Data (Re)-use: Scientific Workflow

Nicola Fiore

LIFEWATCH ERIC, SERVICE CENTRE ICT COORDINATOR

DATA FAIRNESS

FAIR

To be Re-usable:

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

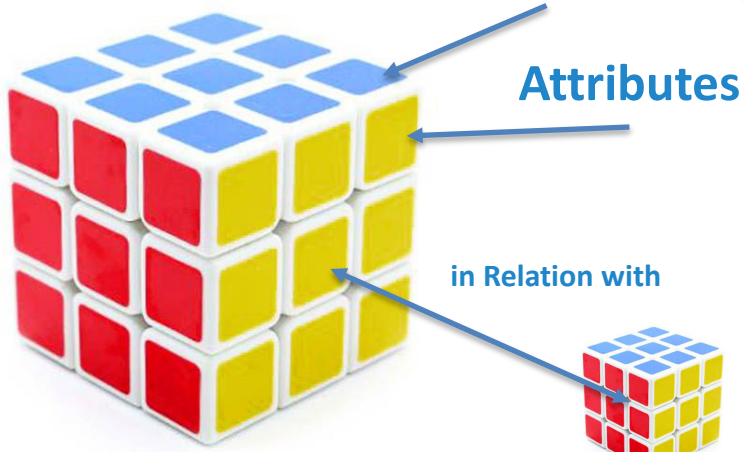
R1.1. (meta)data are released with a clear and accessible data usage license.

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

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

Re-Using What?

Digital Object

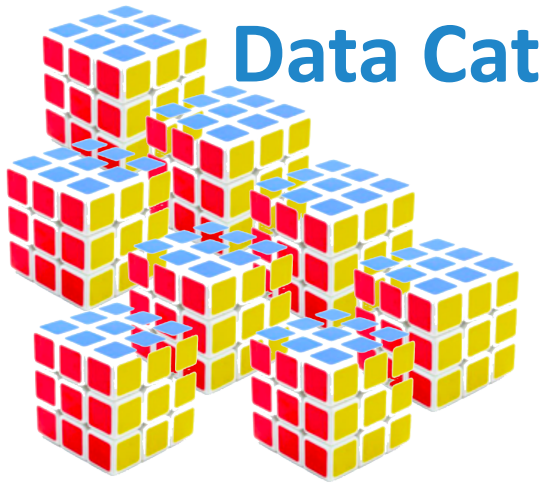


PId
Metadata
Policy – Usage Licence
Authors / Provenance

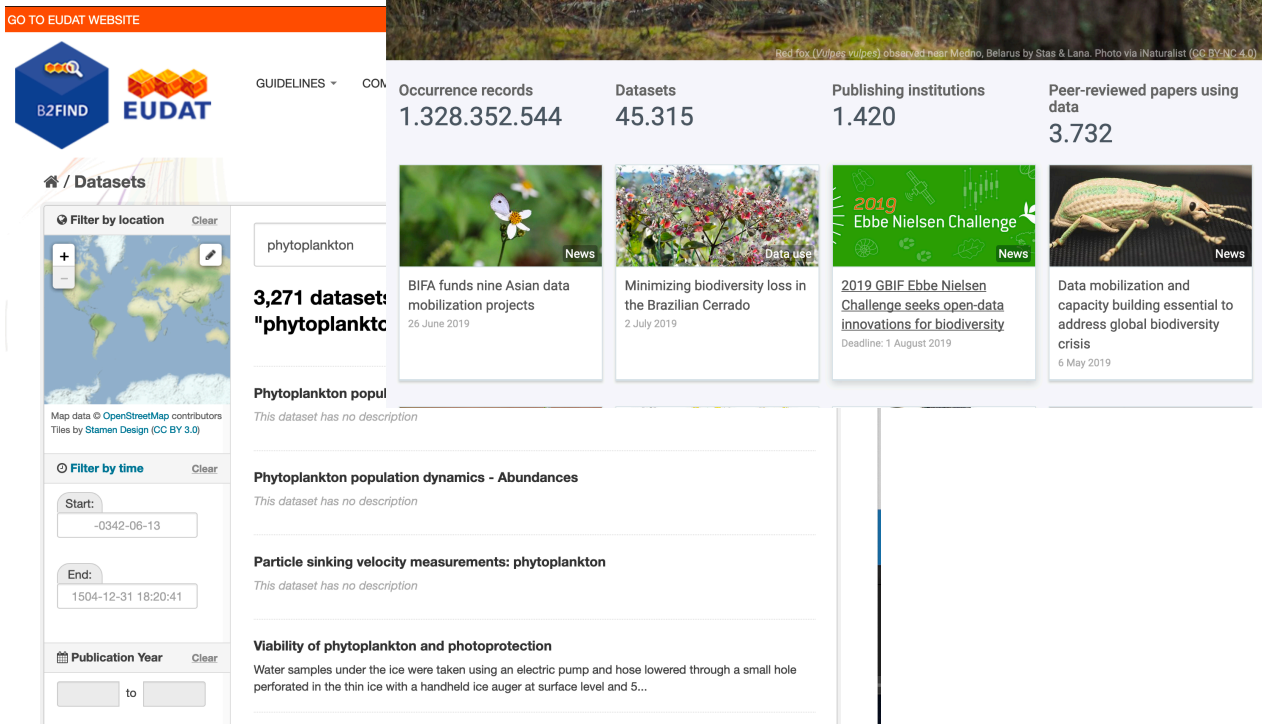
Standard Format based on Thesauri & Ontologies

DATA FAIRNESS

Re-Using How?



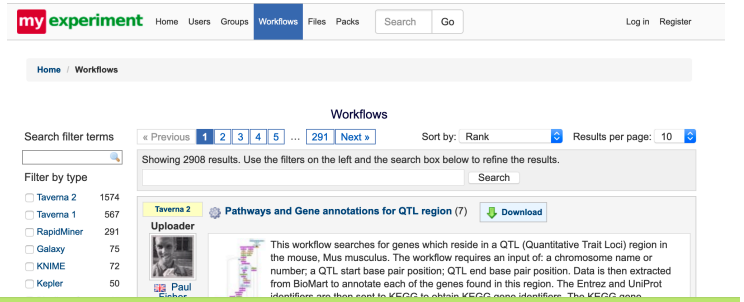
Data Catalogues

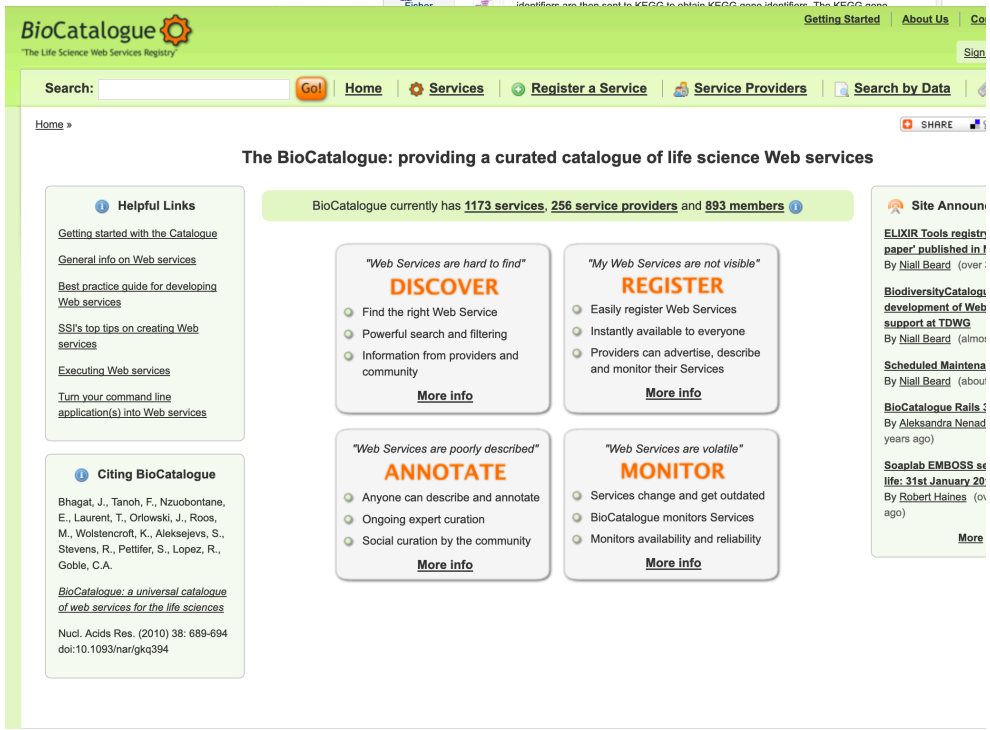


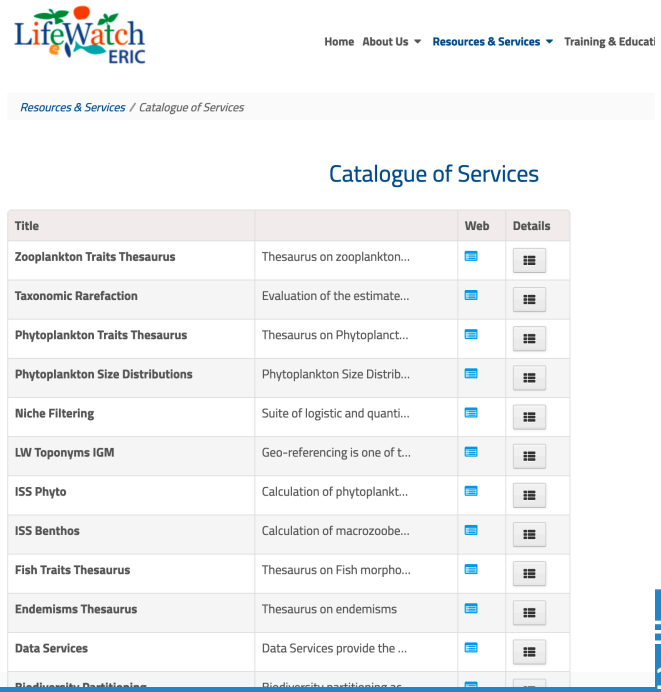
DATA FAIRNESS

Re-Using How?

Service Catalogues







Scientific Workflow

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.

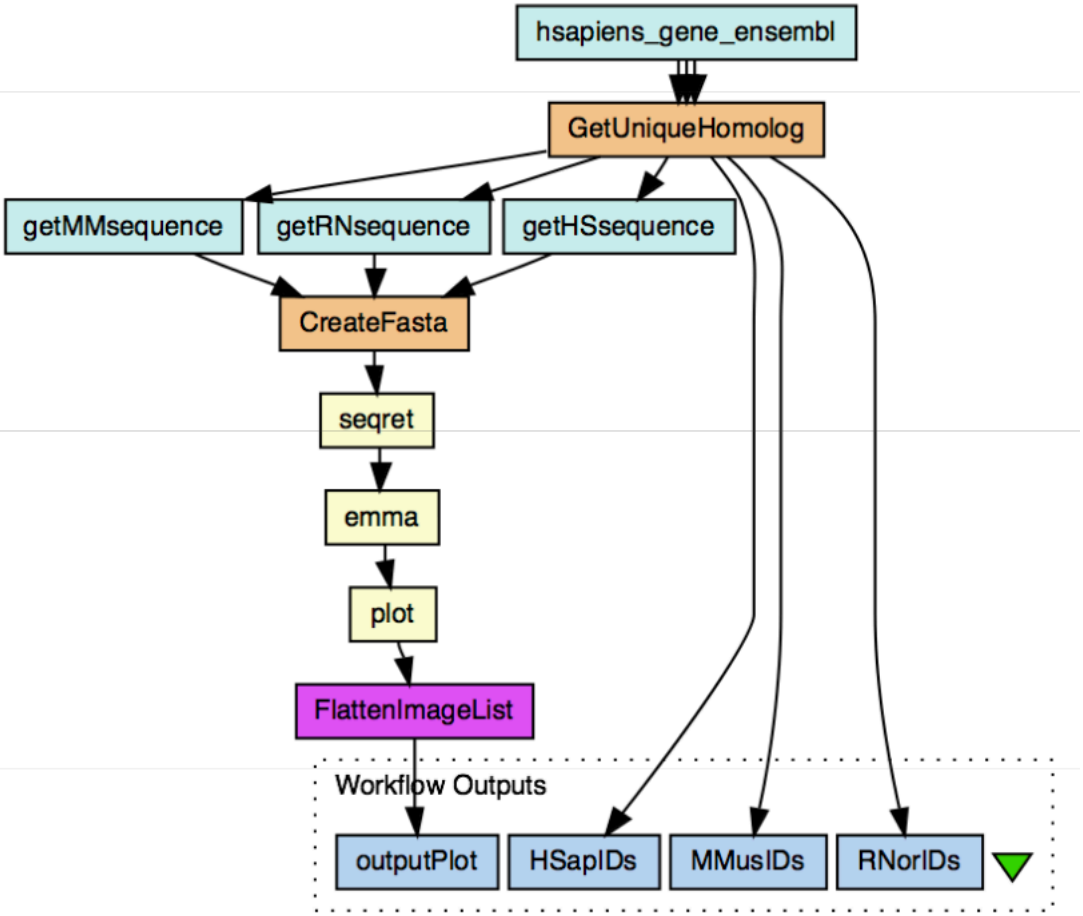
Scientific Workflow

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

Scientific Workflow



Example workflow represented in the Taverna workflow system. This workflow extractsgene IDs from human chromosome 22 with mappings to disease functions and homologues in mouseand rat; fetches base pairs of the associated DNA sequences; combines the sequences into a FASTAfile; performs a multiple sequence alignment; and renders the result. The workflow uses threesoaplab-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

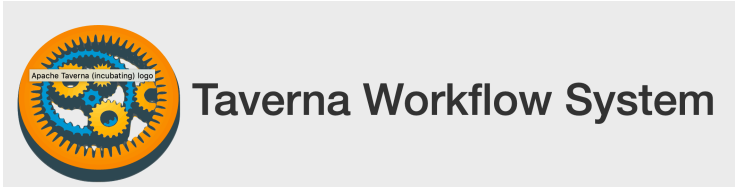
Scientific workflow systems (SWFSs) – Technical Requirements

- 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

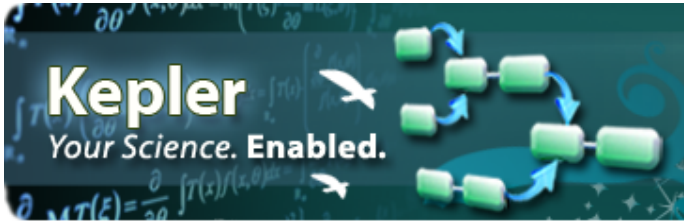
- No need to learn a programming language
- Visual representation of what workflow does
- Allows you to monitor workflow execution
- Enables user interaction
- Facilitates sharing workflows

Scientific workflow systems (SWFSs) – TOOL



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

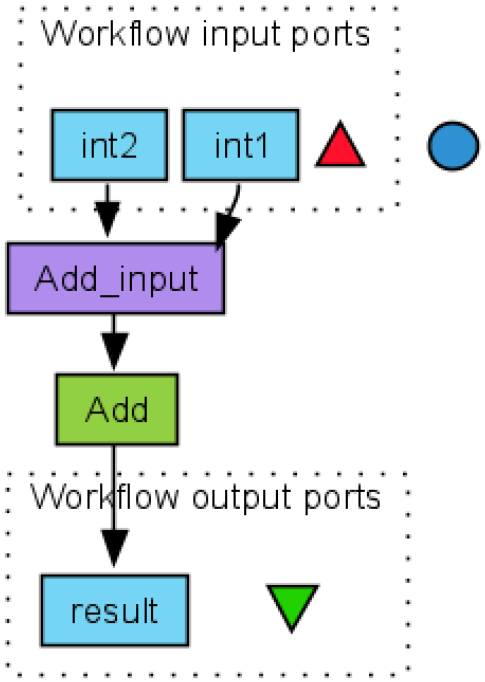
https://galaxyproject.org/



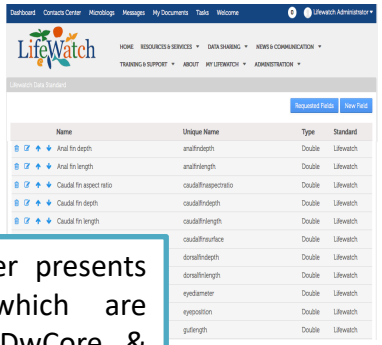
https://kepler-project.org/

Scientific workflow systems (SWFSs) – TAVERNA WF EXAMPLE

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



Data Life Cycle



Data is generated, collected or received

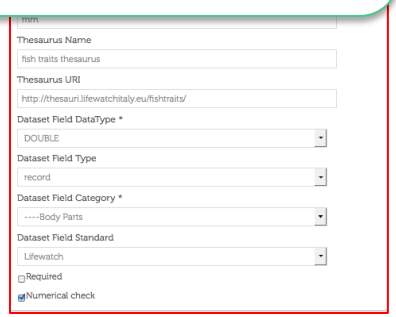
Each parameter presents information which are imported by DwCore & LifeWatch-ITA Thesauri.

Acquisition

Data is curated by adding metadata like: originating experiment, persistent identifiers, QA annotations, etc.

Curation

- Follows the INSPIRE metadata regulation and the ISO 19115 metadata standard
- Contains 31 elements of which 20 are mandatory



Publishing

Data is published, making it accessible to the environmental research community

- LifeWatch-ITA Data Portal
- Semantic Search

Processing

Services for transformation, collation, and analysis of data are provided

VREs
e.g. Phytoplankton VRE

Use

Researchers use the data, potentially producing new research data

RESEARCH DATA

Nicola Fiore

LIFEWATCH ERIC

e-mail: nicola.fiore@lifewatch.eu