

Data (Re)-use: Scientific Workflow

Nicola Fiore

LIFEWATCH ERIC, SERVICE CENTRE ICT COORDINATOR

To be Re-usable:

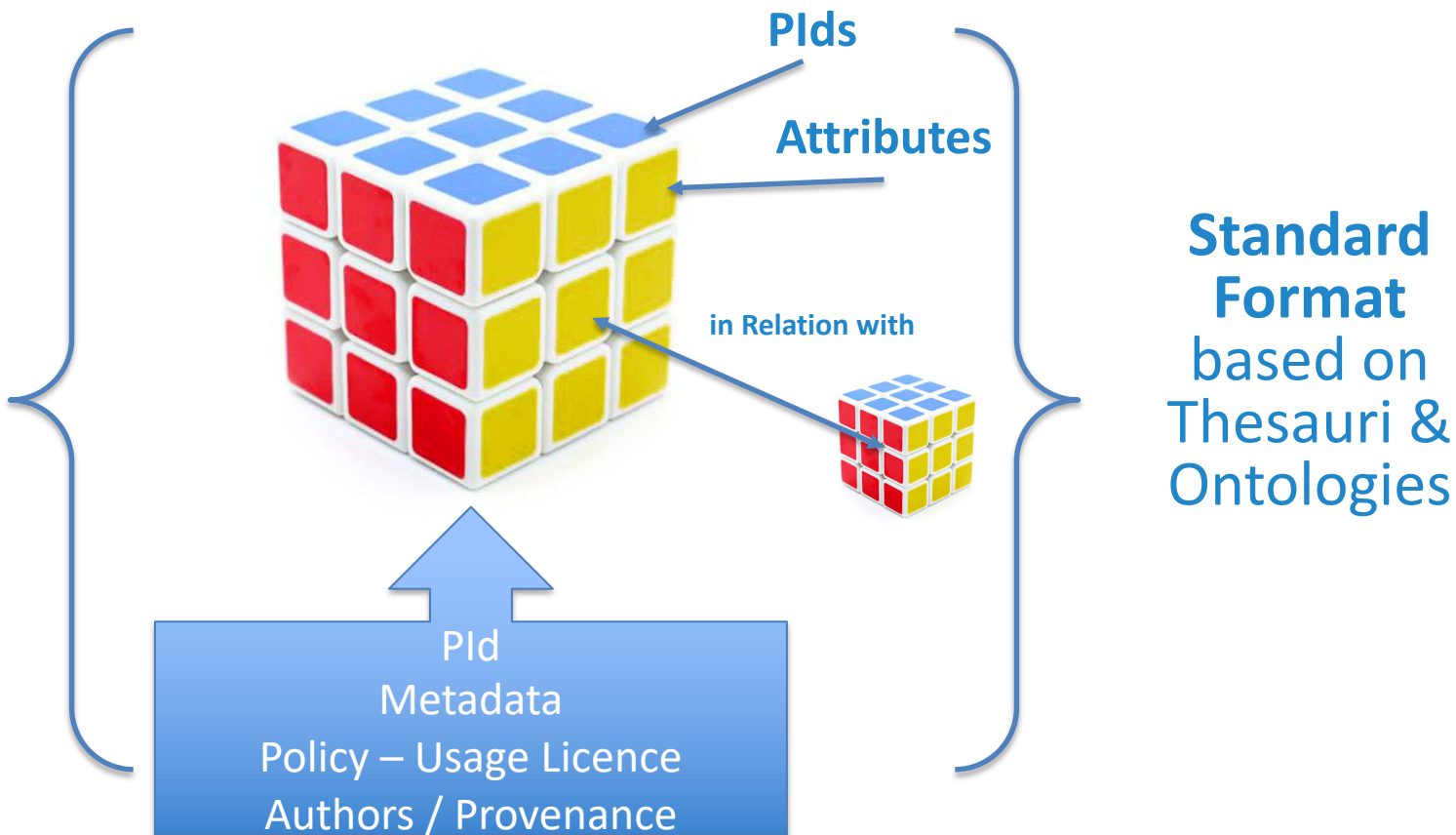
R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

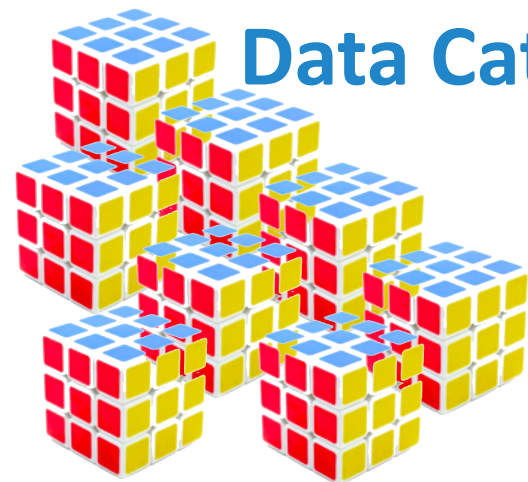
R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.


Digital Object




Data Catalogues



GO TO EUDAT WEBSITE



B2FIND



EUDAT

GUIDELINES

Occurrence records 1.328.352.544	Datasets 45.315	Publishing institutions 1.420	Peer-reviewed papers using data 3.732
--	---------------------------	---	---

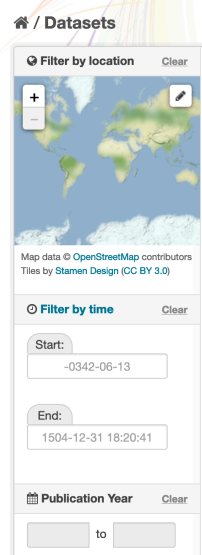
3,271 dataset: "phytoplankton"


Phytoplankton popul

Phytoplankton population dynamics - Abundances


Particle sinking velocity measurements: phytoplankton

Viability of phytoplankton and photoprotection







BIFA funds nine Asian data mobilization projects
26 June 2019



Minimizing biodiversity loss in the Brazilian Cerrado
2 July 2019

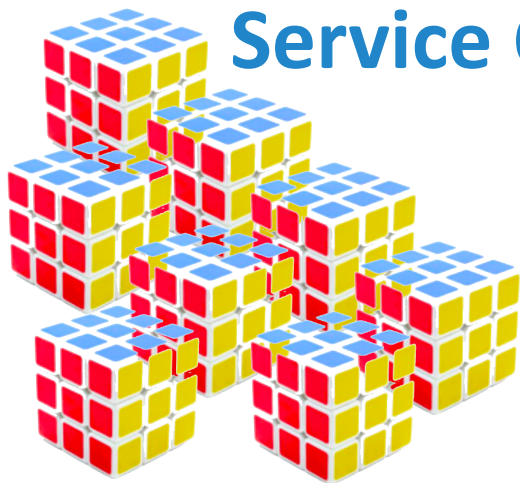


2019 GBIF Ebbe Nielsen Challenge seeks open-data innovations for biodiversity
Deadline: 1 August 2019



Data mobilization and capacity building essential to address global biodiversity crisis
6 May 2019

Service Catalogues




Home About Us Resources & Services Training & Educat

Resources & Services / Catalogue of Services

Catalogue of Services

Title	Web	Details
Zooplankton Traits Thesaurus	Thesaurus on zooplankton...	
Taxonomic Rarefaction	Evaluation of the estimate...	
Phytoplankton Traits Thesaurus	Thesaurus on Phytoplankt...	
Phytoplankton Size Distributions	Phytoplankton Size Distrib...	
Niche Filtering	Suite of logistic and quanti...	
LW Toponyms IGM	Geo-referencing is one of t...	
ISS Phyto	Calculation of phytoplankt...	
ISS Benthos	Calculation of macrozoobe...	
Fish Traits Thesaurus	Thesaurus on Fish morpho...	
Endemisms Thesaurus	Thesaurus on endemisms	
Data Services	Data Services provide the ...	
Biodiversity DataManag	Biodiversity partitioning a...	

Scientific workflows allow users to easily express multi-step computational tasks.

Goals: automate a scientist's repetitive data management and analysis tasks

Typical Phases:

Data access, scheduling, generation, transformation, aggregation, analysis, visualization

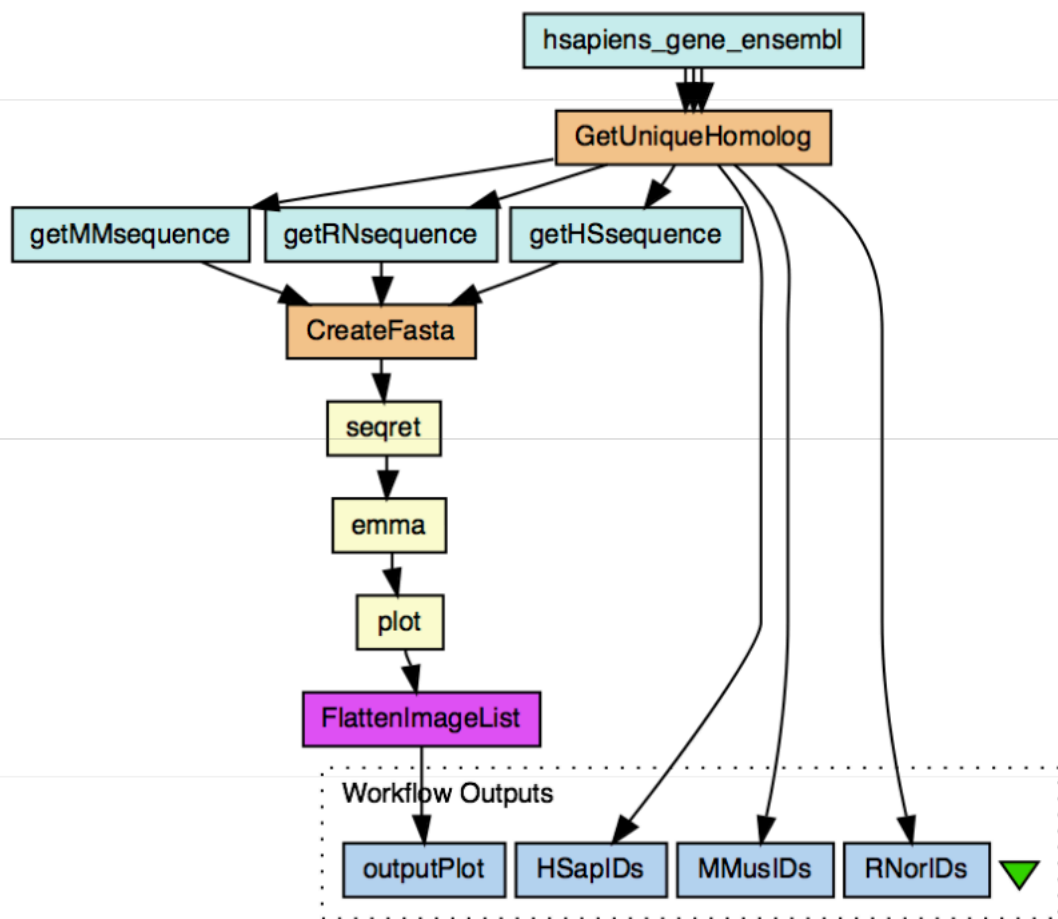
Design, test, share, deploy, execute, reuse SWF's

for example retrieve data from a catalogue or an instrument, reformat the data, and run an analysis.

Scientific workflows are often visually represented as directed graphs linking atomic tasks or composite components, so-called **subworkflows**.

Tasks can include native functions of the workflow system, but often correspond to invocations of **local applications**, **remote (web)services**, or **subworkflows**.

There is currently no standard scientific workflow language



Example workflow represented in the Taverna workflow system. This workflow extracts gene IDs from human chromosome 22 with mappings to disease functions and homologues in mouse and rat; fetches base pairs of the associated DNA sequences; combines the sequences into a FASTA file; performs a multiple sequence alignment; and renders the result. The workflow uses three soaplab-based analysis operations (seqret, emma, plot) that run on the EBI compute cluster.

Scientific workflow systems (SWFSs) – User Requirements

- Design tools – especially for non-expert users
Need to look into how scientists define processes
- Ease of use – fairly simple user interface having more complex features hidden in background
- Reusable generic features
- Generic enough to serve different communities but specific enough to serve one domain
- Extensibility for the expert user – almost a visual programming interface
- Registration and publication of data products and “process products” (workflows); provenance

Scientific workflow systems (SWFSs) – Technical Requirements

- Error detection and recovery from failure
- Logging information for each workflow
- Allow data-intensive and compute-intensive tasks (maybe at the same time)
- Data management/integration
- Allow status checks and on the fly updates
- Visualization
- Semantics and metadata based dataset access
- Certification, trust, security

Scientific workflow systems (SWFSs) – Why a GUI

- No need to learn a programming language
- Visual representation of what workflow does
- Allows you to monitor workflow execution
- Enables user interaction
- Facilitates sharing workflows

Scientific workflow systems (SWFSs) – TOOL

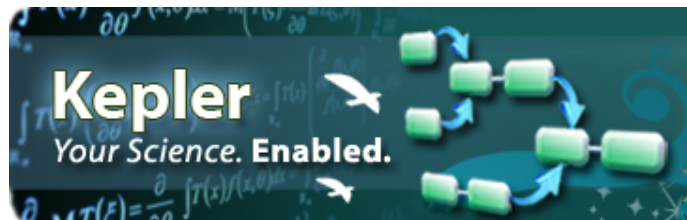


Taverna Workflow System

<https://taverna.incubator.apache.org/>

 Galaxy
COMMUNITY HUB

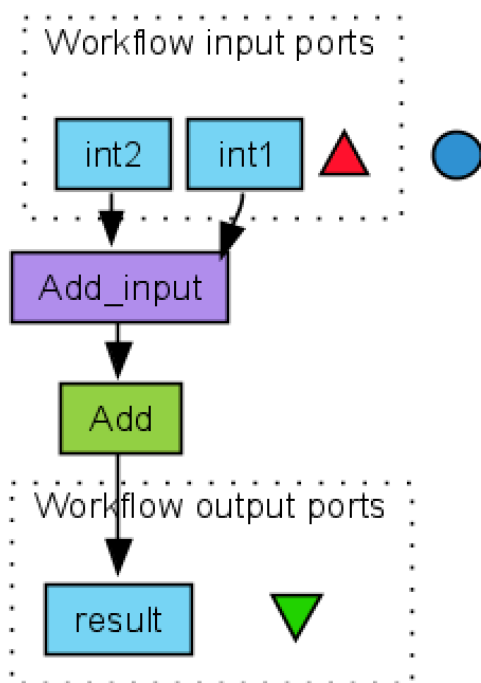
<https://galaxyproject.org/>



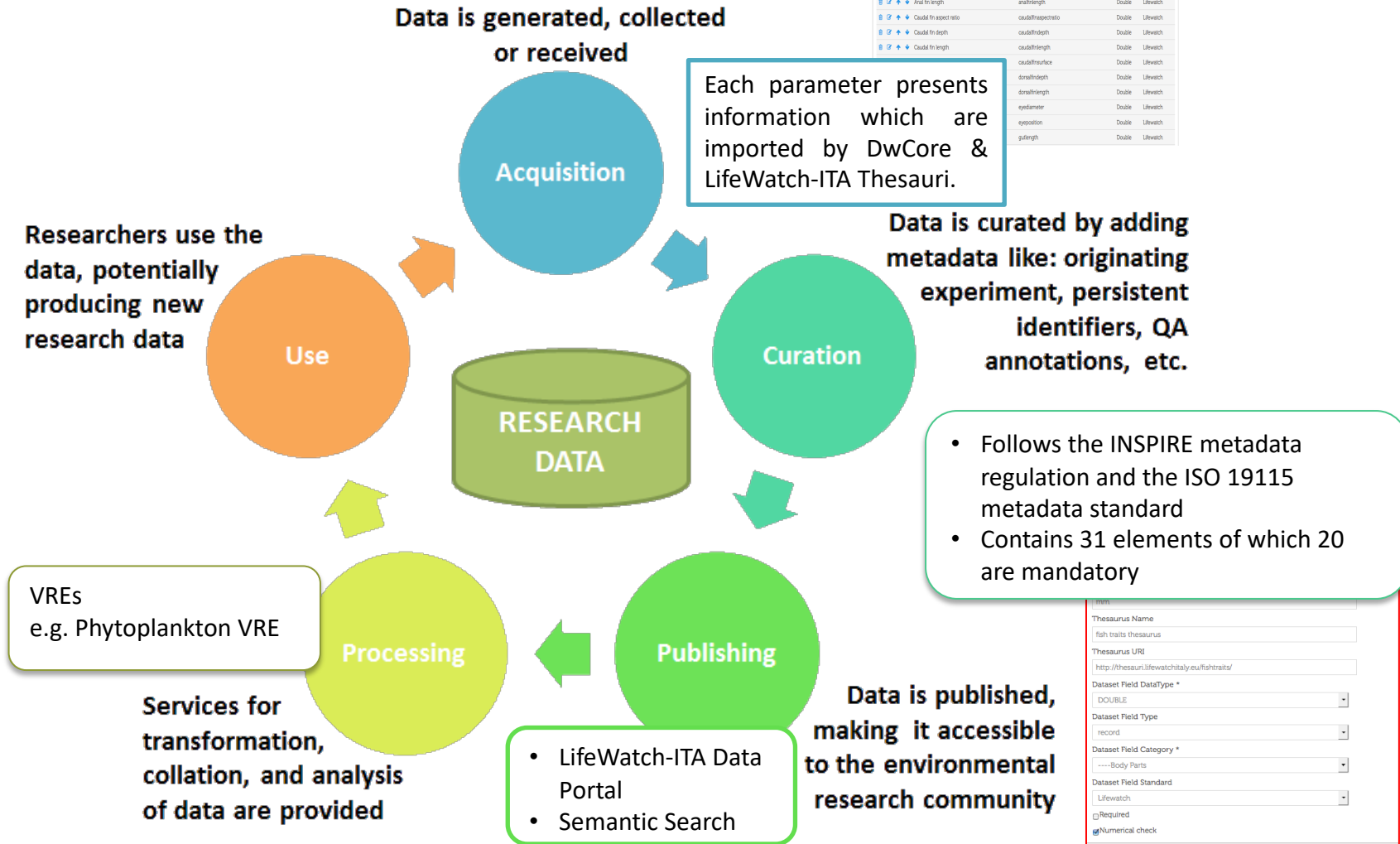
<https://kepler-project.org/>

Scientific workflow systems (SWFSs) – TAVERNA WF EXAMPLE

WSDL TEST URL: <http://www.dneonline.com/calculator.asmx?wsdl>



Name	Unique Name	Type	Standard
Anal fin depth	analfindepth	Double	Lifewatch
Anal fin length	analfinlength	Double	Lifewatch
Caudal fin aspect ratio	caudalfinaspectratio	Double	Lifewatch
Caudal fin depth	caudalfindepth	Double	Lifewatch
Caudal fin length	caudalfinlength	Double	Lifewatch
Caudal fin surface	caudalfinsurface	Double	Lifewatch
Dorsal fin depth	dorsalfindepth	Double	Lifewatch
Dorsal fin length	dorsalfinlength	Double	Lifewatch
eyediameter	eyediameter	Double	Lifewatch
eyeposition	eyeposition	Double	Lifewatch
gulf length	gulf length	Double	Lifewatch



Thesaurus Name
fish traits thesaurus

Thesaurus URI
<http://thesauri.lifewatchitaly.eu/fishtraits/>

Dataset Field DataType *
DOUBLE

Dataset Field Type
record

Dataset Field Category *
---Body Parts

Dataset Field Standard
Lifewatch

Required

Numerical check



Nicola Fiore

LIFEWATCH ERIC

e-mail: nicola.fiore@lifewatch.eu