose repositories, but there has been less support for "big data". Outlined here are our experiences in attempting to tackle the gaps in publishing large-scale, computationally intensive research. *GigaScience* is an open-access, open-data journal aiming to revolutionize large-scale biological data dissemination, organization and re-use. Through use of the data handling infrastructure of the genomics centre BGI, *GigaScience* links standard manuscript publication with an integrated database (GigaDB) that hosts all associated data, and provides additional data analysis tools and computing resources. Furthermore, the supporting workflows and methods are also integrated to make published articles more transparent and open. GigaDB has released many new and previously unpublished datasets and data types, including as urgently needed data to tackle infectious disease outbreaks, cancer and the growing food crisis. Other "executable" research objects, such as workflows, virtual machines and software from several *GigaScience* articles have been archived and shared in reproducible, transparent and usable formats. With data citation producing evidence of, and credit for, its us
GigaScience demonstra
publications. Here data
upon by users without c
tational infrastructure in

Keywords Reproducib
Computational biology

1 Introduction

In a world where zetta
now produced globally
to this information is t
realizing its potential f
For scientific data in p
ing access to enable n
transparency and self-c
tive and rapid progress,
questions—revealing p
tions across datasets.

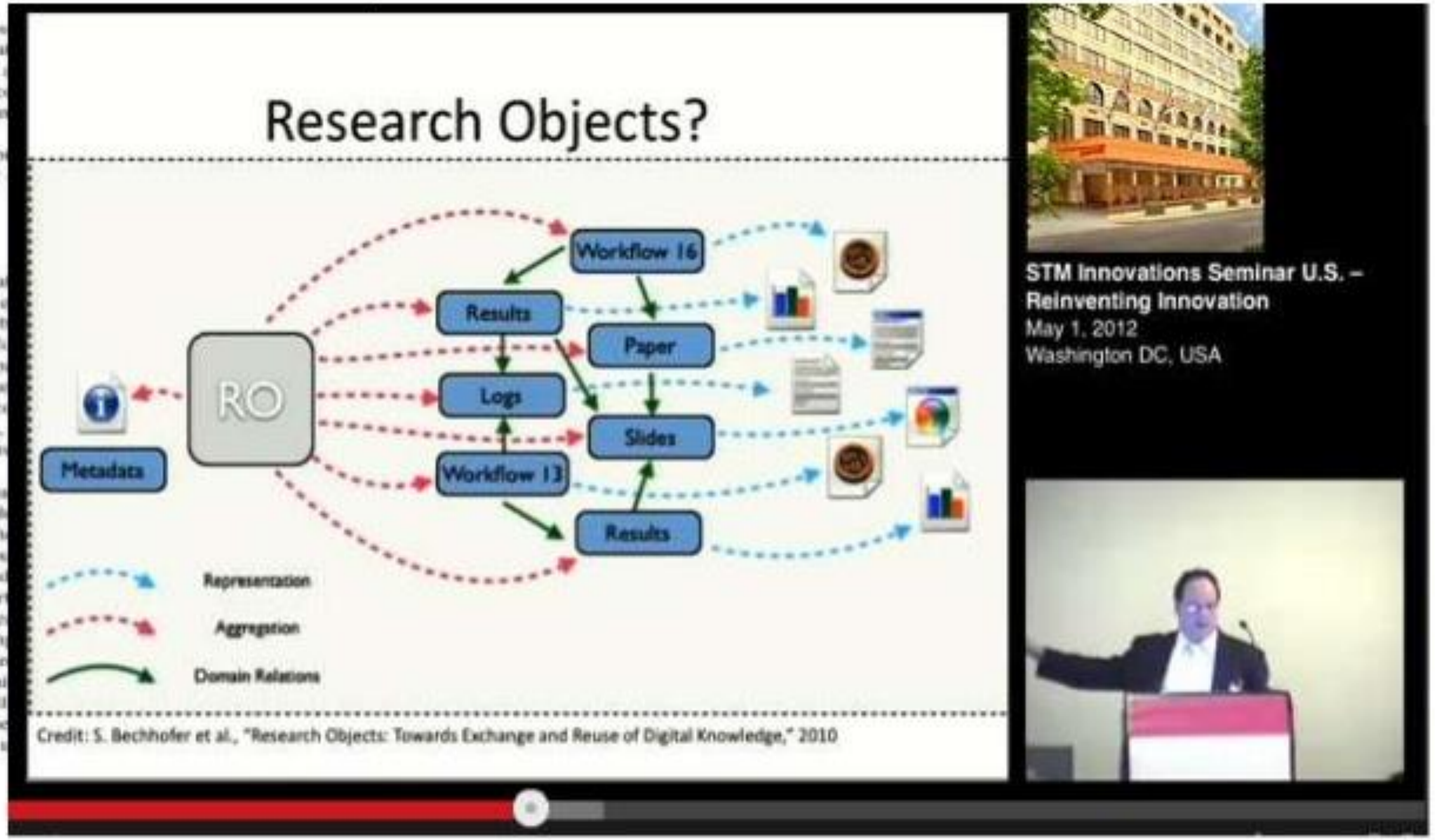
On top of a citation
has had other measurabl
rice research [90]. Furth
of a growing global popu
loss of biodiversity, and
and rapid action. Unfor
in much of the world me
data that is already bein
much as possible. Ther
of open access publishi
half the papers current
read [83]. Browsing the
key to maximizing the

✉ Scott C Edmunds
scott@gigasciencejournal.com

[1] GigaScience, BGI-Hong Kong Co, Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong SAR, China

[2] Office for National Statistics, Duffryn, Government Buildings, Cardiff Rd, Newport NP10 8XG, UK

Howard Ratner,
Chair STM Future Labs Committee, CEO EVP Nature Publishing Group
Director of Development for CHORUS

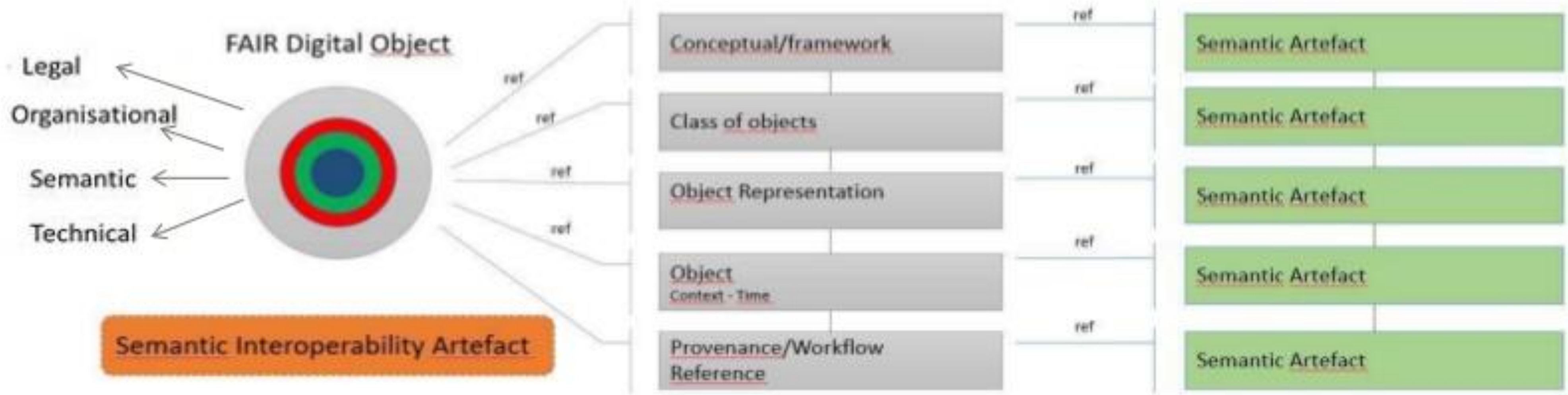


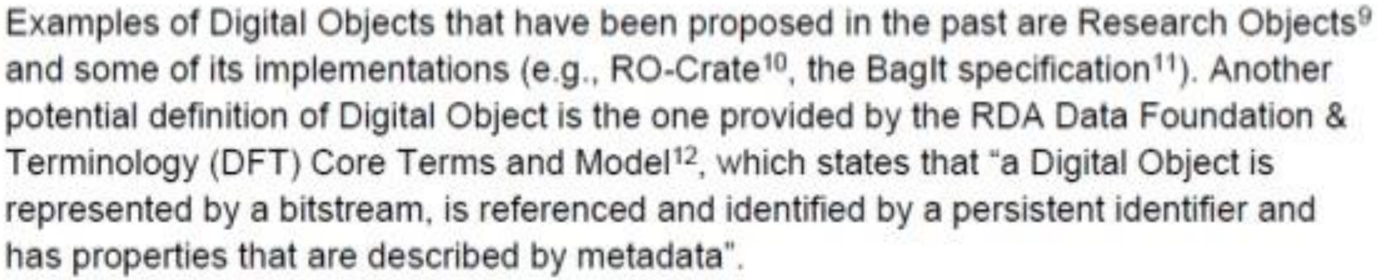
Research Objects?

Credit: S. Bechhofer et al., "Research Objects: Towards Exchange and Reuse of Digital Knowledge," 2010

STM Innovations Seminar U.S. – Reinventing Innovation
May 1, 2012
Washington DC, USA

European Open Science Cloud Interoperability Framework




Examples of Digital Objects that have been proposed in the past are Research Objects[9] and some of its implementations (e.g., RO-Crate[10], the BagIt specification[11]). Another potential definition of Digital Object is the one provided by the RDA Data Foundation & Terminology (DFT) Core Terms and Model[12], which states that "a Digital Object is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata".

FAIR DO ecosystem

EOSC Interoperability Framework (v1.0)
May 2020, Draft for community consultation
Chair: Oscar Corcho, UPM

2021 we start to combine

Used? Yes ……….

NIH Data Commons transferring and archiving **very large HTS datasets in a location-independent** way

keep the context of data content together when its scattered. Scalability

https://doi.org/10.1109/BigData.2016.7840618

NIH DataSTAGE **RO Composer** to **exchange** between Seven Bridges Platform genomics platform and the Mendeley Data repository

A framework for **standardizing and sharing computations and analyses** generated from High-throughput genome sequencing.

Standardized as IEEE 2791-2020

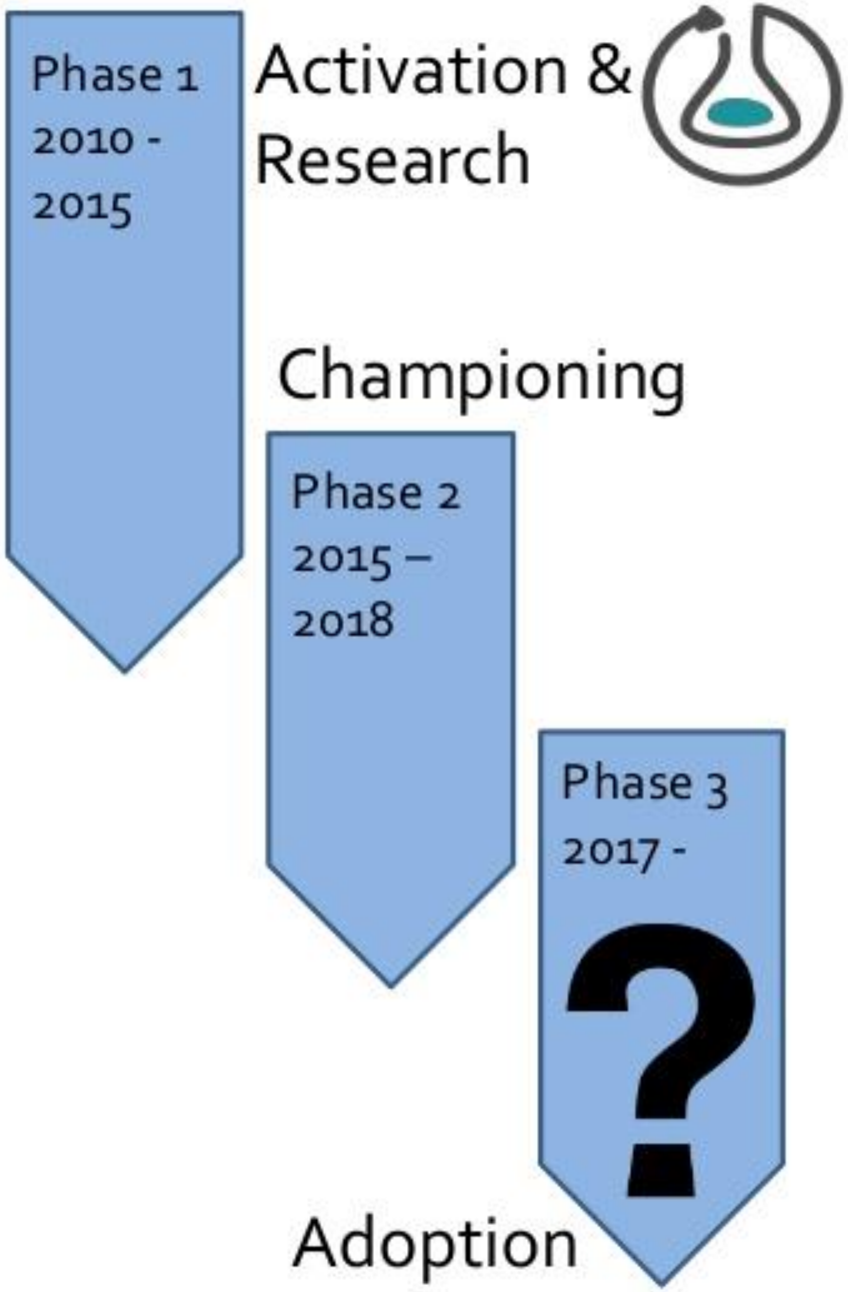
https://doi.org/10.1371/journal.pbio.3000099

Virtualized collaborative working environment for **Earth Science researchers** to share resources (data, workflows), ideas, knowledge, and results.

https://doi.org/10.1016/j.future.2019.03.046

Used? Yes ……….

	BIG DATA BAG	MENDELEY DATA Seven Bridges	BioCompute Objects FDA	everest ROHUB
Exchange	✓	✓	✓	✓
Reproducibility			✓	✓
Archival	✓	✓	✓	✓
Active Objects		✓	✓	✓



Activation & Research

Phase 1
2010 -
2015

Championing

Phase 2
2015 –
2018

Phase 3
2017 -

?

Adoption

time to reflect….

Desiderata & Norms
Balance and prioritise

Machine-processable

Standards

EXAMPLES

Low tech Incremental

Multi-platform

Graceful degradation

Commodity tooling

Technology Independent

Keep it simple

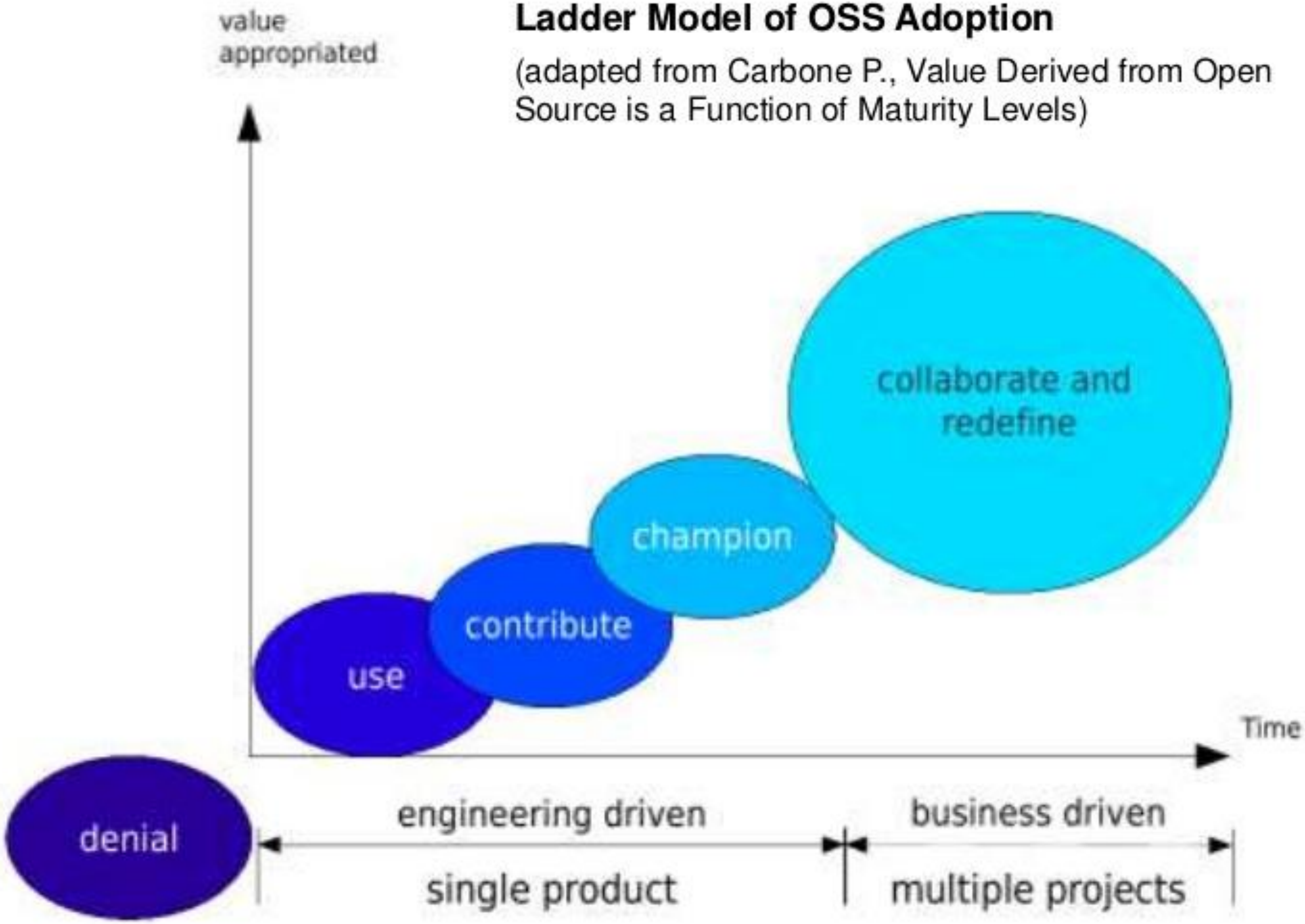
Developer friendliness

- "just enough complexity" or “just enough standards” so…
- **sufficient extra benefits** from what already exists (Linked Data, vocabularies, tooling, validation, transformation
- **without compromising the developer entry-level experience** so much that they rather do their own thing.

Research Object Tensions

Research Infrastructures sit in the middle

Academic Viewpoint	Infrastructure Viewpoint
Green field site	Pre-existing platforms
Theoretical purity	Practicality
Use latest thing	Use things that work
Proof of concept	Production
Sophistication	Simplicity
Narrow developer audience	Wide developer audience
Strive for super generic	Several specific is ok!
The end	The means
Exposing the tech	Hiding the tech



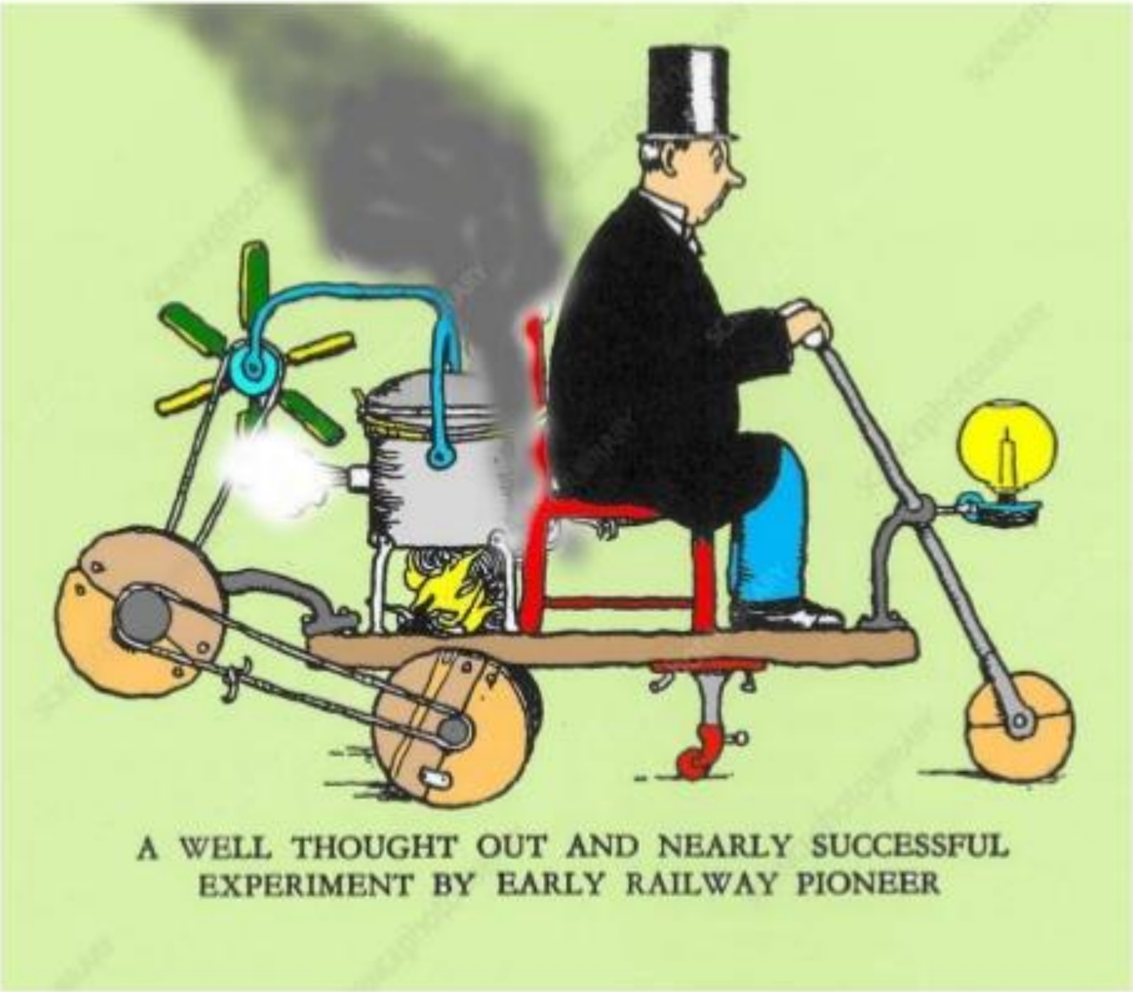
"it's better, initially, to make a small number of users really love you than a large number kind of like you"
Paul Buchheit
paulbuchheit.blogspot.com

[FLOSS@Sycracuse]

Not really mortal developer friendly

- "Easy to make, hard to use…"
- Daunting Linked Data tech stack
- Being too clever
  - Infer what is in the object and what kind of object it is
  - Massive reuse of ontologies
- Make developers (and researchers) lives easier not more demanding….

Developer friendliness matters

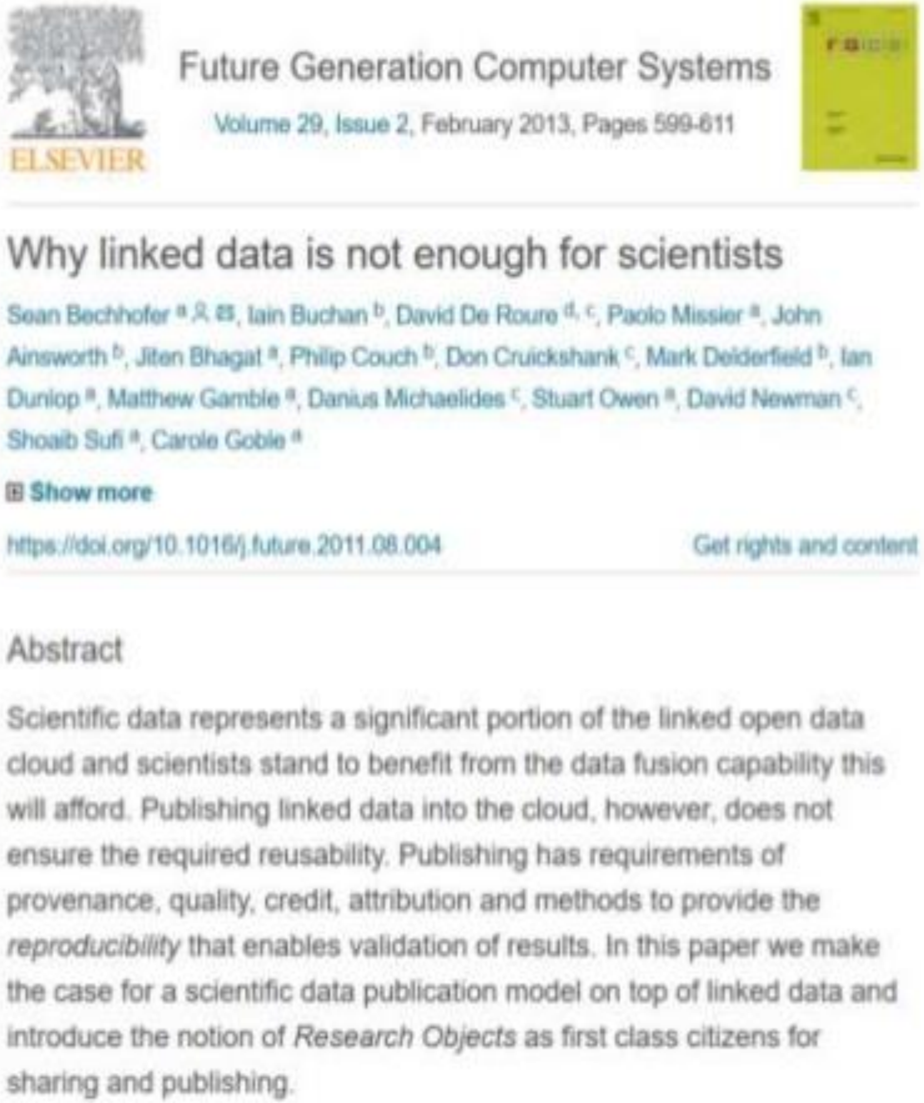

A WELL THOUGHT OUT AND NEARLY SUCCESSFUL
EXPERIMENT BY EARLY RAILWAY PIONEER

Reinvent with fewer features

Easy to understand and simple conceptually…

… with strong opinionated guide to current best practices

… using software stacks widely used on the Web



Future Generation Computer Systems

Volume 29, Issue 2, February 2013, Pages 599-611

Why linked data is not enough for scientists

Sean Bechhofer [a], Iain Buchan [b], David De Roure [d, c], Paolo Missier [a], John Ainsworth [b], Jiten Bhagat [a], Philip Couch [b], Don Cruickshank [c], Mark Delderfield [b], Ian Dunlop [a], Matthew Gamble [a], Danius Michaelides [c], Stuart Owen [a], David Newman [c], Shoaib Sufi [a], Carole Goble [a]

⊞ Show more

https://doi.org/10.1016/j.future.2011.08.004

Get rights and content

Abstract

Scientific data represents a significant portion of the linked open data cloud and scientists stand to benefit from the data fusion capability this will afford. Publishing linked data into the cloud, however, does not ensure the required reusability. Publishing has requirements of provenance, quality, credit, attribution and methods to provide the *reproducibility* that enables validation of results. In this paper we make the case for a scientific data publication model on top of linked data and introduce the notion of *Research Objects* as first class citizens for sharing and publishing.

Indeed.

Linked Data is not enough.

Research Infrastructures:
"digital technologies (hardware, software), resources (data, services, digital libraries, standards), comms (protocols, access rights, networks), **people and organisational structures**"

Linked Data and a Spec is not enough

and sometimes too much

Use Driver

Community

Reference examples

Tools

Guides

Exchange, reproducibility, executable objects

Portability between platforms, Archiving

Platform & user buy-in & consensus

Passionate, dedicated leadership

Active engaged community, seed Support

Developer friendly – so possible

Incentives – so rewarding

Adoption path – so acceptable

Metadata capture

Early benefit

Research Object Reboot

Sponsored

Community

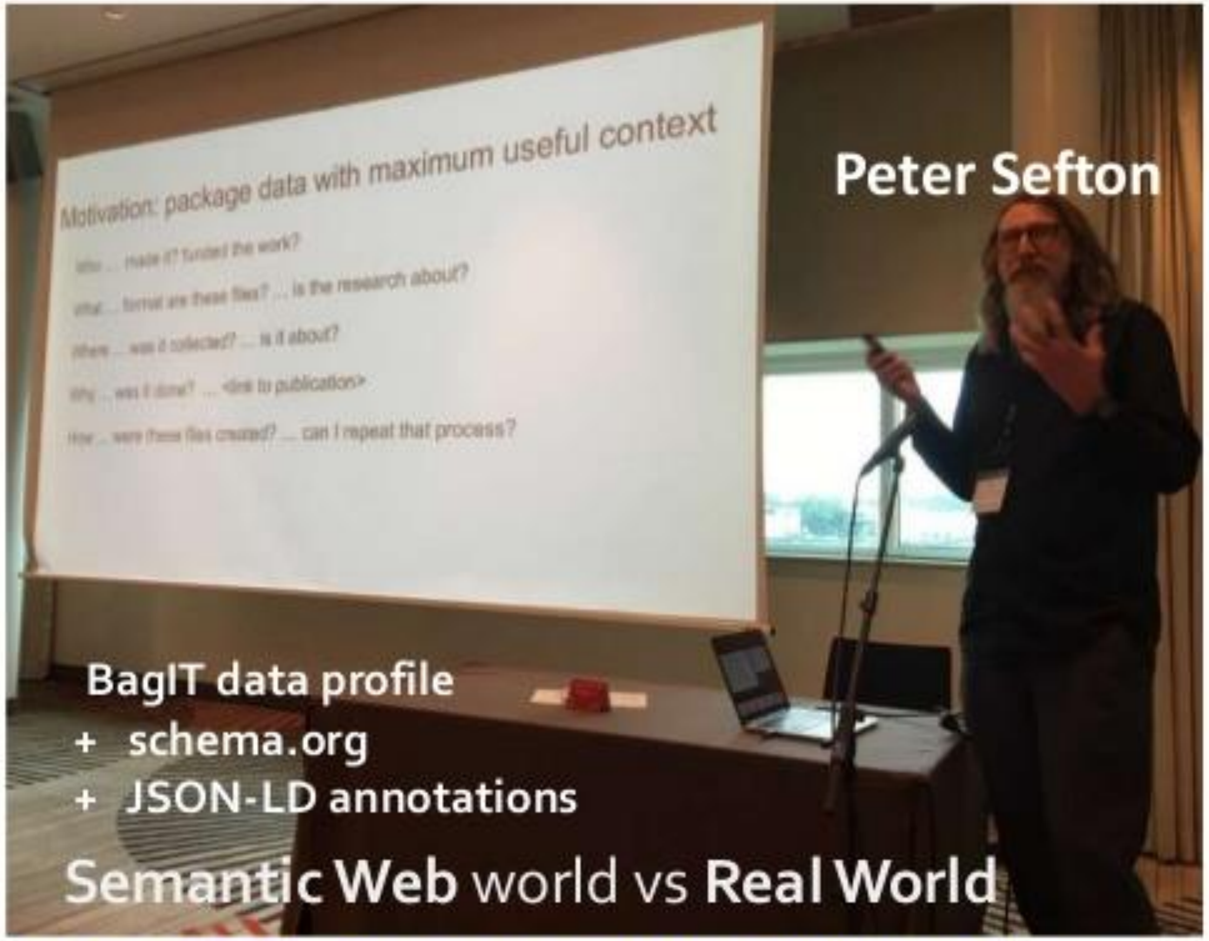
RO2018, Amsterdam
e-Science Conference

BDBags Keynote

Research Objects for Everyday Use
Carl Kesselman
Dean's Professor, Industrial and Systems Engineering
University of Southern California
Information Sciences Institute

Swing Back to Basics

DataCrate from the Open Repository community



Motivation: package data with maximum useful context

Who ... made it? funded the work?

What ... format are these files? ... is the research about?

Where ... was it collected? ... is it about?

Why ... was it done? ... <link to publication>

How ... were these files created? ... can I repeat that process?

Peter Sefton

BagIT data profile
+ schema.org
+ JSON-LD annotations

Semantic Web world vs **Real World**

"As a researcher…I'm a bit b****y fed up with Data Management", Cameron Neylon

Be Humble

Archivist and library people know the importance of metadata and standards…

… and for things to work 5, 10, 20 years later.

End-users need to have their own way to "bypass the system"…

…. their field, repositories, institutions , journals etc. will always be lagging behind the curve

Most who want to make their data is FAIR …

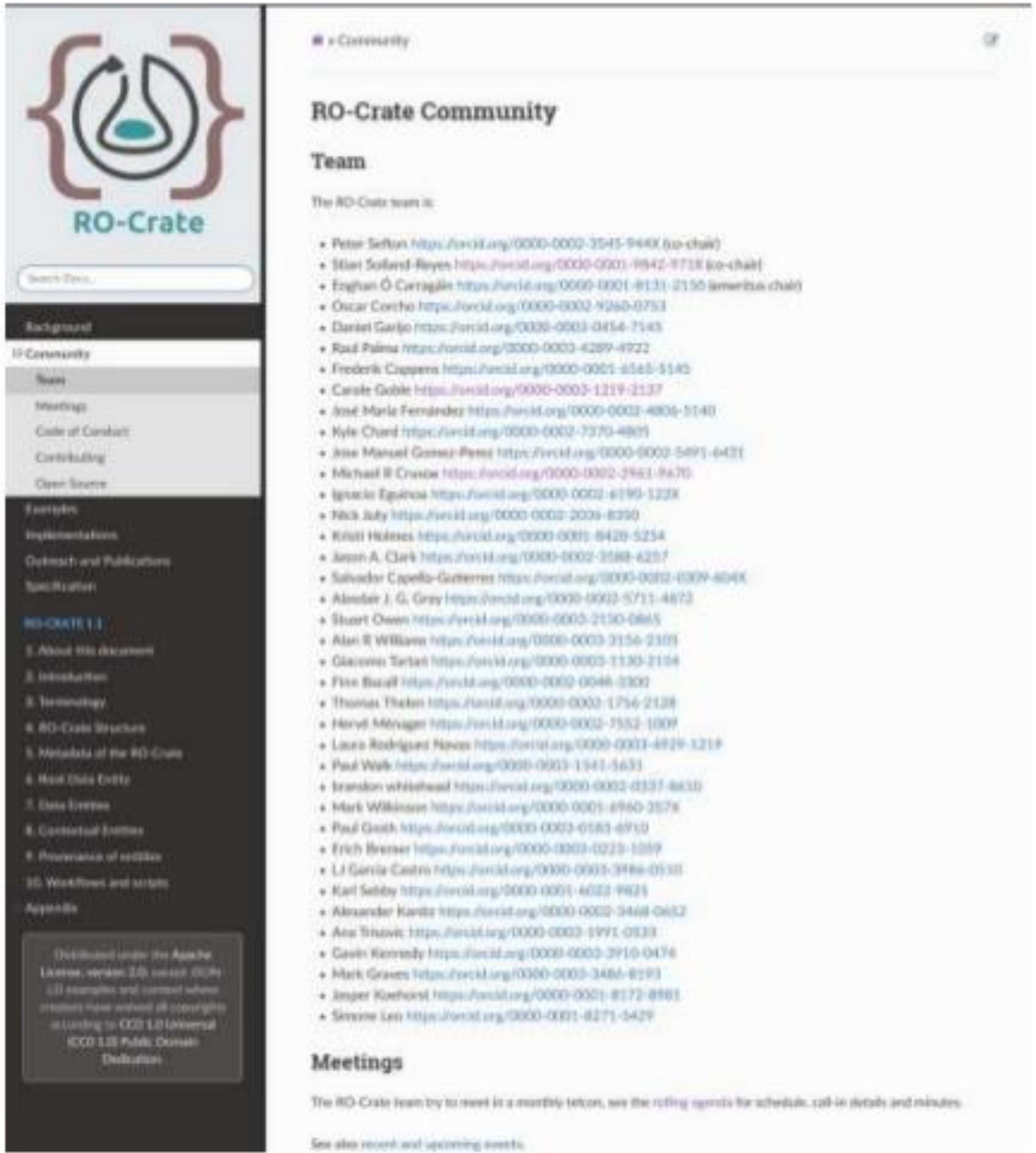
… do not have the resources or knowledge to start championing all of this to all levels & need tools and ramps.

http://www.lisbdnet.com/

https://ischools.org/

A RO-Crate Community!

A Merger



Distributed under the Apache License, version 2.0; except JSON-LD examples and content where creators have waived all copyrights according to CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

» Community

RO-Crate Community

Team

The RO-Crate team is:

- Peter Sefton https://orcid.org/0000-0002-3545-944X (co-chair)
- Stian Soiland-Reyes https://orcid.org/0000-0001-9842-9718 (co-chair)
- Eoghan Ó Carragáin https://orcid.org/0000-0001-8131-2150 (emeritus chair)
- Oscar Corcho https://orcid.org/0000-0002-9260-0753
- Daniel Garijo https://orcid.org/0000-0003-0454-7145
- Raul Palma https://orcid.org/0000-0003-4289-4922
- Frederik Coppens https://orcid.org/0000-0001-6565-5145
- Carole Goble https://orcid.org/0000-0003-1219-2137
- José María Fernández https://orcid.org/0000-0002-4806-5140
- Kyle Chard https://orcid.org/0000-0002-7370-4805
- Jose Manuel Gomez-Perez https://orcid.org/0000-0002-5491-6431
- Michael R Crusoe https://orcid.org/0000-0002-2961-9670
- Ignacio Eguinoa https://orcid.org/0000-0002-6190-122X
- Nick Juty https://orcid.org/0000-0002-2036-8350
- Kristi Holmes https://orcid.org/0000-0001-8420-5254
- Jason A. Clark https://orcid.org/0000-0002-3588-6257
- Salvador Capella-Gutierrez https://orcid.org/0000-0002-0309-604X
- Alasdair J. G. Gray https://orcid.org/0000-0002-5711-4872
- Stuart Owen https://orcid.org/0000-0003-2130-0865
- Alan R Williams https://orcid.org/0000-0003-3156-2105
- Giacomo Tartari https://orcid.org/0000-0003-1130-2154
- Finn Bacall https://orcid.org/0000-0002-0048-3300
- Thomas Thelen https://orcid.org/0000-0002-1756-2128
- Hervé Ménager https://orcid.org/0000-0002-7552-1009
- Laura Rodríguez Navas https://orcid.org/0000-0003-4929-1219
- Paul Walk https://orcid.org/0000-0003-1541-5631
- brandon whitehead https://orcid.org/0000-0002-0337-8610
- Mark Wilkinson https://orcid.org/0000-0001-6960-357X
- Paul Groth https://orcid.org/0000-0003-0183-6910
- Erich Bremer https://orcid.org/0000-0003-0223-1059
- LJ Garcia Castro https://orcid.org/0000-0003-3986-0510
- Karl Sebby https://orcid.org/0000-0001-6022-9825
- Alexander Kanitz https://orcid.org/0000-0002-3468-0652
- Ana Trisovic https://orcid.org/0000-0003-1991-0533
- Gavin Kennedy https://orcid.org/0000-0003-3910-0474
- Mark Graves https://orcid.org/0000-0003-3486-8193
- Jasper Koehorst https://orcid.org/0000-0001-8172-8981
- Simone Leo https://orcid.org/0000-0001-8271-5429

Meetings

The RO-Crate team try to meet in a monthly telcon, see the rolling agenda for schedule, call-in details and minutes.

See also recent and upcoming events.

https://www.researchobject.org/ro-crate/#contribute

https://github.com/researchobject/ro-crate/issues/1

- A diverse set of people
- A variety of stakeholders
- A set of collective norms
- A open platform that facilitates communication (GitHub, Google Docs, monthly telcons)

RO-Crate

Specifications and Tooling

It is **recommended** that new Research Object users adapt the RO-Crate specification.

RO-Crate is a **community effort** to establish a **lightweight** approach to **packaging research data** with their **metadata**.

It is based on schema.org annotations in JSON-LD, and aims to make best-practice in formal metadata description **accessible** and practical for use in a wider variety of situations, from an individual researcher working with a **folder of data**, to large data-intensive computational research environments.

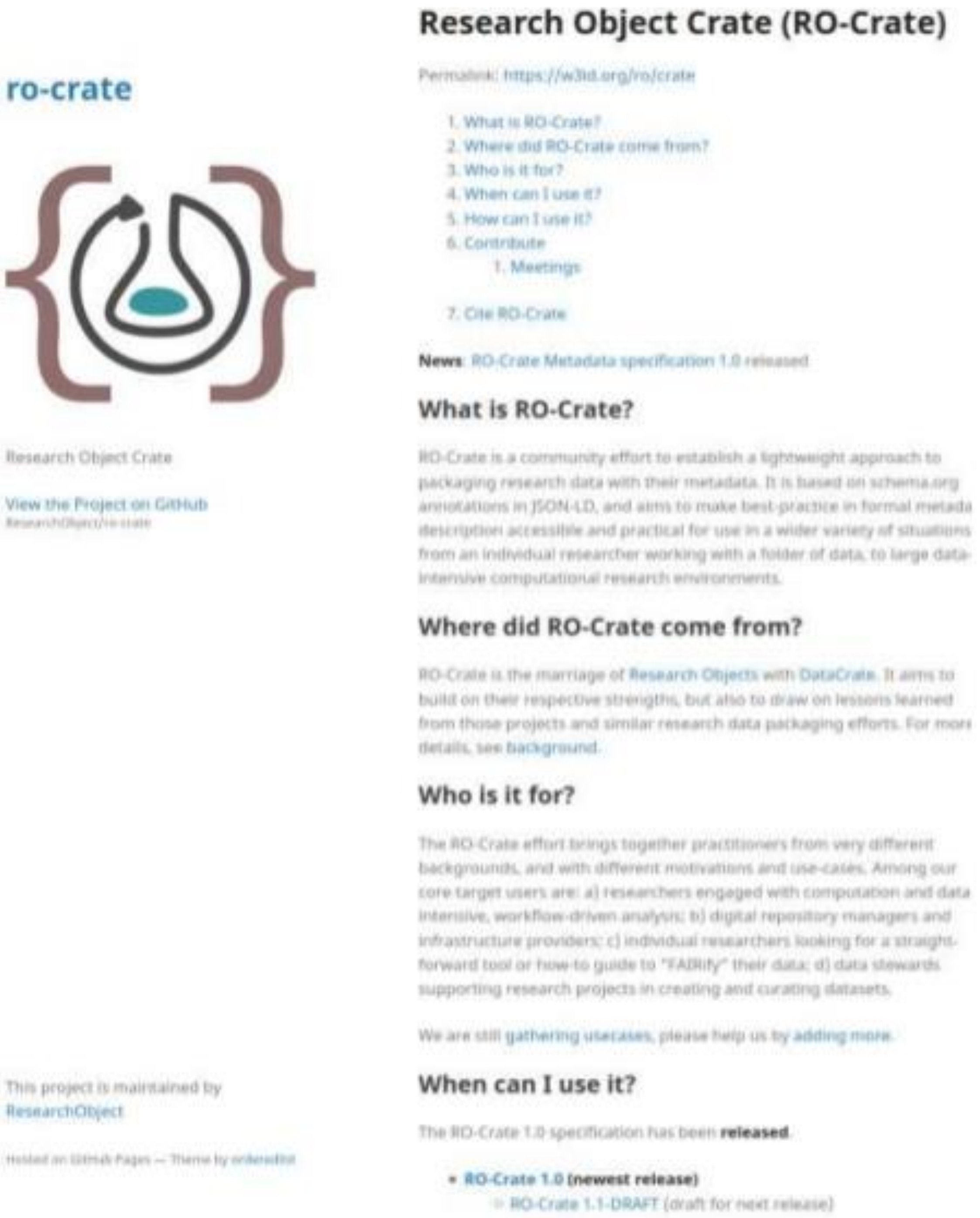
RO-Crate is the **marriage of Research Objects with DataCrate**. It aims to build on their respective strengths, but also to draw on lessons learned from those projects and similar research data packaging efforts. For more details, see RO-Crate background.

The RO-Crate specification details how to capture a set of files and resources as a dataset with associated metadata – including **contextual entities** like *people, organizations, publishers, funding, licensing, provenance, workflows, geographical places, subjects and repositories*.

A growing list of RO-Crate tools and libraries simplify creation and consumption of RO-Crates, including the graphical interface Describo.

The RO-Crate community help shape the specification or get help with using it!

https://w3id.org/ro/crate



ro-crate

Research Object Crate

View the Project on GitHub
ResearchObject/ro-crate

This project is maintained by
ResearchObject

Hosted on GitHub Pages — Theme by orderedlist

Research Object Crate (RO-Crate)

Permalink: https://w3id.org/ro/crate

News: RO-Crate Metadata specification 1.0 released

What is RO-Crate?

RO-Crate is a community effort to establish a lightweight approach to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD, and aims to make best-practice in formal metada description accessible and practical for use in a wider variety of situations from an individual researcher working with a folder of data, to large data-intensive computational research environments.

Where did RO-Crate come from?

RO-Crate is the marriage of Research Objects with DataCrate. It aims to build on their respective strengths, but also to draw on lessons learned from those projects and similar research data packaging efforts. For more details, see background.

Who is it for?

The RO-Crate effort brings together practitioners from very different backgrounds, and with different motivations and use-cases. Among our core target users are: a) researchers engaged with computation and data intensive, workflow-driven analysis; b) digital repository managers and infrastructure providers; c) individual researchers looking for a straight-forward tool or how-to guide to "FAIRify" their data; d) data stewards supporting research projects in creating and curating datasets.

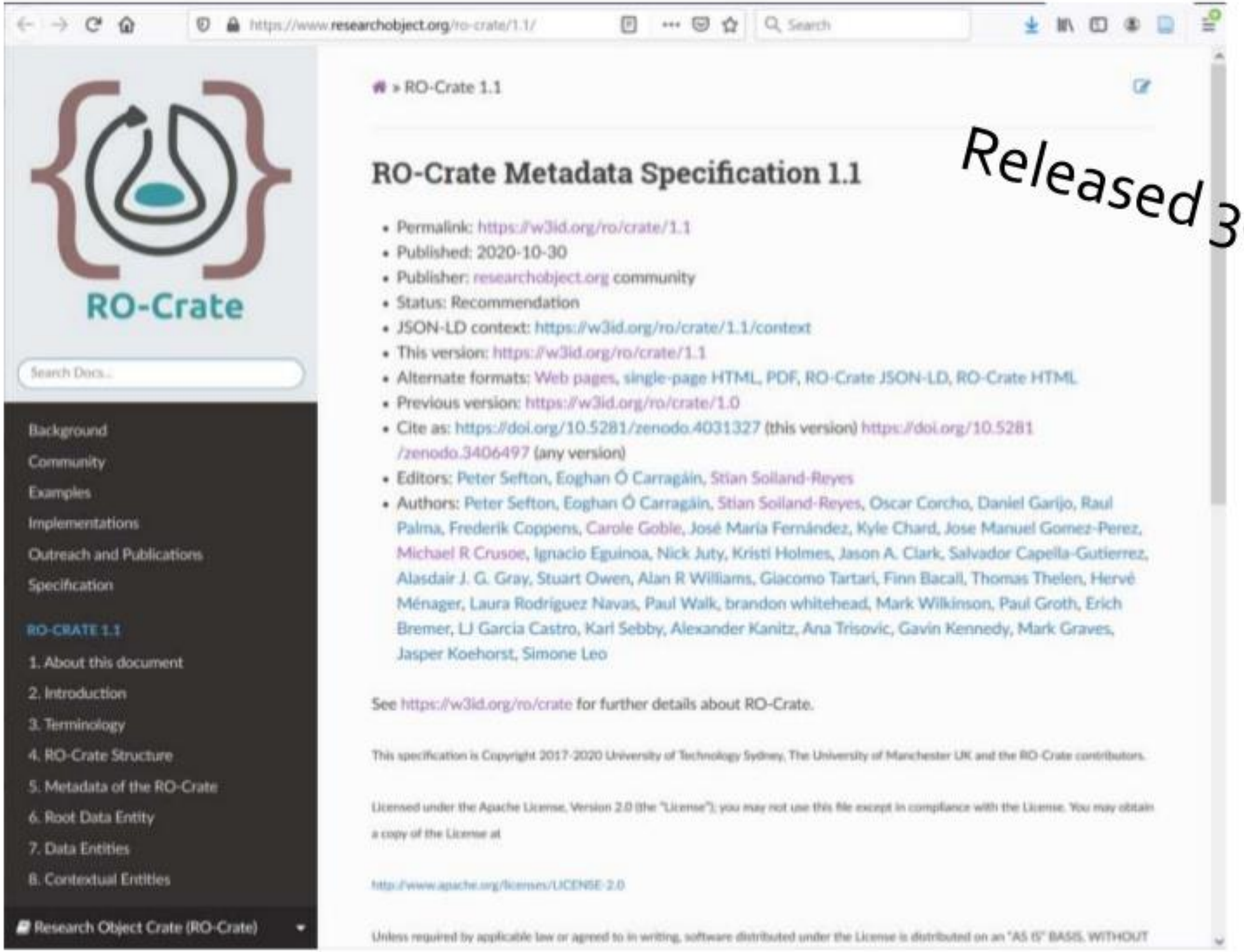
We are still gathering usecases, please help us by adding more.

When can I use it?

The RO-Crate 1.0 specification has been **released**.

- RO-Crate 1.0 **(newest release)**
    - RO-Crate 1.1-DRAFT (draft for next release)

Released 30th October 2020



RO-Crate Metadata Specification 1.1

- Permalink: https://w3id.org/ro/crate/1.1
- Published: 2020-10-30
- Publisher: researchobject.org community
- Status: Recommendation
- JSON-LD context: https://w3id.org/ro/crate/1.1/context
- This version: https://w3id.org/ro/crate/1.1
- Alternate formats: Web pages, single-page HTML, PDF, RO-Crate JSON-LD, RO-Crate HTML
- Previous version: https://w3id.org/ro/crate/1.0
- Cite as: https://doi.org/10.5281/zenodo.4031327 (this version) https://doi.org/10.5281/zenodo.3406497 (any version)
- Editors: Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes
- Authors: Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes, Oscar Corcho, Daniel Garijo, Raul Palma, Frederik Coppens, Carole Goble, José María Fernández, Kyle Chard, Jose Manuel Gomez-Perez, Michael R Crusoe, Ignacio Eguinoa, Nick Juty, Kristi Holmes, Jason A. Clark, Salvador Capella-Gutierrez, Alasdair J. G. Gray, Stuart Owen, Alan R Williams, Giacomo Tartari, Finn Bacall, Thomas Thelen, Hervé Ménager, Laura Rodriguez Navas, Paul Walk, brandon whitehead, Mark Wilkinson, Paul Groth, Erich Bremer, LJ Garcia Castro, Karl Sebby, Alexander Kanitz, Ana Trisovic, Gavin Kennedy, Mark Graves, Jasper Koehorst, Simone Leo

See https://w3id.org/ro/crate for further details about RO-Crate.

Reinvent as Lightweight Underware
Linked data but **Developer Friendly**

Easy to understand and simple conceptually…

Data entities are files/directories or web resources
Boundness of elements is explicit
Single graph, data structure depth 1

… with strong opinionated guide to current best practices

Example driven rather than strict specification
Implementers add additional metadata using schema.org types and properties

… using software stacks widely used on the Web

BagIT data profile , schema.org, JSON-LD, JSONSchema
Flattened compacted JSON-LD, no need for RDF libraries

Swung a bit far….
and swung back…

Tooling!

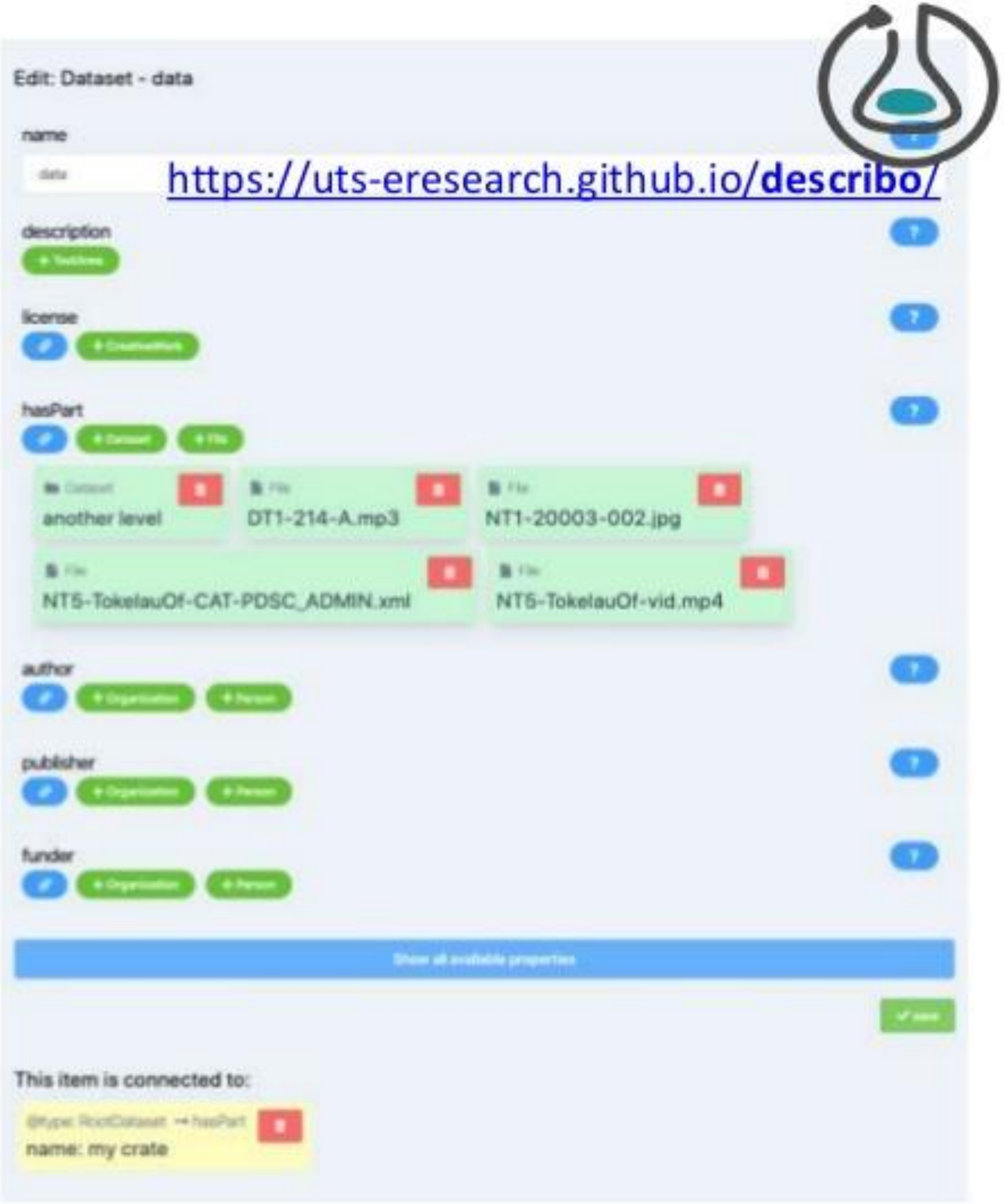
How can I use it?

While we're mostly focusing on the specification, some tools already exist for working with RO-Crates:

- Describo interactive **desktop application** to create, update and export RO-Crates for different profiles. (~ *beta*)
- CalcyteJS is a command-line tool to help create RO-Crates and HTML-readable rendering (~ *beta*)
- ro-crate - JavaScript/NodeJS library for RO-Crate rendering as HTML. (~ *beta*)
- ro-crate-js - utility to render HTML from RO-Crate (~ *alpha*)
- ro-crate-ruby Ruby library to consume/produce RO-Crates (~ *alpha*)
- ro-crate-py **Python library** to consume/produce RO-Crates (~ *planning*)

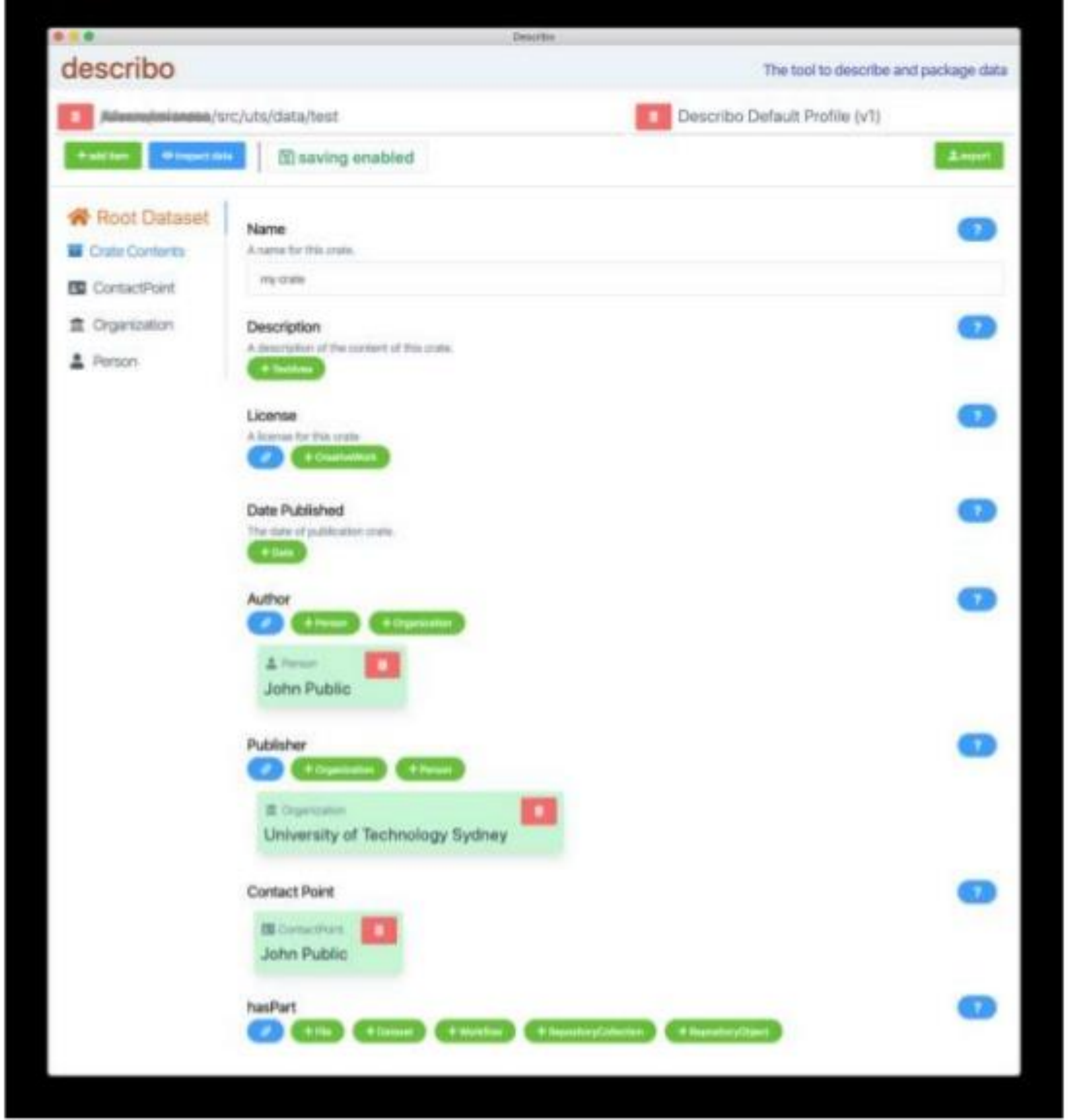
These applications use or expose RO-Crates:

- Workflow Hub imports and exports Workflow RO-Crates
- OCFL-indexer NodeJS application that walks the Oxford Common File Layout on the file system, validate RO-Crate Metadata Files and parse into objects registered in Elasticsearch. (~ *alpha*)
- ONI indexer
- ocfl-tools
- ocfl-viewer
- Research Object Composer is a REST API for gradually building and depositing Research Objects according to a pre-defined profile. (RO-Crate support *alpha*)
- ... (yours?)



https://uts-eresearch.github.io/**describo/**

Under-ware



[redacted]/src/uts/data/test

Describo Default Profile (v1)

Name

A name for this crate.

my crate

Description

A description of the content of this crate.

License

A license for this crate

Date Published

The date of publication crate.

Author

Person

John Public

Publisher

Organization

University of Technology Sydney

Contact Point

ContactPoint

John Public

hasPart

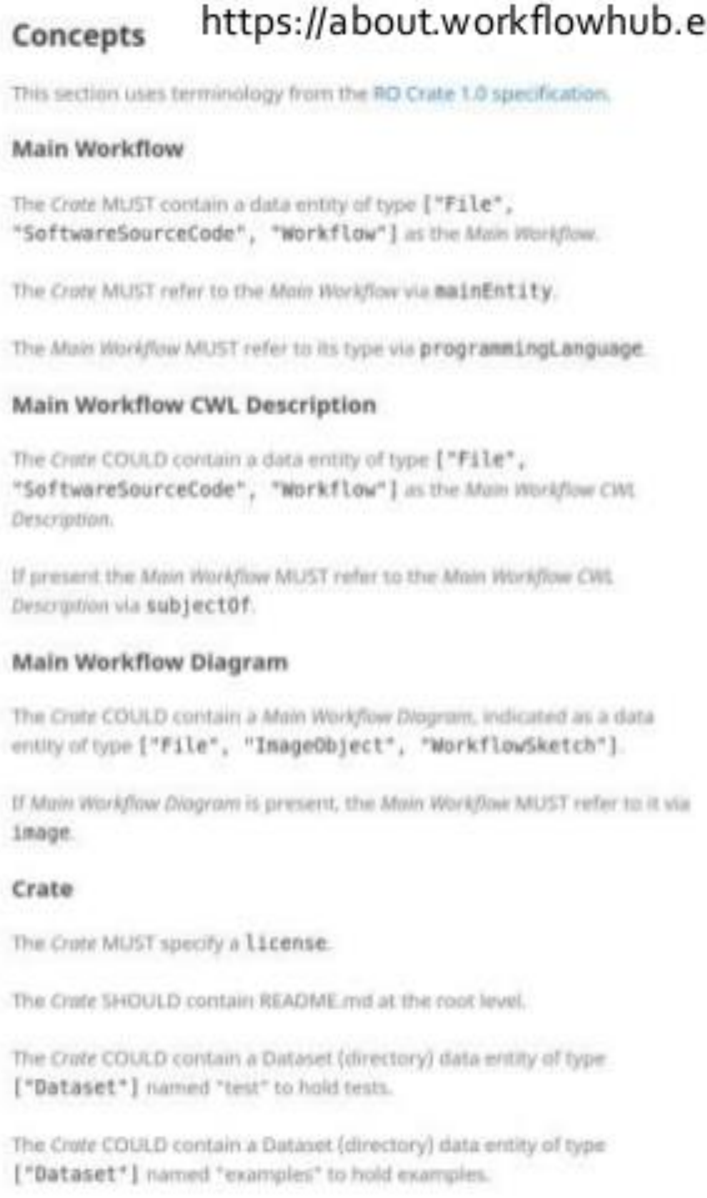
- RDF and schema.org but you don't need to know.
- Extend RO-Crate ….
  - Add your own schema.org types and properties.
  - Add in your own ontologies

…and it still works!

https://arkisto-platform.github.io/case-studies/

Driver! Profile for workflows

https://about.workflowhub.eu/Workflow-RO-Crate/



Concepts

This section uses terminology from the RO Crate 1.0 specification.

Main Workflow

The *Crate* MUST contain a data entity of type `["File", "SoftwareSourceCode", "Workflow"]` as the *Main Workflow*.

The *Crate* MUST refer to the *Main Workflow* via `mainEntity`.

The *Main Workflow* MUST refer to its type via `programmingLanguage`.

Main Workflow CWL Description

The *Crate* COULD contain a data entity of type `["File", "SoftwareSourceCode", "Workflow"]` as the *Main Workflow CWL Description*.

If present the *Main Workflow* MUST refer to the *Main Workflow CWL Description* via `subjectOf`.

Main Workflow Diagram

The *Crate* COULD contain a *Main Workflow Diagram*, indicated as a data entity of type `["File", "ImageObject", "WorkflowSketch"]`.

If *Main Workflow Diagram* is present, the *Main Workflow* MUST refer to it via `image`.

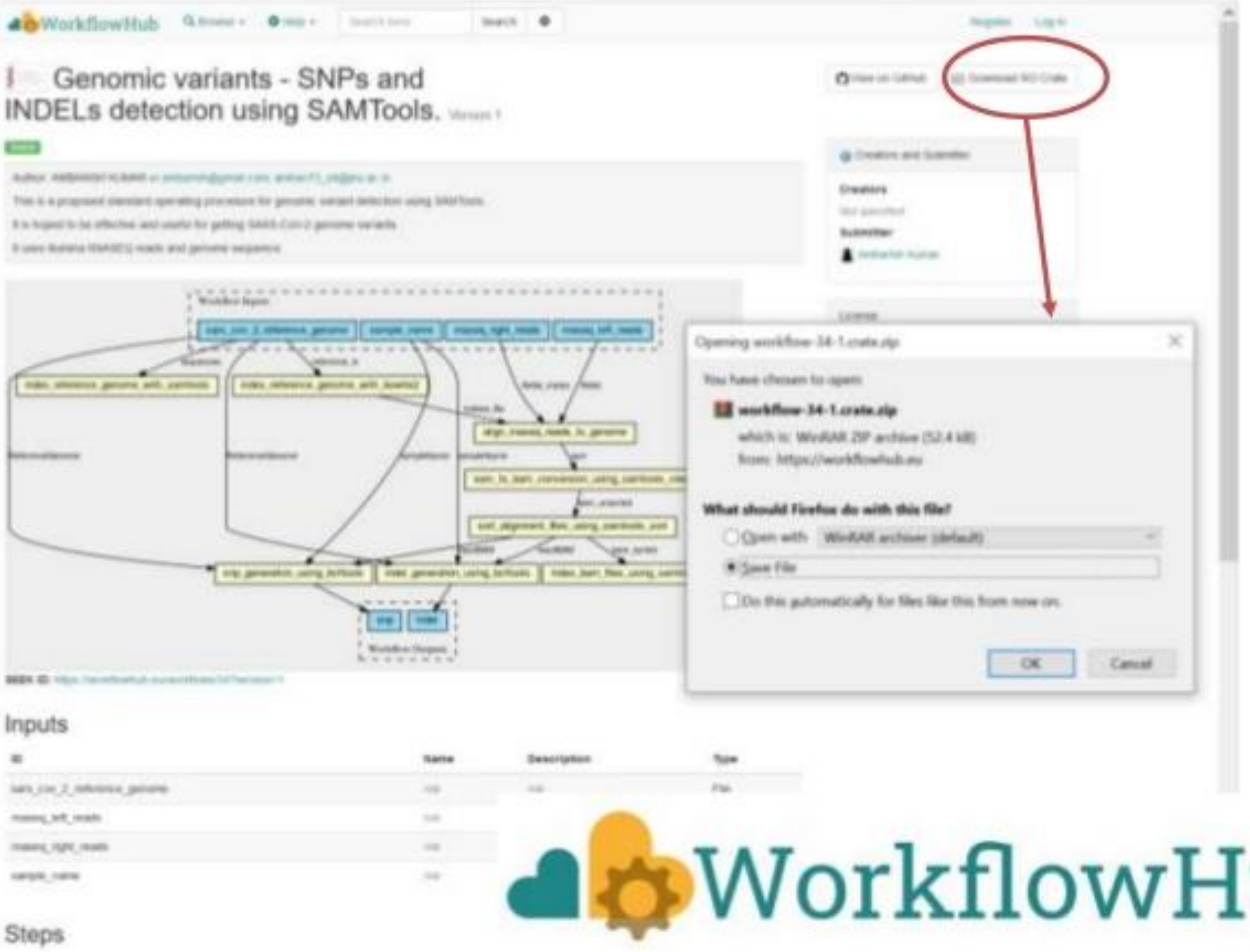
Crate

The *Crate* MUST specify a `license`.

The *Crate* SHOULD contain README.md at the root level.

The *Crate* COULD contain a Dataset (directory) data entity of type `["Dataset"]` named "test" to hold tests.

The *Crate* COULD contain a Dataset (directory) data entity of type `["Dataset"]` named "examples" to hold examples.



Genomic variants - SNPs and INDELs detection using SAMTools. Version 1

This is a proposed standard operating procedure for genomic variant detection using SAMTools.

It is hoped to be effective and useful for getting SARS-CoV-2 genome variants.

It uses Illumina RNASEQ reads and genome sequence.

Opening workflow-34-1.crate.zip

You have chosen to open:

workflow-34-1.crate.zip

which is: WinRAR ZIP archive (52.4 kB)

from: https://workflowhub.eu

What should Firefox do with this file?

○ Open with WinRAR archiver (default)

● Save File

□ Do this automatically for files like this from now on.

OK Cancel

Driver! Profile for workflows

Infrastructure families

EOSC-Life bioexcel DISSCO

On-boarding developers

Galaxy EUROPE nextflow COMMON WORKFLOW LANGUAGE jupyter

OEB OPENEBENCH Life Monitor Snakemake WorkflowHub

Web and dev friendly

Bioschemas.org schema.org

Workflow-RO-Crate profiles

RO in practice

External references – logically & physically contained – versions, snapshots, multi-typed, active, multi-stewarded, multi-authored, governance…

More than plain JSON, Just Enough Linked Data

Retain **benefits of Linked Data in the toolbox**

- querying, graph stores, vocabularies, clickable URI as identifiers)
- customization and conventions

Plus all the other stuff a **developer expects**

- documentation, examples, libraries, tools
- simplifications rather than generalizations (less flexibility frees up developers)
- "Just enough standards" cf. schema.org

Linked Data "exotics" there for when the time is right if needed by the right people.

Keep your eye on the target.....

How do we make RO's normative?

- Propaganda and incentive models to scientists, target the Research Infrastructures to deliver.
- Digital library community allies!

Developer friendliness matters

- Underware, incremental, ramps, embed, metadata automation, persuasive design

Linked Data has a role

- As a means but it is not an end.
- Simpler version of Linked Data makes an adoption path (cf. Knowledge Graphs, schema.org, JSON-LD)

FAIR principles for Research Objects....

- Unifying the vision with the practical

http://researchobject.org

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
Stuart Owen
Finn Bacall
Bert Droebeke
Laura Rodríguez Navas
Ignacio Eguinoa

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
Peter Sefton
Eoghan Ó Carragáin
Frederik Coppens
Jasper Koehorst
Simone Leo
Nick Juty
LJ Garcia Castro
Karl Sebby
Alexander Kanitz
Ana Trisovic

Gavin Kennedy
Mark Graves
José María Fernández Jose
Manuel Gomez-Perez
Jason A. Clark
Salvador Capella-Gutierrez
Alasdair J. G. Gray
Kristi Holmes
Giacomo Tartari
Hervé Ménager
Paul Walk
Brandon Whitehead
Erich Bremer
Mark Wilkinson
Jen Harrow

And many more….

Acknowledgements

MANCHESTER 1824
The University of Manchester

SYNTHESYS+

EOSC-Life

FAIRDOM

FAIR4CURES

elixir

IBISBA

BioCompute Objects

bioexcel

COMMON WORKFLOW LANGUAGE

Workflow 4Ever

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Federal Ministry of Education and Research

DCPPC

BBSRC
bioscience for the future

NIH
NIH Big Data to Knowledge (BD2K)

MRC
Medical Research Council