Accelerating Discovery: Status and Trajectory of Supporting Tools and Systems

Carole Goble
The University of Manchester, UK
European Life Science Research Infrastructures ELIXIR & IBISBA
European Open Science Cloud Life Cluster
(co-author of the infamous FAIR paper ☺)
carole.goble@manchester.ac.uk

NAS Workshop Opportunities for Accelerating Scientific Discovery: Realizing the Potential of Advanced and Automated Workflows, 16-17 March 2020

Summary of Background
20 years of scientific workflow activity

Development

Workflow Management Systems

Apache Taverna
Open source & domain independent tools for designing and executing scientific workflows
http://taverna.incubator.apache.org/

Support for the Workflow Ecosystem

Workflow 4Ever

Principles of wf sharing, lifecycles, preservation, packaging

myexperiment
Workflow Hub

WfMS agnostic repository / registries

BioCatalogue

Registries of services

Metadata & identifier schemes for data, tools & workflows Standards

Bioschemas.org

COMMON WORKFLOW LANGUAGE

W3C PROV

Workflow RO-Crate

researchobject.org

BioCompute Objects

Deployment / Adoption

Workflows in scientific domains

BioVeL
biodiversity

HELIO
astrophysics

elixir
lico
life science 'omics

SYNTHESYS+
Synthesis of Systematic Resources
natural history specimens

IBISBA 1.0
Industrial Biotechnology Innovation and Synthetic Biology Accelerator

SCAPE
SCAlable Preservation Environments
digital libraries

sys and syn biology
Industrial Biotech

NeISS
social sciences


VPH

bioexcel
Center of Excellence for Computational Biomolecular Research
biomolecular modelling at exascale

Accelerating information exchange

Summary of Background

Distributed EU Research Infrastructures for the Life Sciences & the European Open Science Cloud.

23 nations, 230+ orgs
incl. EMBL-EBI

"Data for Life" Data Infrastructure for Life Sciences: coordinate, integrate and sustains resources across member states; enable users to access services.

9 nations, 16 orgs

"Together innovating for sustainable biotechnology" Accelerate end-to-end bioprocess development, linking distributed R&D facilities: computer-assisted design of biocomponents, construction of enzymes and microbial strains, fermentation, downstream processing.

Everyone

"A trusted virtual environment" to store, share & re-use research information to serve Europe's 1.7 million researchers and 70 million science and tech professionals and to foster interdisciplinary working.

Example of a Galaxy pipeline. Human in the Loop!!

https://www.ibisba.eu/

IBISBA 1.0

RP2Path — *Delepine et al. **Met. Eng.** 2018*

Selenzyme — *Carbonell et al. **Bioinformatics** 2018*, *Mellor et al. **ACS Synth. Biol.** 2016*

Pathway ranker, FBA — *Carbonell et al. **NAR** 2014*

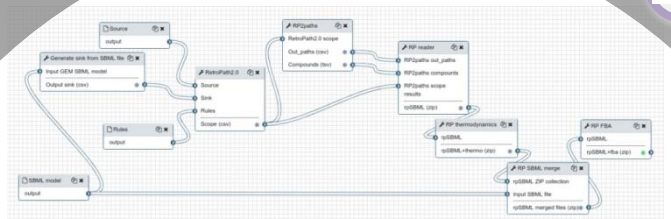
*Delepine et al. **Met. Eng.** 2018*

DoE — *Carbonell et al. **IWBDA** 2019*

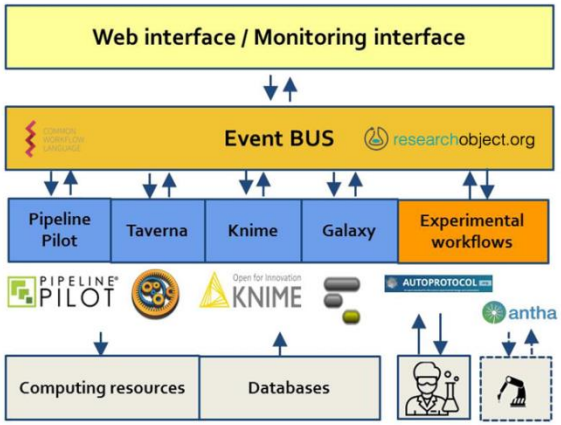
*Duigou et al. **NAR.** 2018*

PathwayGenie — *Swainston et al. **Bioinformatics** 2018*

SBML · SBML · SBML · SMART · SBOL · FASTA · FASTA



The stages may be computational or lab-based (SOPS, Lab Protocols).



Active Learning — *Borkowski et al. **BioRxiv,** 2019*

Data analysis — *Koch et al. **J Cheminform**. 2017*

MS screening

Product extraction

Culture growth

Chassis transformation
Chassis Engineering

Sequence verification

Pathway assembly — *Carbonell et al. **Commun Biol.** 2018*

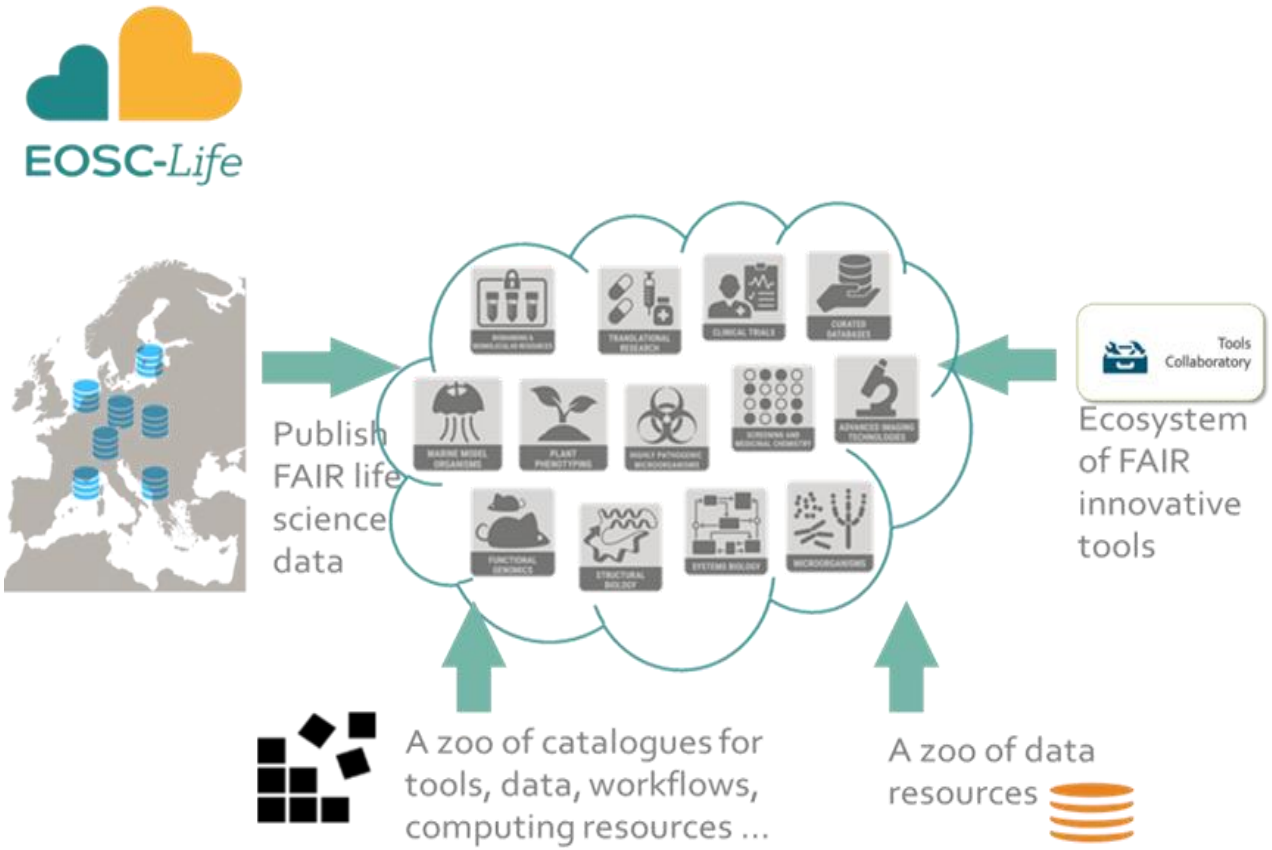
DNA synthesis

BIOBOOST

elixir Microbial Biotech Community

EOSC Life FAIR Workflow Collaboratory

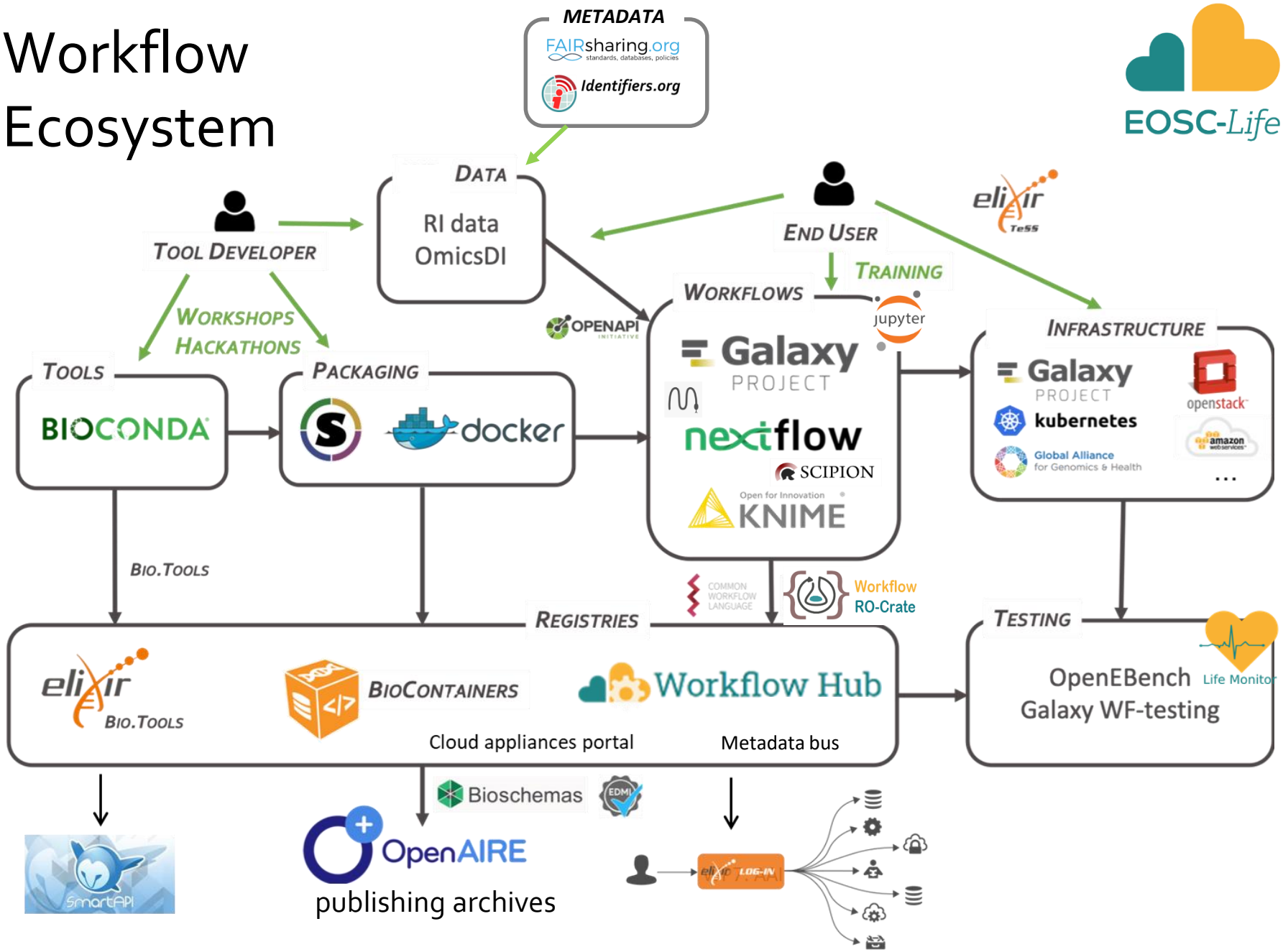
The Workflow Ecosystem for 13+1 European Research Infrastructures in the Life Sciences



- Democratising the use of workflows
- Accelerating the adoption, exchange and use of workflow methodology
- Accelerating the creation of workflows.
- Supporting the scientific workflow for different communities in the wild, at the reality coal face
- Building workflow capacity

https://elixir-europe.org/

Workflow Ecosystem



Challenge: Interoperating and Sustaining

We finally have a tech stack

EOSC-Life

Registries, Repositories, Portals
Find & launch

Description for finding, running, reporting

Interoperability

Portable packaging

Algorithm Design & Execution using WfMS

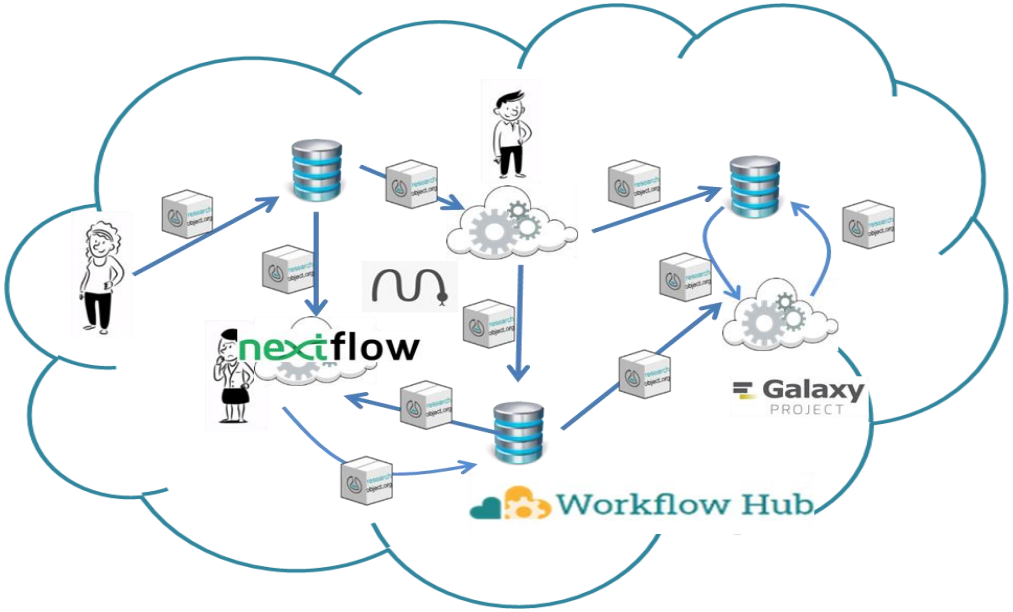
Versioning
Development

FAIR Commons using Research Objects as the Currency of Science, supporting interoperability, a metadata bus and Knowledge Graph of workflows and data

EOSC-*Life*

COMMON WORKFLOW LANGUAGE

Describes workflows to be **portable, scalable & interoperable** with different workflow systems and containerised tools (GA4GH Genomic Data Toolkit)

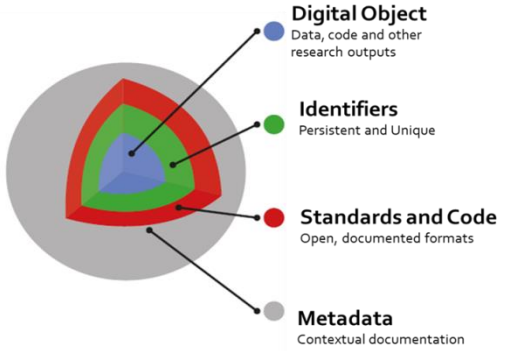


researchobject.org

Bundles descriptions, references, files
Adds context, provenance, examples, data
Relates data collections, SOPs, lab protocols
Links descriptions with native workflows

Metadata packaging framework

FAIR Digital Object Framework

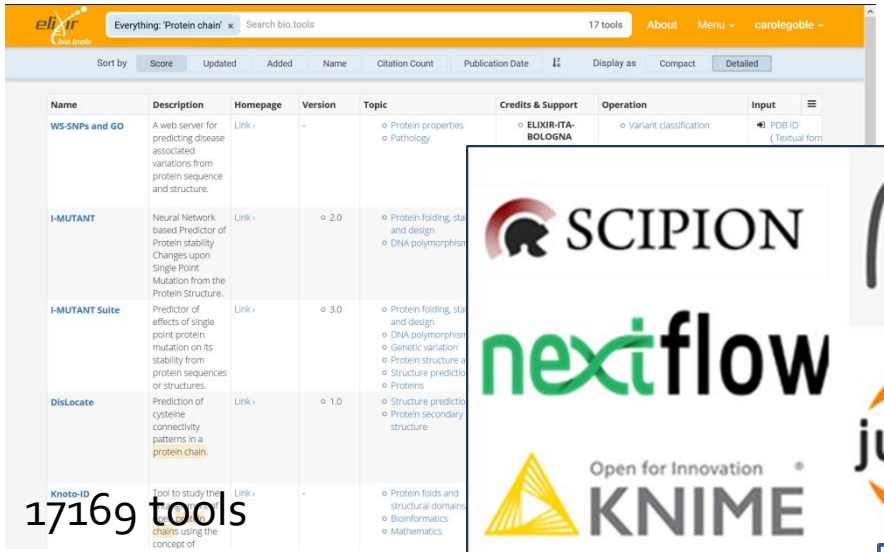


EC "Turning FAIR into Reality, 2018

https://www.commonwl.org/ http://www.researchobject.org/

What workflows tools / systems have you worked on or been involved with?

EOSC-Life



17169 tools

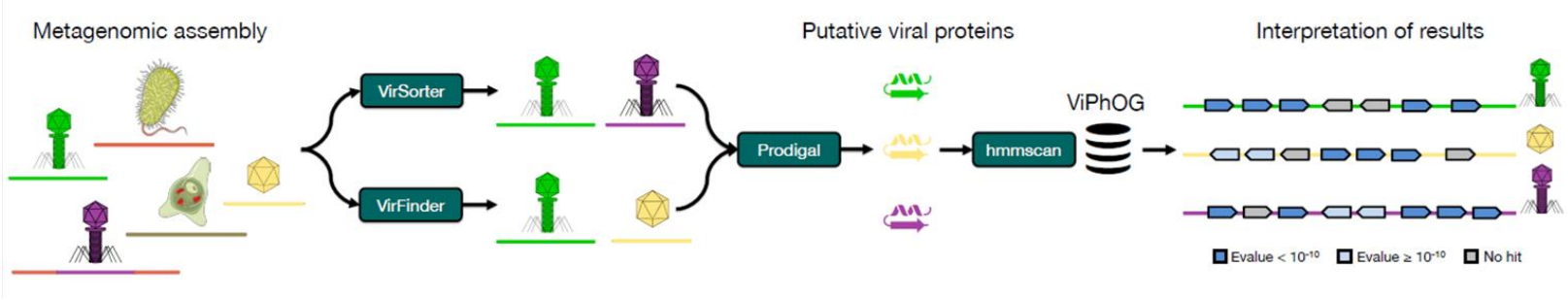
EDAM Ontology

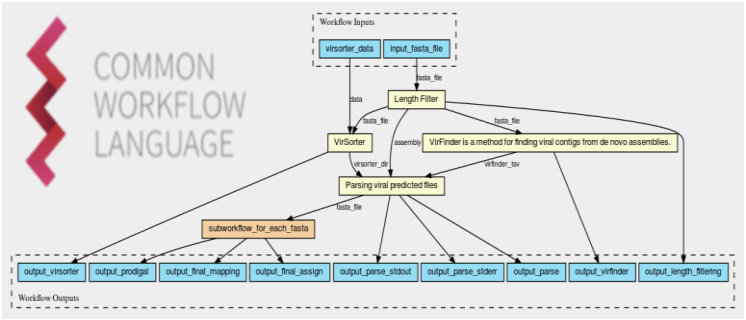
COMMON WORKFLOW LANGUAGE

A Zoo!

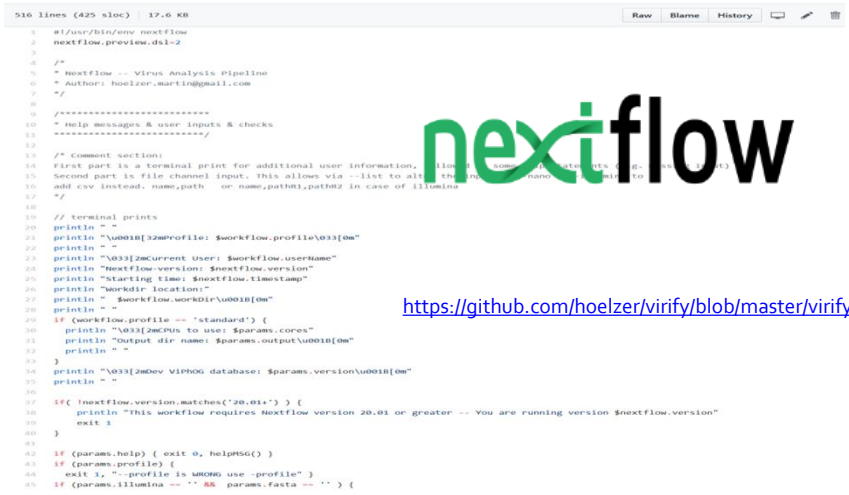
And don't forget the different data formats, identifiers, repositories & catalogues!

https://s.apache.org/existing-workflow-systems





https://github.com/EBI-Metagenomics/emg-virify-pipeline/blob/master/CWL/WorkFlow/pipeline.cwl



https://github.com/hoelzer/virify/blob/master/virify.nf

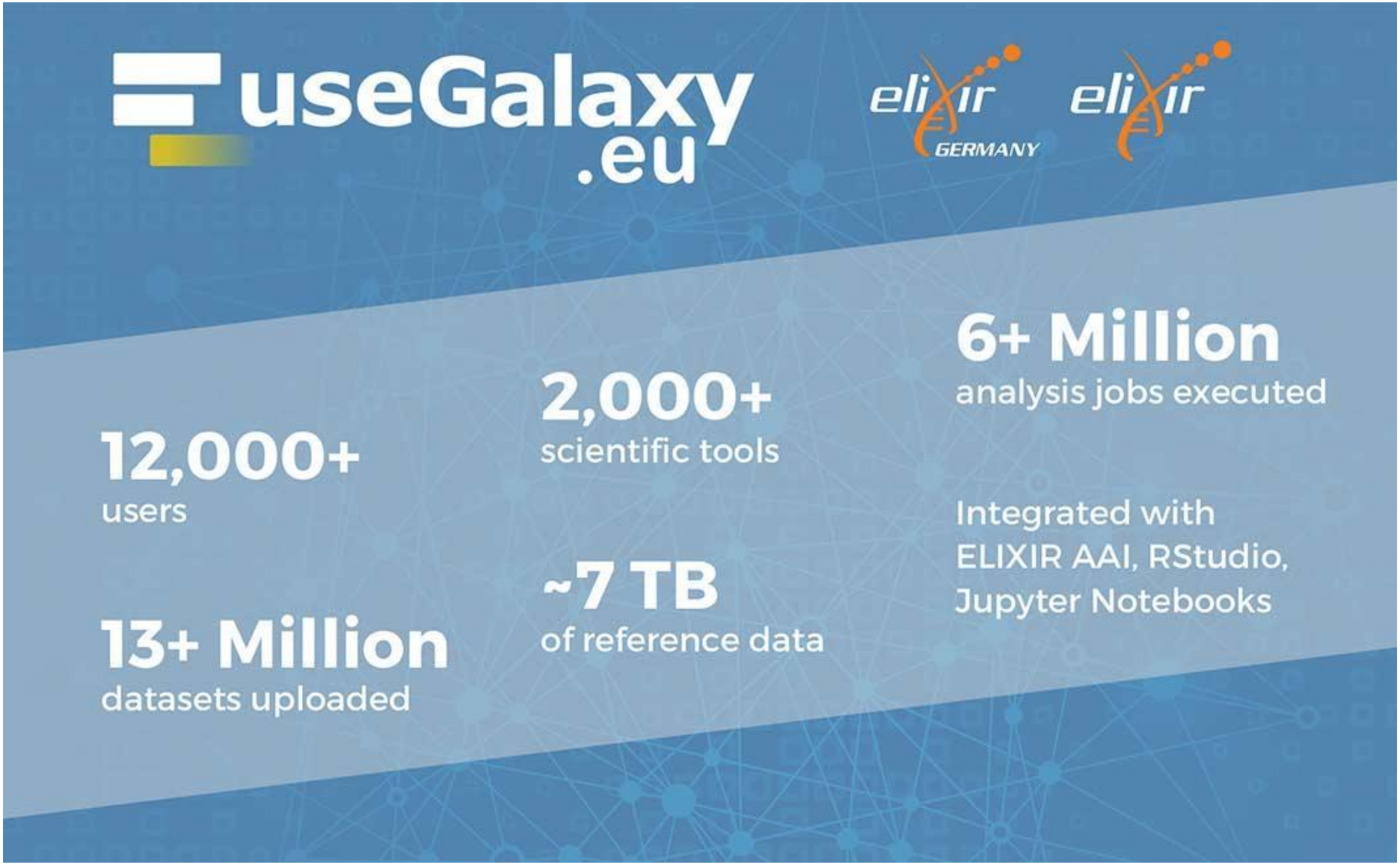
CWL Execution Engine

CWLTool, CWL Exec

Metagenomic Assemblies

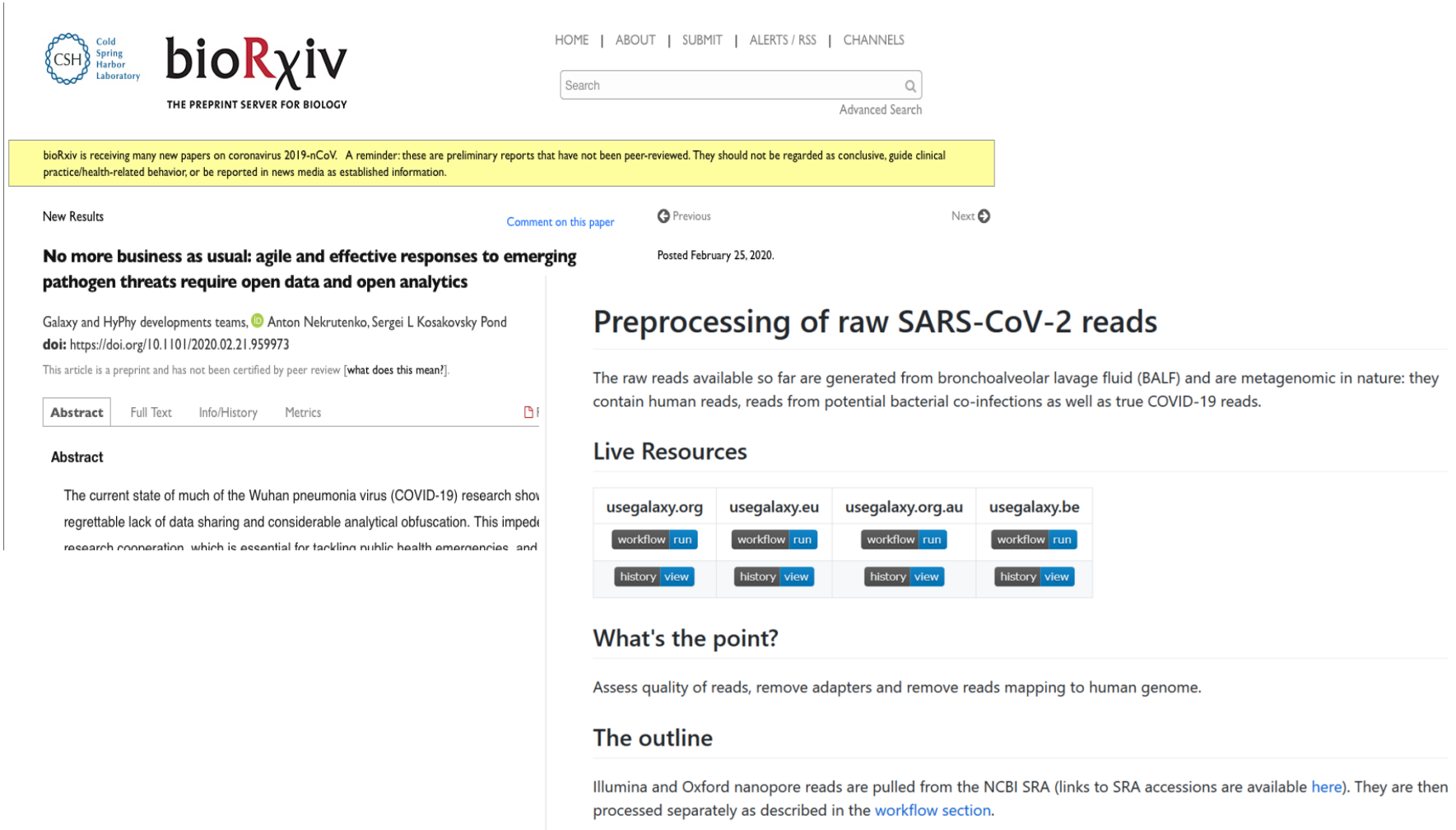
Thanks to Rob Finn, Mgnify, EBI

Designing and building reusable libraries



Disruptor: Yes – workflow environments ARE used.
Sustained, Human in the Loop, Scale out, Professionalisation, RWEs

If you make workflow environments usable and professionally curated and managed then they are used and make a difference



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

bioRxiv is receiving many new papers on coronavirus 2019-nCoV. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

Comment on this paper

Previous

Next

No more business as usual: agile and effective responses to emerging pathogen threats require open data and open analytics

Posted February 25, 2020.

Galaxy and HyPhy developments teams, Anton Nekrutenko, Sergei L Kosakovsky Pond

doi: https://doi.org/10.1101/2020.02.21.959973

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract | Full Text | Info/History | Metrics

Abstract

The current state of much of the Wuhan pneumonia virus (COVID-19) research shov
regrettable lack of data sharing and considerable analytical obfuscation. This impede
research cooperation, which is essential for tackling public health emergencies, and

Preprocessing of raw SARS-CoV-2 reads

The raw reads available so far are generated from bronchoalveolar lavage fluid (BALF) and are metagenomic in nature: they contain human reads, reads from potential bacterial co-infections as well as true COVID-19 reads.

Live Resources

usegalaxy.org	usegalaxy.eu	usegalaxy.org.au	usegalaxy.be
workflow run	workflow run	workflow run	workflow run
history view	history view	history view	history view

What's the point?

Assess quality of reads, remove adapters and remove reads mapping to human genome.

The outline

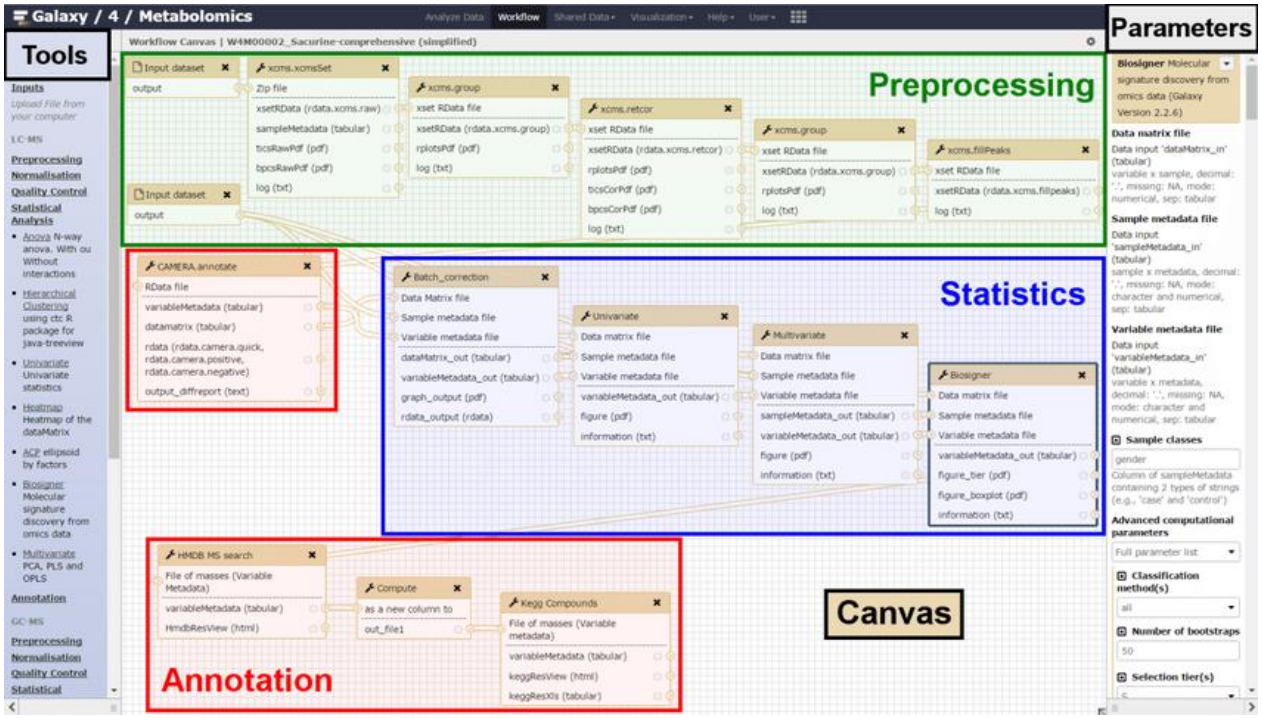
Illumina and Oxford nanopore reads are pulled from the NCBI SRA (links to SRA accessions are available here). They are then processed separately as described in the workflow section.

https://github.com/galaxyproject/SARS-CoV-2/

Clustering around platforms for economies of scale, skill and sustainability

EOSC-*Life*

Workflow collections. Workflow as a Service

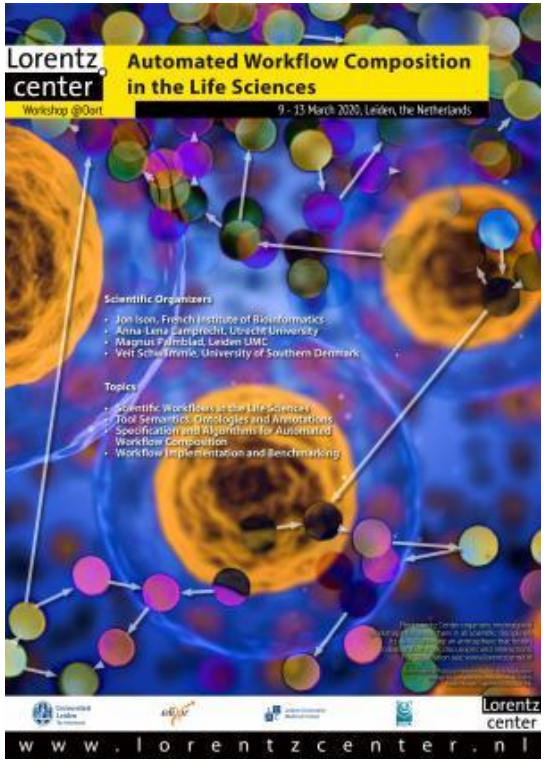


- rna.usegalaxy.eu
- clipseq.usegalaxy.eu
- metagenomics.usegalaxy.eu
- hicexplorer.usegalaxy.eu
- cheminformatics.usegalaxy.eu
- proteomics.usegalaxy.eu
- Imaging.usegalaxy.eu
- metabolomics.usegalaxy.eu
- ecology.usegalaxy.eu
- nanopore.usegalaxy.eu
- singlecellomics.usegalaxy.eu
- humancellatlas.usegalaxy.eu
- streetscience.usegalaxy.eu

Galaxy interactive workflow collection for Metabolomics

Workflow making is still largely artisanal

Producing the **FAIR ecosystem** so we can semi-automatically build workflows that can help automate beyond "heroic" one-offs



- AI's role in building/running the ecosystem
- Recommend workflows, matchmaking….
- Automated discovery of workflows / data
- The capacity and capability of researchers to use and reuse workflows
- Professionalisation of workflow making/mgt

Example: Galaxy Workflow Platform

- Galaxy-ML (https://github.com/goeckslab/Galaxy-ML)
- Tool recommender system in Galaxy using deep learning (https://doi.org/10.1101/838599)

- AI steps in the workflows

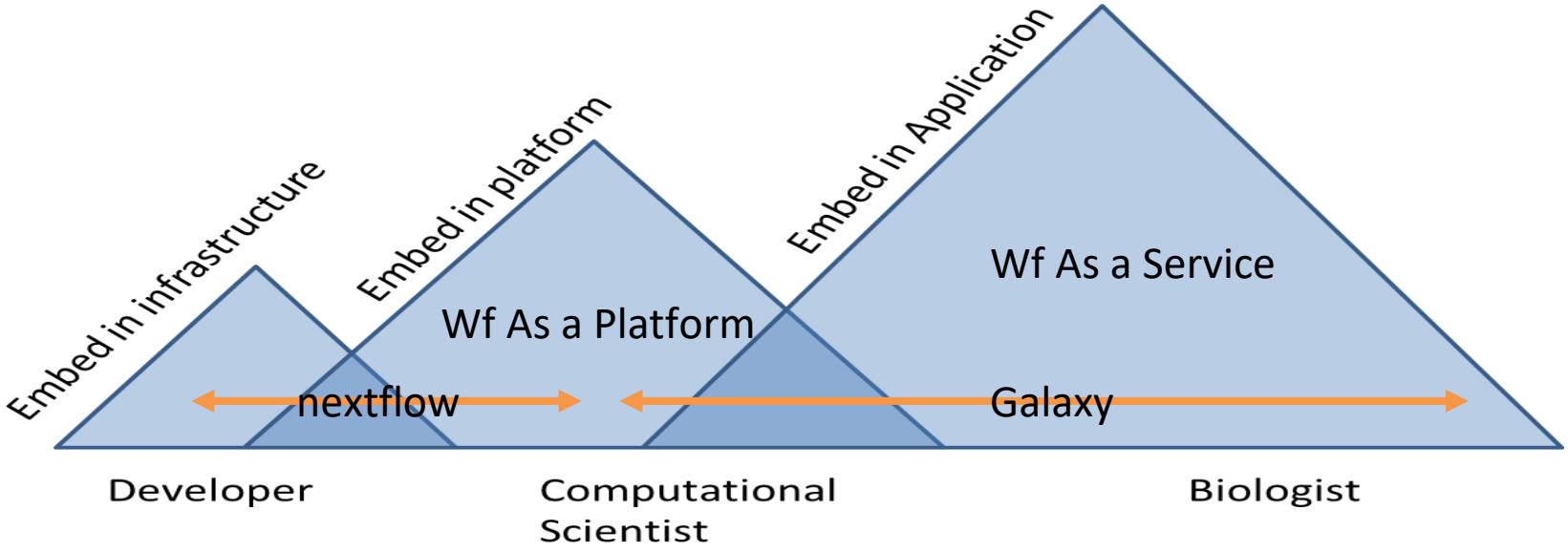
- AI enabled wf design
- AI enabled wf set up
- AI enabled wf execution
- AU enabled wf analytics
- AI enabled wf workflow ecosystem

FAIR Workflows

Citation: C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M.R. Crusoe, K. Peters & D. Schober. FAIR computational workflows. Data Intelligence 2(2020), 108–121. doi: 10.1162/dint_a_000

Who are the Workflow Users?

Capacity Building, Professionalisation, Economics



different systems for different users, spanning the spectrum, enabling researchers to across the spectrum

Trajectories & Challenges: FAIR Workflows

- **A slew of workflow systems**
  - Domain specific languages are on the rise, Script-like workflows & problem-specific ad-hoc data models + Adoption of industry standards.
  - Take up dependent on the "plugged-in" availability of data type specific codes, optimised processing and everyone else using it
- **At the same time take-up of professionally managed cloud platforms**
  - Professionalisation of Workflow Engineering
- **Catalogues of workflow components (and data and SOPs)**
  - Designed to work together, Quality Assured, curated and maintained
  - tools / sub workflows, and the data they operate over accessible and licensed.
- **Interoperability: neat and tidy interfaces**
  - Shared (standardised) data formats, identifier schemes and interfaces for tool compatibility, incl. ids for workflows
- **Just enough metadata on everything**
  - Semantic representations of Workflows, Data and Tasks: schema.org, JSON-LD
- **Portable / accessible execution – containers**
  - tools can execute next to each other independent of WfMS or cluster installation

Acknowledgements

Ian Cottam
Frederik Coppens
Björn Grüning
Mark Wilkinson
Stian Soiland-Reyes
Nick Juty
Rob Finn
Michel Dumontier
Michael Crusoe
Finn Bacall
Alan Williams
Katy Wolstencroft
Pinar Alper
Sarah Cohen-Boulakia
Stuart Owen
Jiten Bhagat
Don Cruickshank
David De Roure
Niall Beard
Aleks Nenandic
Paolo Missier
Marco Roos
Khalid Belhajjame
Sarah Cohen-Boulakia
Yolanda Gil
Ilkay Altintas
Daniel Garijo
Melchior Du-Lac

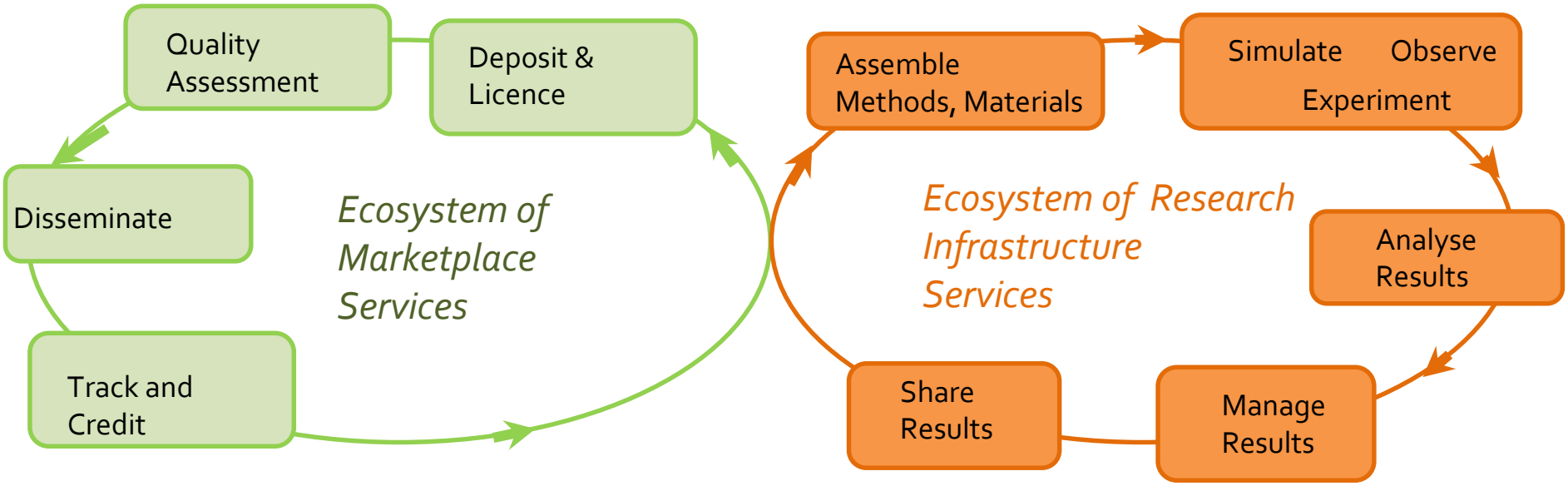
And all the projects and many more

BONUS SLIDES

What are we accelerating?

- Accelerating workflow creation
- Accelerating workflow use and reuse, sharing and exchange
- Accelerating the workflows themselves

Building a FAIR Research Commons



Fixing this to be an 'in flight" feedback loop

Science 2.0 Repositories: Time for a Change in Scholarly Communication
Assante, Candela, Castelli, Manghi, Pagano DOI: 10.1045/january2015-assante

Mesirov, J. Accessible Reproducible Research Science 327(5964), 415-416 (2010)

For the Panel

- What are your observations on the impact of systems and tools to automate workflows, with respect to scientific discovery?
- What is the role of artificial intelligence (AI) and automation in scientific discovery? We'd like you to consider both AI-driven workflows as well as workflows that include AI as part of the scientific discovery process.
- What are the current challenges faced by these systems?
- What do you see as disruptors that will affect how workflow tools are being used by the scientific community?
- What is your vision for AI in workflows in the future of scientific discovery? What research is needed to advance this area?