Supporting Data Quality Assessment in eScience

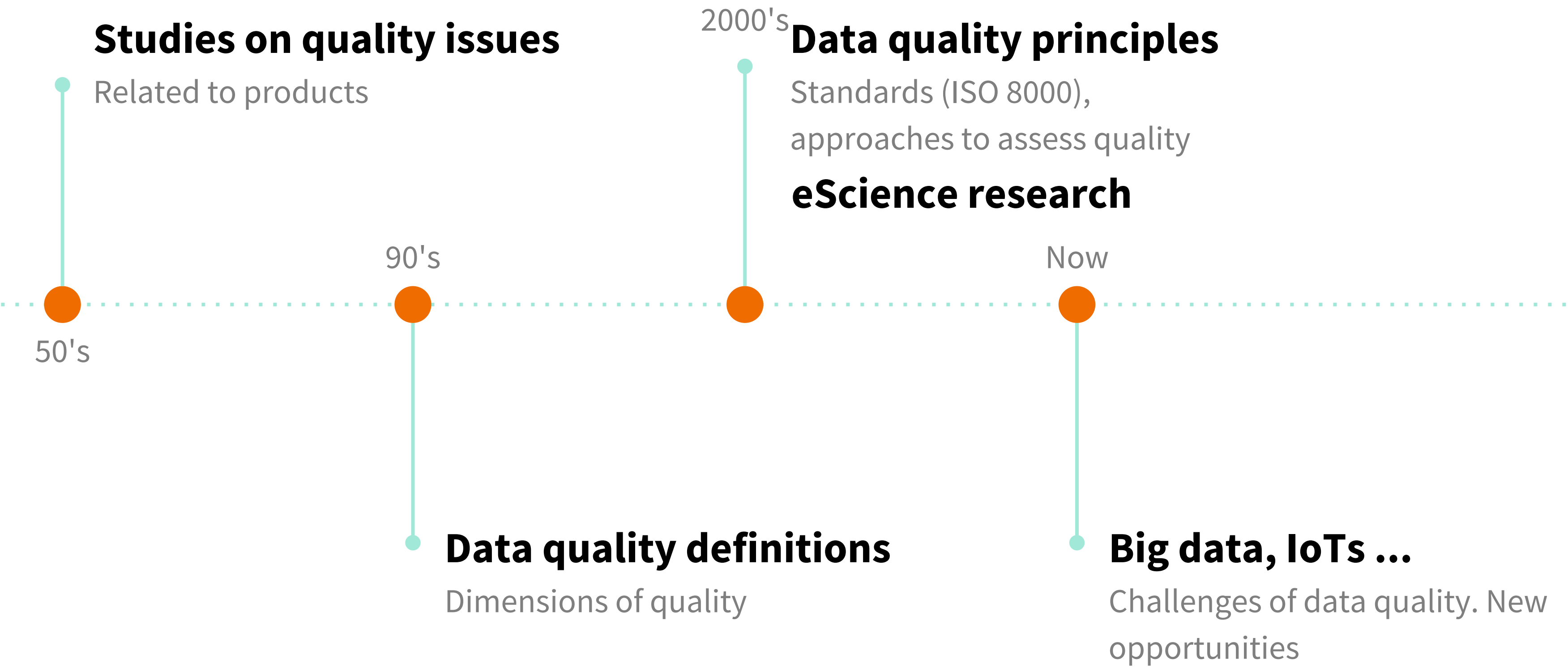
Postdoctoral fellow: Joana Gonzales Malaverri
Supervisor: Prof. Claudia Bauzer Medeiros

IC - Unicamp

Outline

- Foundations
- Our research
  - ProvenFrame
  - Quality Flow
  - w2Share
- Conclusions
- Ongoing work

Data quality research



50's

Studies on quality issues

Related to products

90's

Data quality definitions

Dimensions of quality

2000's

Data quality principles

Standards (ISO 8000), approaches to assess quality

eScience research

Now

Big data, IoTs ...

Challenges of data quality. New opportunities

Problems

- Different formats, standards and scales.
  - data from primary to secondary sources, raw or derived.
  - domain requirements and the intended use of the data.
- Data submitted to different transformation processes.
- Out-of-date data expressed as 'current'.

Issues of quality

- Attributes to represent a particular characteristic of quality.
    - Context-based
- Classification:
    - Qualitatives: reputation, completeness.
    - Quantitatives: consistency, timeliness.
    - Qualitatives & Quantitatives: accuracy.

Reliability
Format
Sufficiency
Flexibility
Accuracy
Conciseness
Timeliness
Currency
Comparability
Scope
Level-of-detail
Precision
Relevance
Completeness
Effeciency
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Consistency
Informativeness
Clarity
Content
Importance
Freedom from bias

Extracted from
https://wq.io/research/quality

Example

Health data

- Temporally and geographically distributed.
- Stored in files (digital or not) and (research) databases.
- Data are fragmented.
- Data in an electronic record should be accurate, up-to-date and complete.
    - clear understanding of the meaning, context and intent of the data.
    - unambiguous and standardized.

Health data quality – a two-edged sword.
Link: https://goo.gl/zTeKxQ

Example

Health data

- Accurate:
  - data or values well reflects the true state of the source information.
  - data or values will not cause ambiguity.
- Up-to-date:
  - data are regularly updated.
  - the time interval from data collection and processing to release meets requirements.

How to assess these (and others) quality dimensions?

Issues of quality

- Different assessment approaches:
  - Attribute-based
    - Manual: user experience, crowdsourcing…
    - Automatic: parsing, validations rules, functions…
  - Provenance-based

Reliability
Format
Sufficiency
Flexibility
Conciseness
Accuracy
Timeliness
Currency
Comparability
Scope
Level-of-detail
Precision
Relevance
Completeness
Effeciency
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Consistency
Informativeness
Clarity
Content
Importance
Freedom from bias

Extracted from
https://wq.io/research/quality

Quality assessment

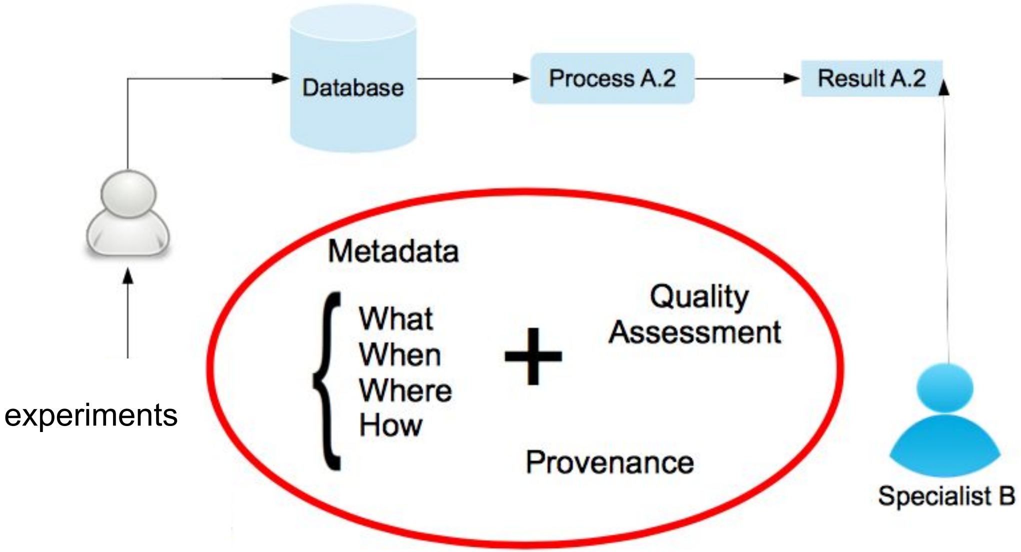
- Dimensions?
  - Timeliness, accuracy, reputation...
- Domain requirements?
- Models?
  - attribute or provenance-based?

Quality assessment

- Dimensions?
  - Timeliness, accuracy, reputation...
- Domain requirements?
- Models?
  - attribute or provenance-based?
- Research:
  - **Provenance** inducing quality assessment

Provenance

Provenance is information about **entities**, **activities**, and **people** involved in **producing** a piece of **data** or **thing**, which can be **used** to form **assessments** about its **quality**, reliability or trustworthiness (W3C).



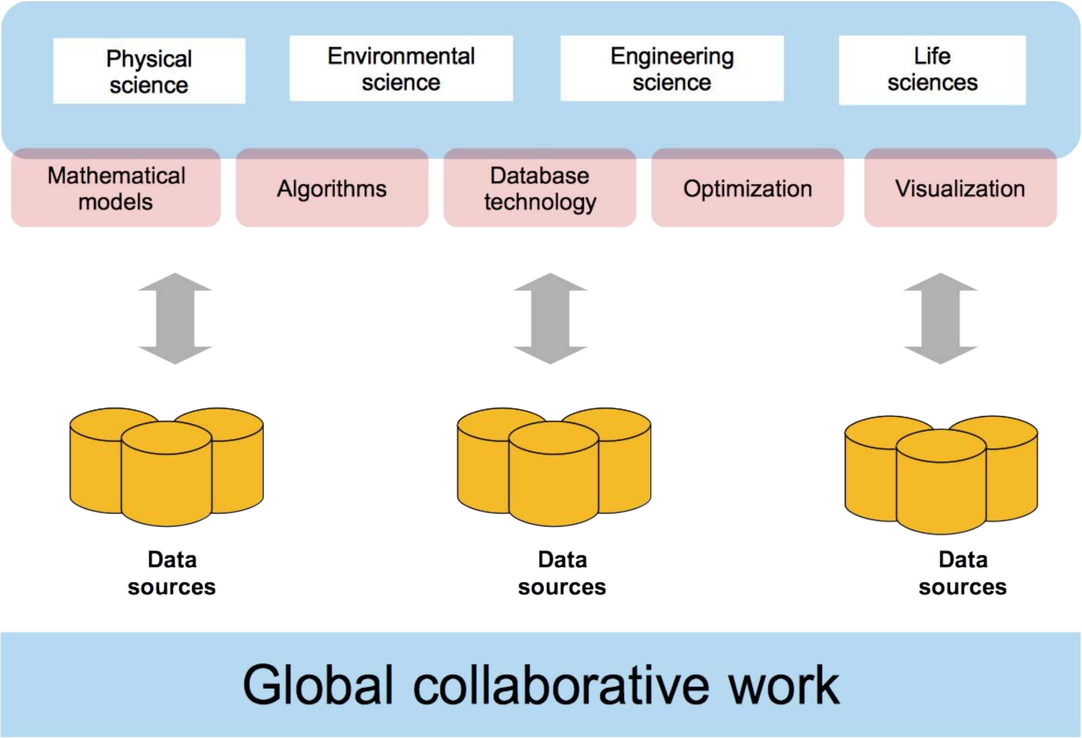
Our research

Supporting Data Quality Assessment in eScience

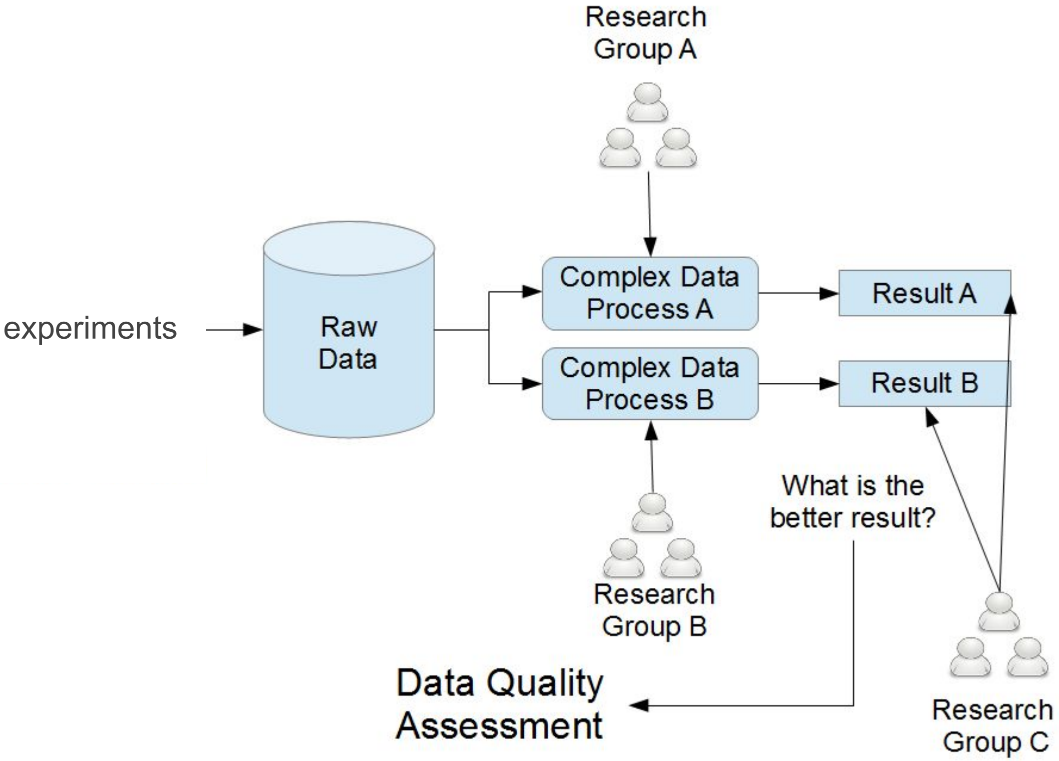
- ProvenFrame
- Quality Flow
- w2Share

What is eScience?

eScience


Physical science
Environmental science
Engineering science
Life sciences
Mathematical models
Algorithms
Database technology
Optimization
Visualization
Data sources
Data sources
Data sources
Global collaborative work

eScience - Example of a typical scenario



eScience - Challenges

- How to lead with data **heterogeneity** issues?
- How to support the **integration** and **sharing** of data?
- How to **evaluate** and **ensure** the **quality** of data?
- How to ensure the **reuse** and **reproducibility** of experiments?

What is scientific data?

- Any (digital) input to experiments.
- Any (digital) result of scientific experiments.

Scientific Workflow Management Systems (SFMSs)?

Taverna
www.taverna.org.uk

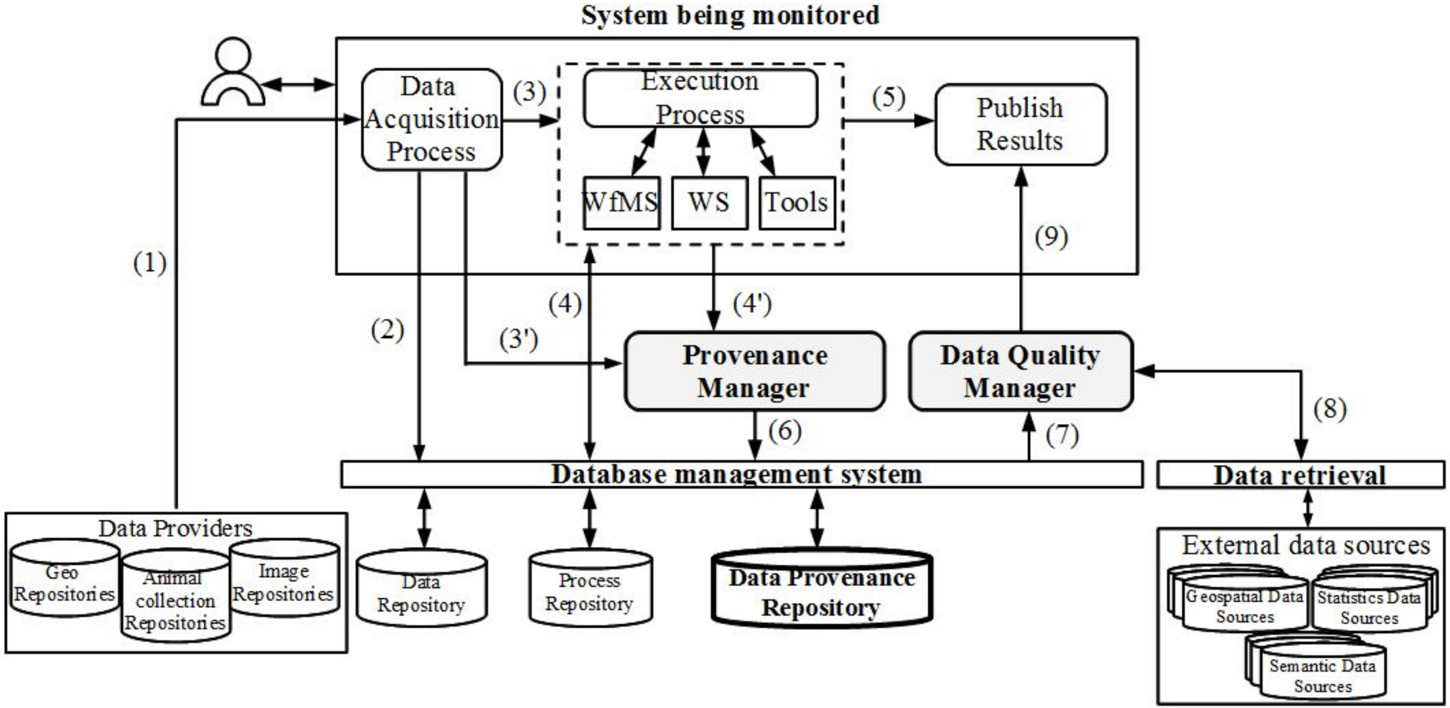
Kepler
https://kepler-project.org

WINGS
www.wings-workflows.org

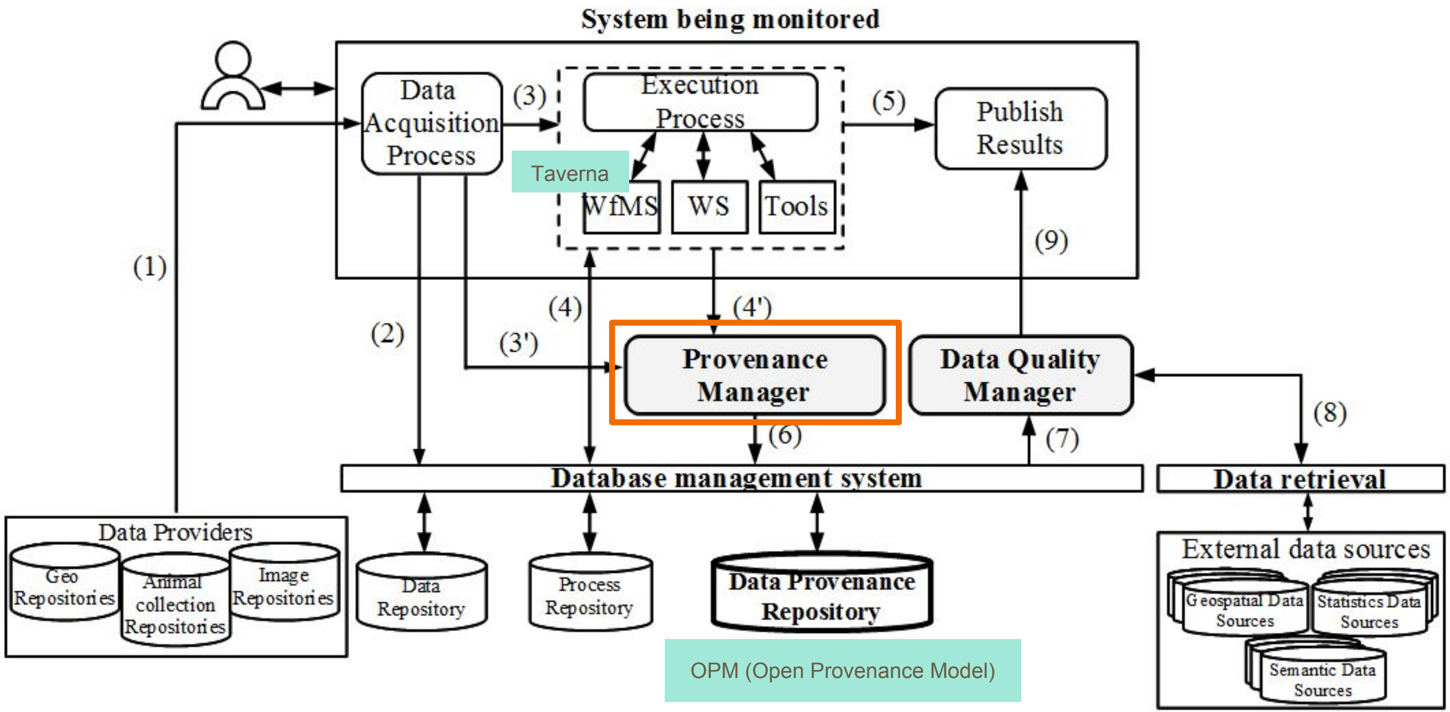
- Workflow
  - a set of inter-dependent steps needed to complete a certain task.
- Scientific workflows
  - specification of design, data capture, integration, processing, and analysis that leads to scientific discovery.
- SWfMS
  - computational tool to model, execute and monitor scientific processes.
  - generation of provenance information of scientific processes.

ProvenFrame

ProvenFrame - architecture



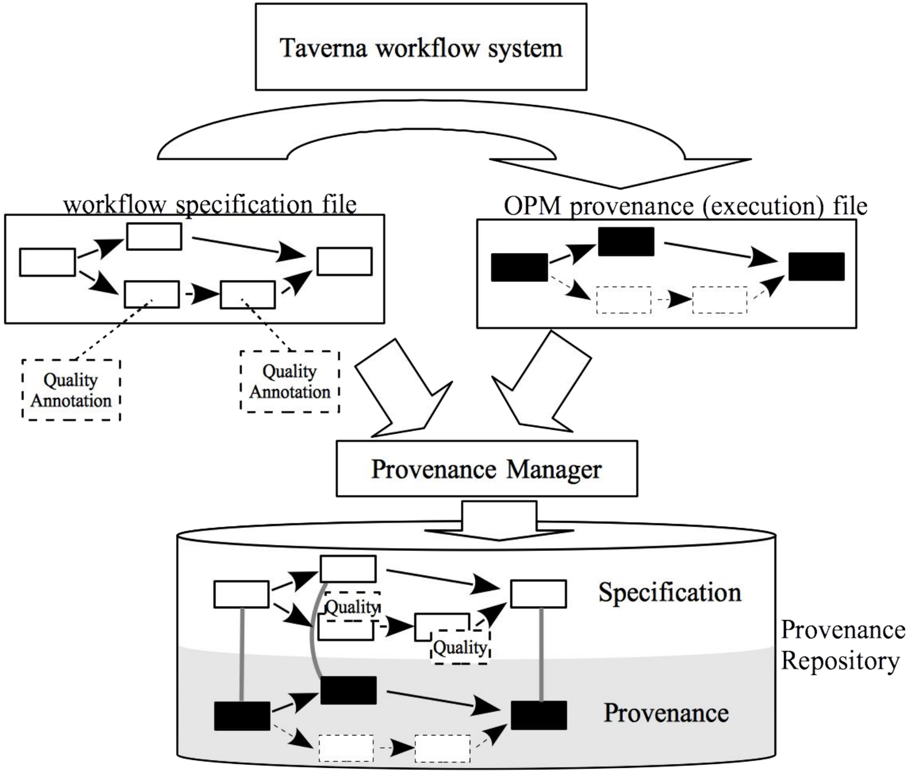
ProvenFrame - architecture



Taverna

OPM (Open Provenance Model)

ProvenFrame interacting with the Taverna WfMS

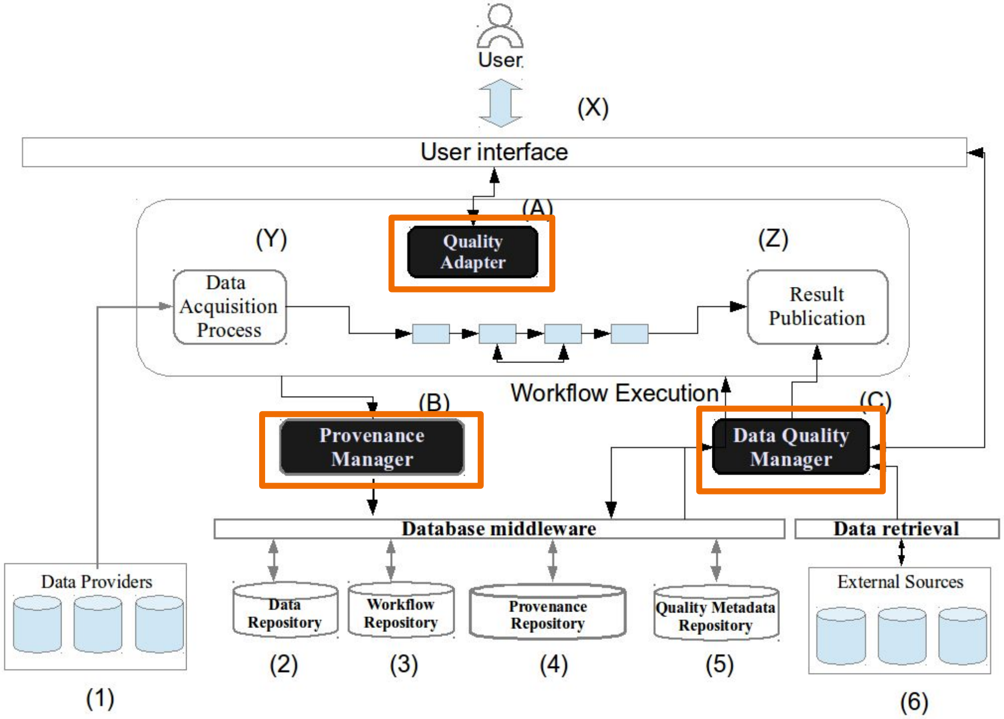


Quality Flow
Sousa, Renato B. Master thesis

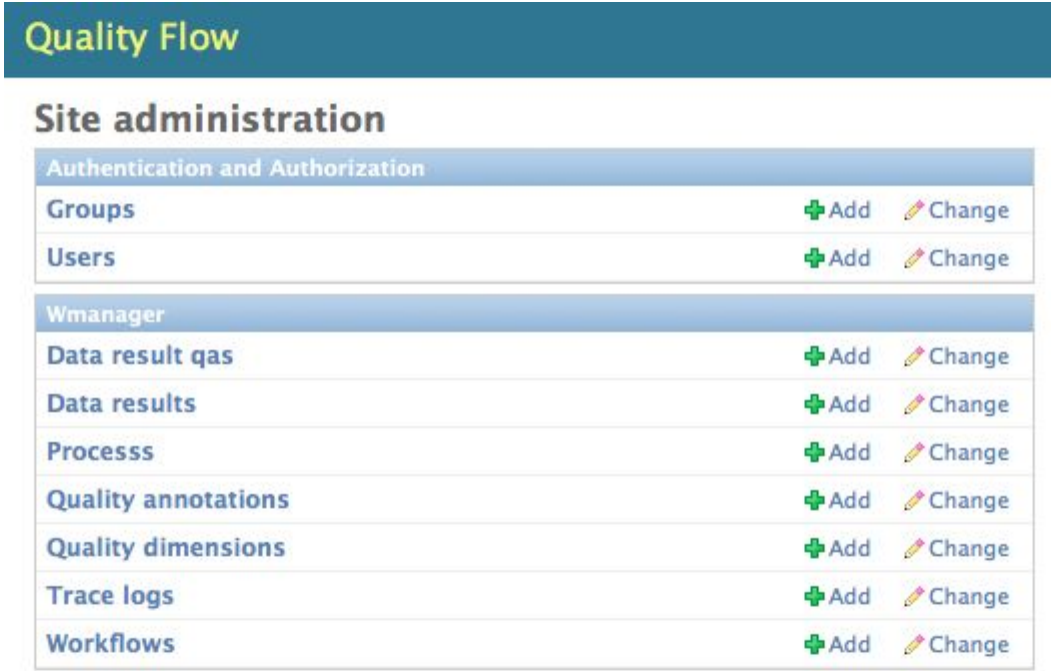
Quality Flow (Sousa, Renato B. Master thesis)

- Improvement and instantiation of ProvenFrame.
- A workflow-based system for data quality assessment of scientific experiments.
- Case study: Long term data preservation and curation.
  - Why? It requires ensuring (meta)data quality.

Quality Flow - architecture



Quality Flow - web prototype



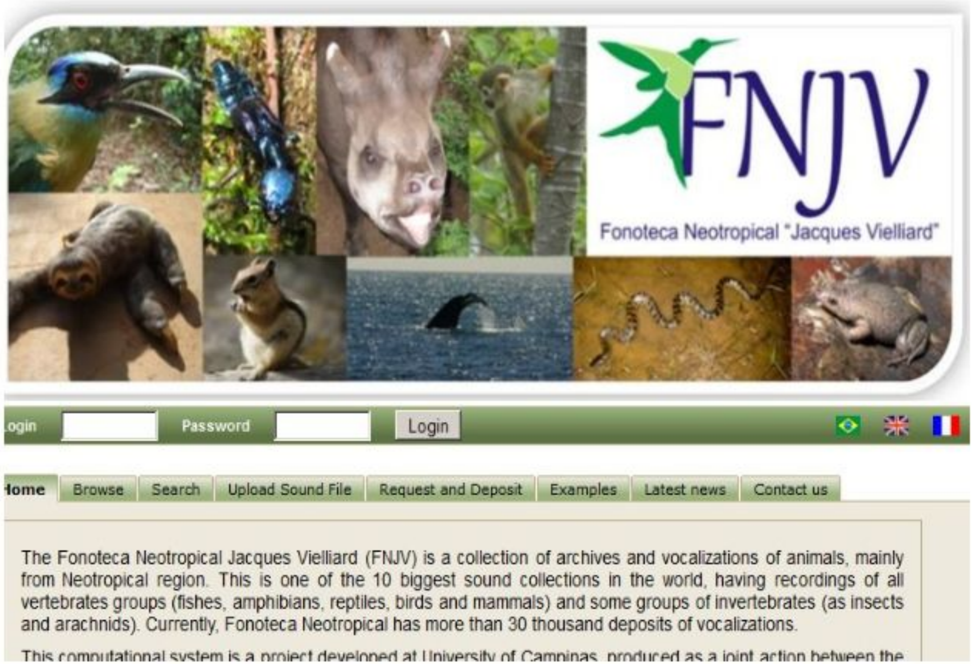
Quality Flow

Site administration

Authentication and Authorization		
Groups	Add	Change
Users	Add	Change

Wmanager		
Data result qas	Add	Change
Data results	Add	Change
Processs	Add	Change
Quality annotations	Add	Change
Quality dimensions	Add	Change
Trace logs	Add	Change
Workflows	Add	Change

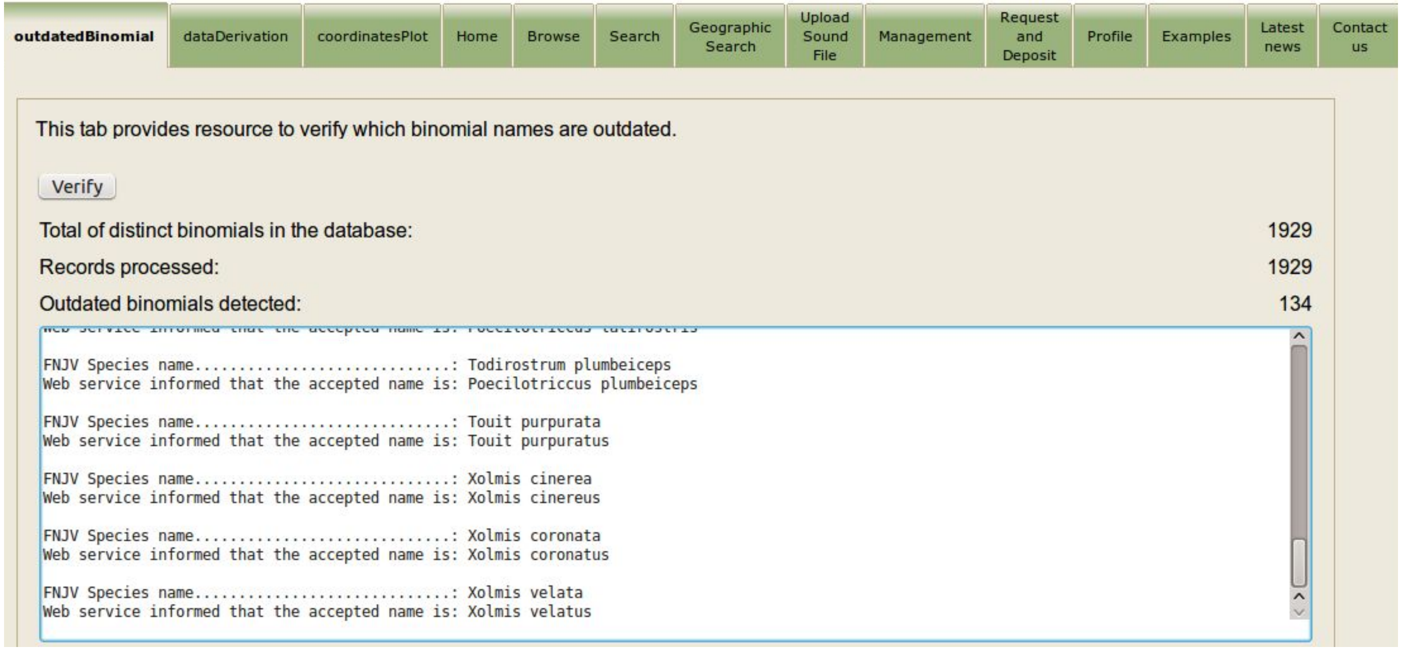
Case study: Long term data preservation and curation



The Fonoteca Neotropical Jacques Vielliard (FNJV) is a collection of archives and vocalizations of animals, mainly from Neotropical region. This is one of the 10 biggest sound collections in the world, having recordings of all vertebrates groups (fishes, amphibians, reptiles, birds and mammals) and some groups of invertebrates (as insects and arachnids). Currently, Fonoteca Neotropical has more than 30 thousand deposits of vocalizations.

This computational system is a project developed at University of Campinas, produced as a joint action between the

Animal sound recording metadata system



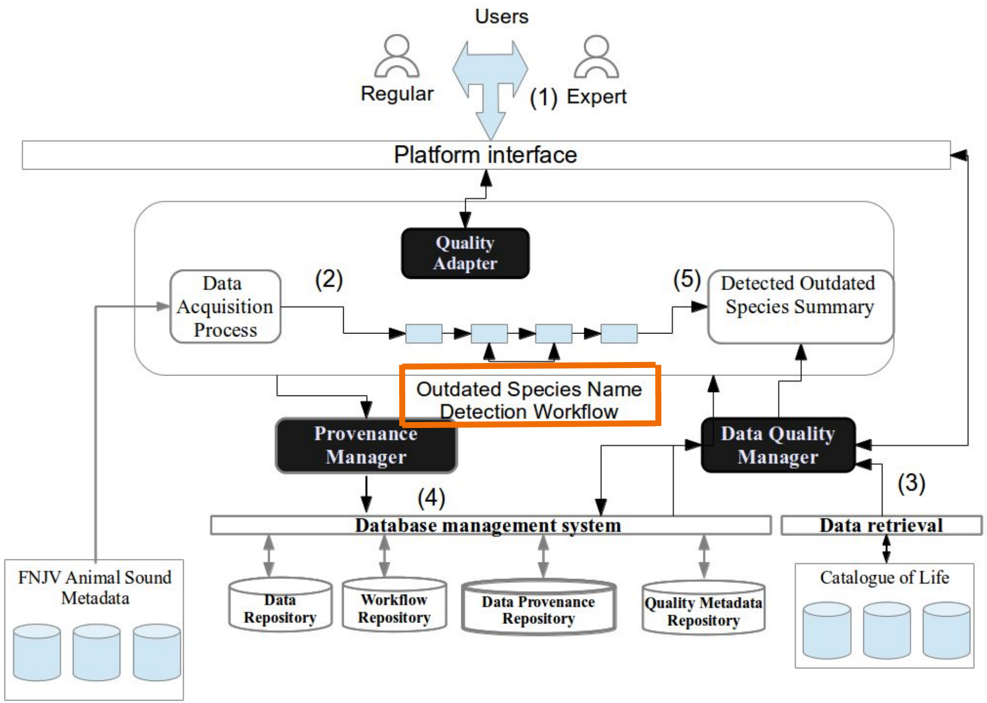
This tab provides resource to verify which binomial names are outdated.

Total of distinct binomials in the database:	1929
Records processed:	1929
Outdated binomials detected:	134

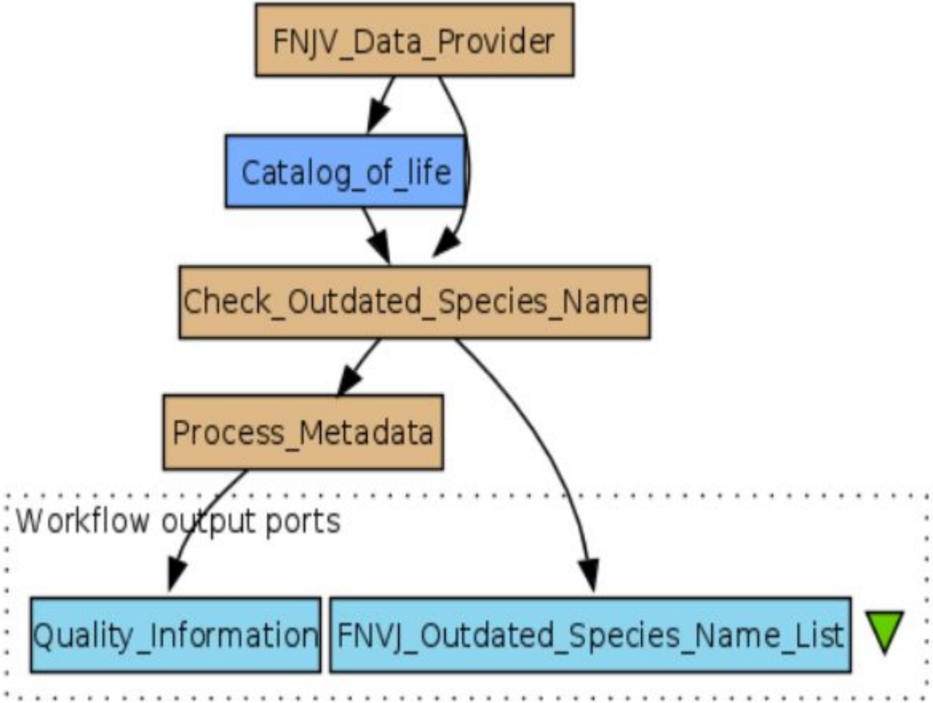
```
Web service informed that the accepted name is: Poecilotriccus latirostris

FNJV Species name............................: Todirostrum plumbeiceps
Web service informed that the accepted name is: Poecilotriccus plumbeiceps

FNJV Species name............................: Touit purpurata
Web service informed that the accepted name is: Touit purpuratus

FNJV Species name............................: Xolmis cinerea
Web service informed that the accepted name is: Xolmis cinereus

FNJV Species name............................: Xolmis coronata
Web service informed that the accepted name is: Xolmis coronatus

FNJV Species name............................: Xolmis velata
Web service informed that the accepted name is: Xolmis velatus
```

Outdated names – 7% (134 species)

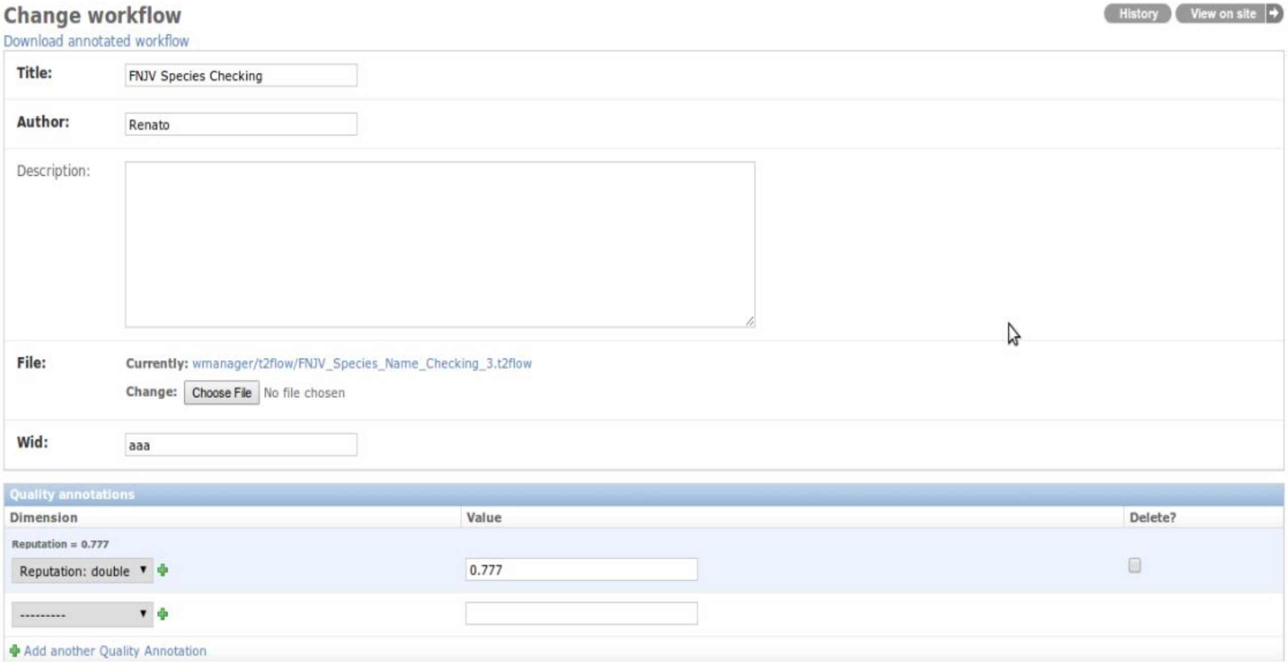
Using Quality Flow



Abstract workflow: Quality Processing



Quality Adapter



Change workflow

Download annotated workflow

Title:	FNJV Species Checking
Author:	Renato
Description:	
File:	**Currently:** wmanager/t2flow/FNJV_Species_Name_Checking_3.t2flow **Change:** Choose File No file chosen
Wid:	aaa

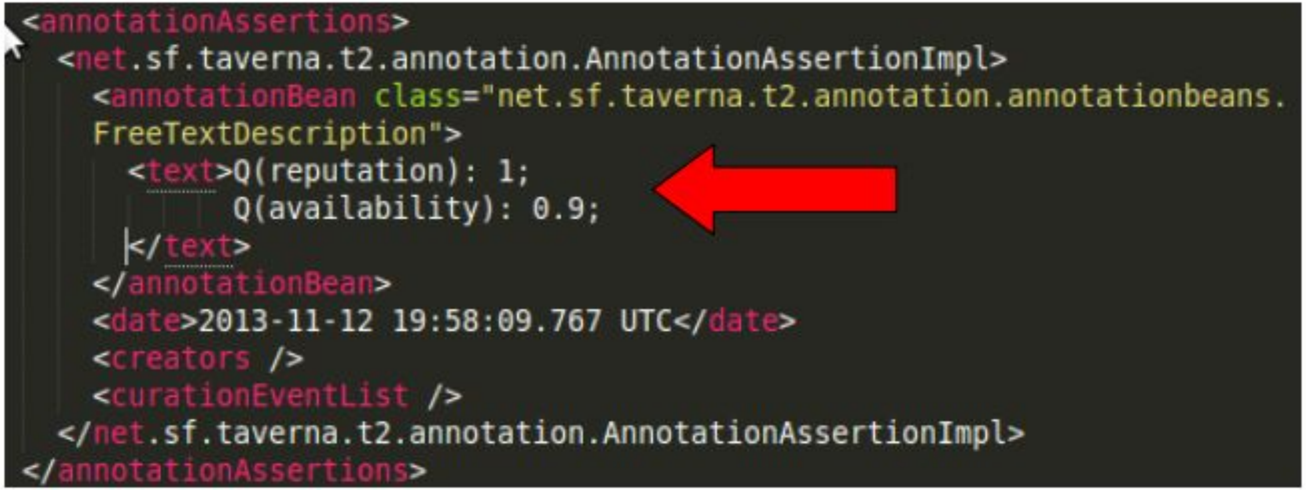
Quality annotations

Dimension	Value	Delete?
Reputation = 0.777 Reputation: double	0.777	□

Add another Quality Annotation

Quality Adapter

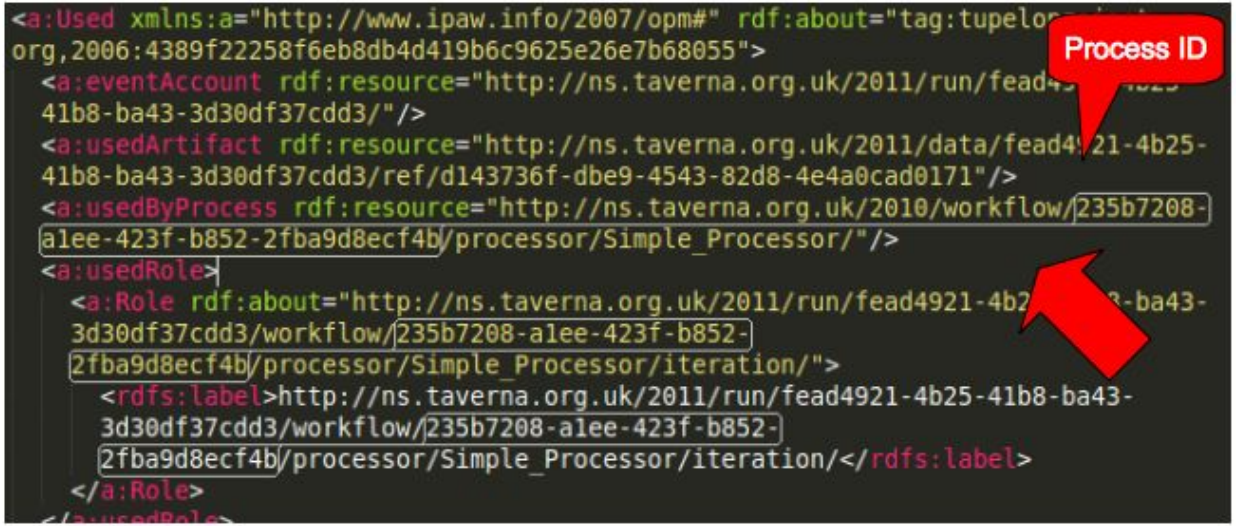
- Insert quality information to a workflow specification
- No changes to workflow model



```
<annotationAssertions>
  <net.sf.taverna.t2.annotation.AnnotationAssertionImpl>
    <annotationBean class="net.sf.taverna.t2.annotation.annotationbeans.
    FreeTextDescription">
      <text>Q(reputation): 1;
            Q(availability): 0.9;
      </text>
    </annotationBean>
    <date>2013-11-12 19:58:09.767 UTC</date>
    <creators />
    <curationEventList />
  </net.sf.taverna.t2.annotation.AnnotationAssertionImpl>
</annotationAssertions>
```

Provenance Manager

Extraction of provenance information from trace logs



```
<a:Used xmlns:a="http://www.ipaw.info/2007/opm#" rdf:about="tag:tupelo
org,2006:4389f22258f6eb8db4d419b6c9625e26e7b68055">
  <a:eventAccount rdf:resource="http://ns.taverna.org.uk/2011/run/fead4
  41b8-ba43-3d30df37cdd3/"/>
  <a:usedArtifact rdf:resource="http://ns.taverna.org.uk/2011/data/fead4921-4b25-
  41b8-ba43-3d30df37cdd3/ref/d143736f-dbe9-4543-82d8-4e4a0cad0171"/>
  <a:usedByProcess rdf:resource="http://ns.taverna.org.uk/2010/workflow/235b7208-
  a1ee-423f-b852-2fba9d8ecf4b/processor/Simple_Processor/"/>
  <a:usedRole>
    <a:Role rdf:about="http://ns.taverna.org.uk/2011/run/fead4921-4b2      8-ba43-
    3d30df37cdd3/workflow/235b7208-a1ee-423f-b852-
    2fba9d8ecf4b/processor/Simple_Processor/iteration/">
      <rdfs:label>http://ns.taverna.org.uk/2011/run/fead4921-4b25-41b8-ba43-
      3d30df37cdd3/workflow/235b7208-a1ee-423f-b852-
      2fba9d8ecf4b/processor/Simple_Processor/iteration/</rdfs:label>
    </a:Role>
  </a:usedRole>
```

Quality Manager

Data quality assessment:

- From provenance.
- From annotations – quality attributes generated by Quality Adapter.

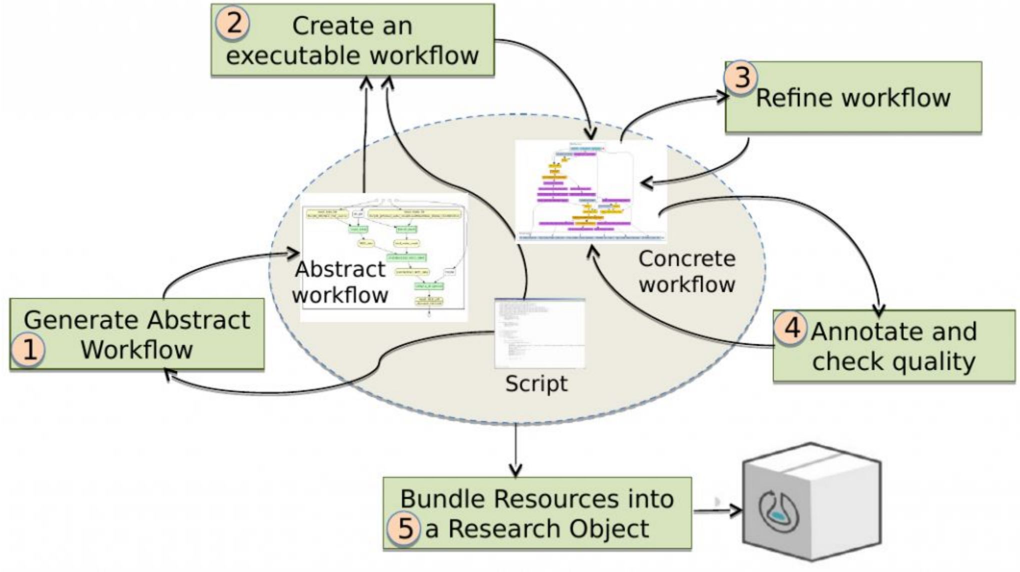
Papers:

- 2017. A User-Sensitive Quality Assessment Approach for Experiments in eScience. Journal of Data and Information Quality (JDIQ). (under review)

w2Share
Carvalho, Lucas PhD (ongoing)

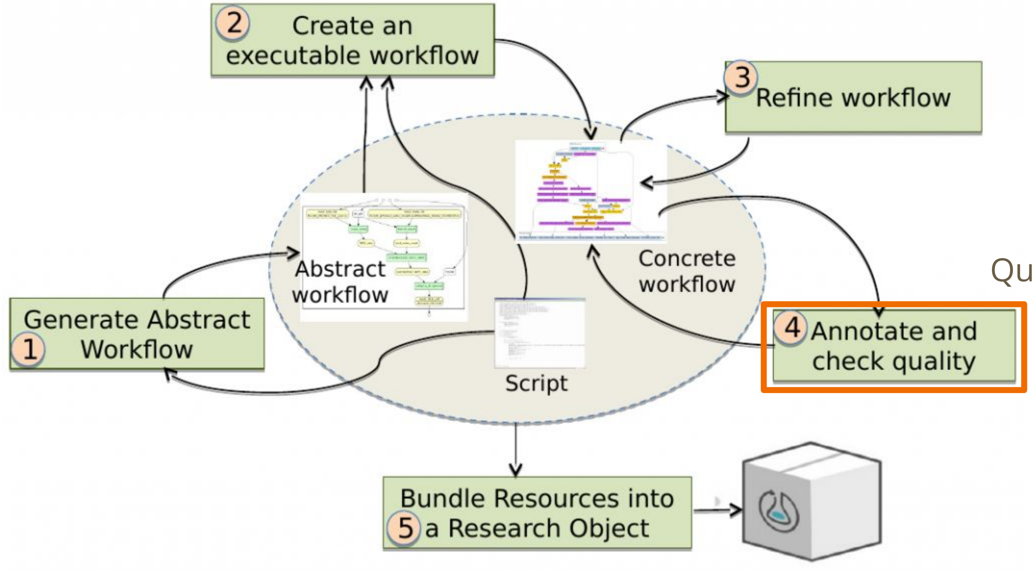
w2Share (Carvalho, Lucas PhD (ongoing))

Reproducibility and workflow management across scientific disciplines


2 Create an executable workflow
3 Refine workflow
Abstract workflow
Concrete workflow
Generate Abstract 1 Workflow
Script
4 Annotate and check quality
Bundle Resources into 5 a Research Object

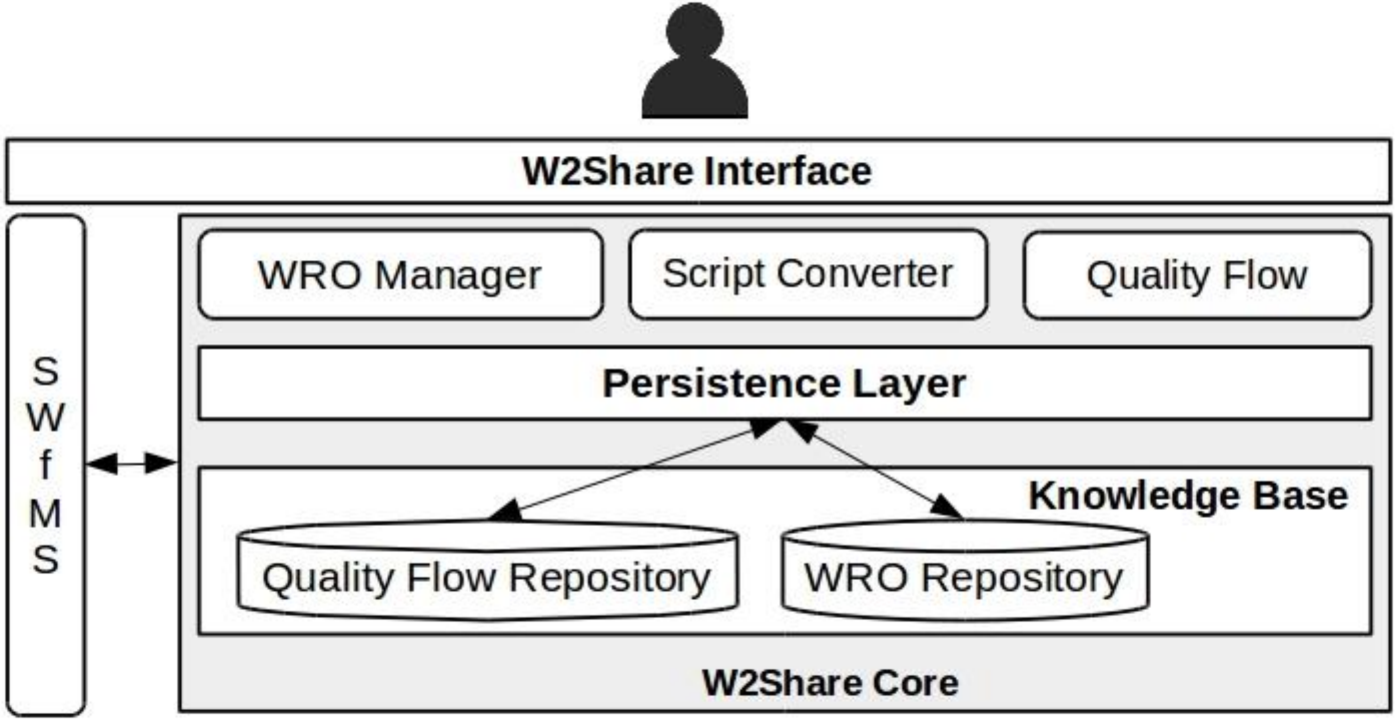
w2Share

Reproducibility and workflow management across scientific disciplines


2 Create an executable workflow
3 Refine workflow
Abstract workflow
Concrete workflow
Generate Abstract 1 Workflow
4 Annotate and check quality
Script
Bundle Resources into 5 a Research Object

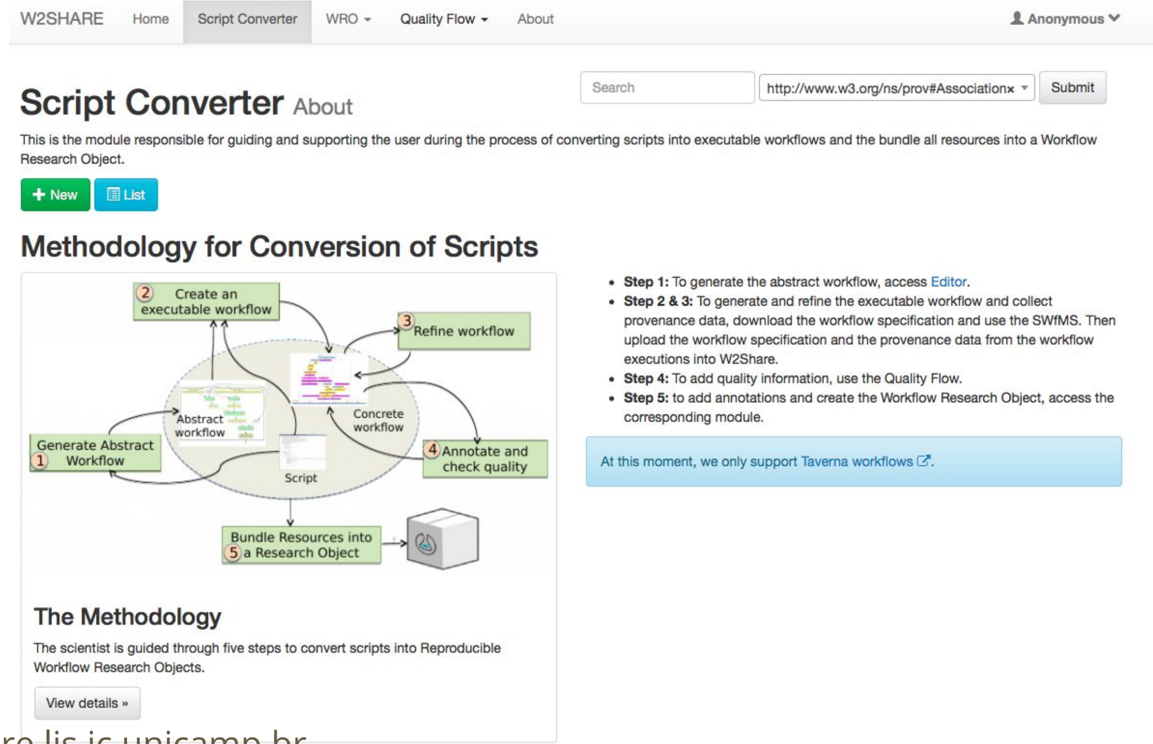
Quality Flow!!!

w2Share



Available at w2share.lis.ic.unicamp.br

w2Share


W2SHARE
Home
Script Converter
WRO
Quality Flow
About
Anonymous
Search
http://www.w3.org/ns/prov#Association×
Submit
New
List
2 Create an executable workflow
3 Refine workflow
Abstract workflow
Concrete workflow
Generate Abstract 1 Workflow
4 Annotate and check quality
Script
Bundle Resources into 5 a Research Object
View details »

Script Converter About

This is the module responsible for guiding and supporting the user during the process of converting scripts into executable workflows and the bundle all resources into a Workflow Research Object.

Methodology for Conversion of Scripts

- **Step 1:** To generate the abstract workflow, access Editor.
- **Step 2 & 3:** To generate and refine the executable workflow and collect provenance data, download the workflow specification and use the SWfMS. Then upload the workflow specification and the provenance data from the workflow executions into W2Share.
- **Step 4:** To add quality information, use the Quality Flow.
- **Step 5:** to add annotations and create the Workflow Research Object, access the corresponding module.

At this moment, we only support Taverna workflows.

The Methodology

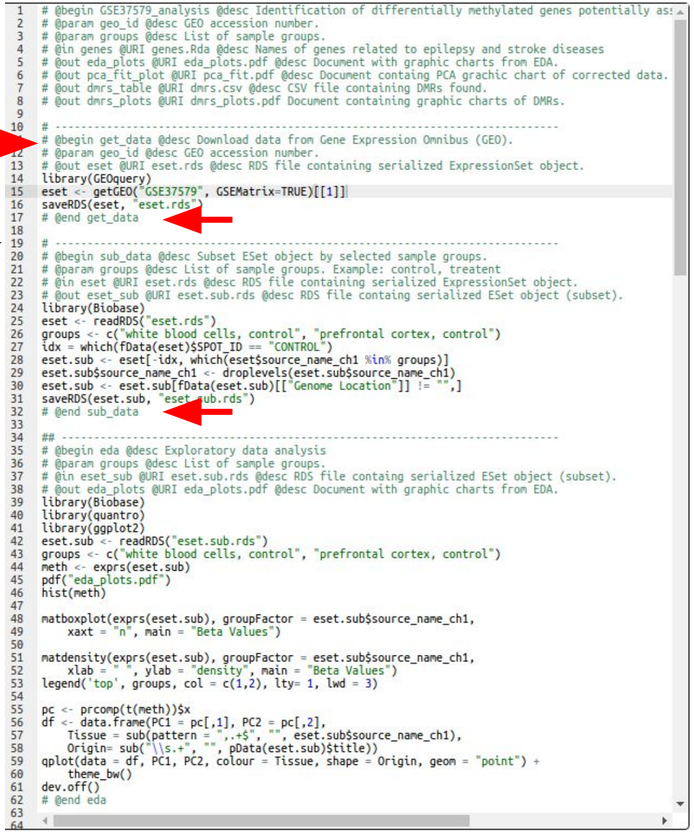
The scientist is guided through five steps to convert scripts into Reproducible Workflow Research Objects.

Available at w2share.lis.ic.unicamp.br

Case study: DNA Methylation Microarray Analysis

w2Share: Generating the abstract workflow

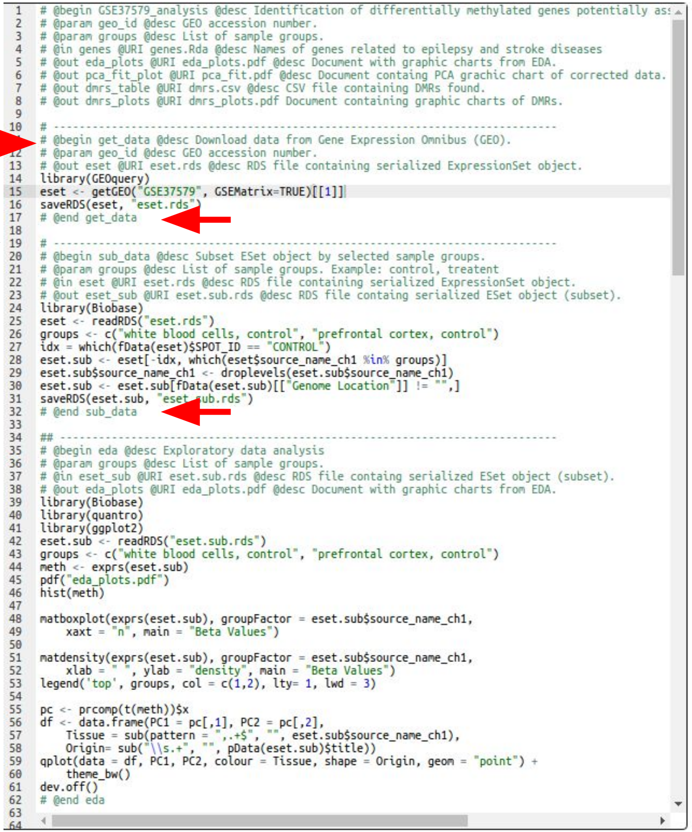
Script annotations



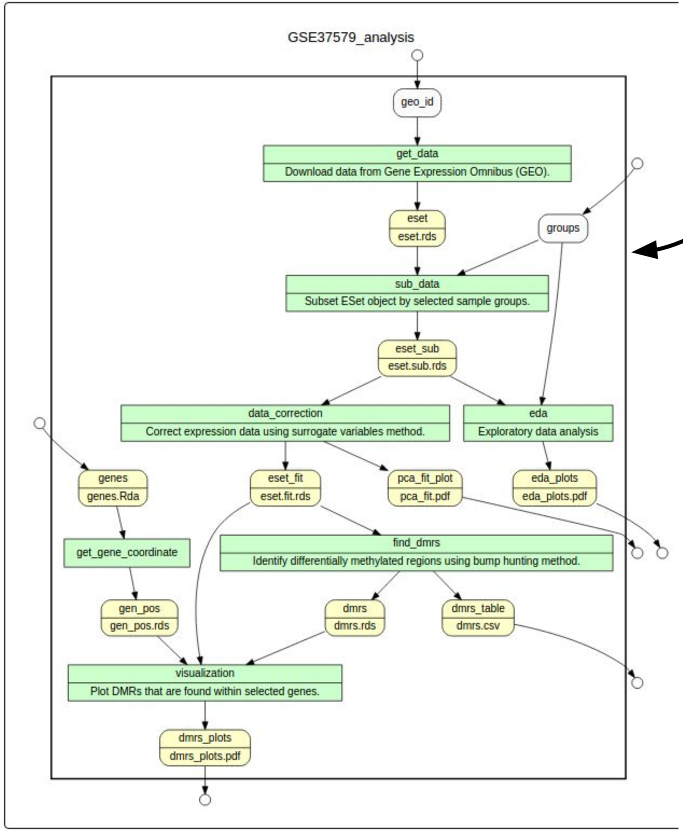
```
# @begin GSE37579_analysis @desc Identification of differentially methylated genes potentially as
# @param geo_id @desc GEO accession number.
# @param groups @desc List of sample groups.
# @in genes @URI genes.Rda @desc Names of genes related to epilepsy and stroke diseases
# @out eda_plots @URI eda_plots.pdf @desc Document with graphic charts from EDA.
# @out pca_fit_plot @URI pca_fit.pdf @desc Document containg PCA grachic chart of corrected data.
# @out dmrs_table @URI dmrs.csv @desc CSV file containing DMRs found.
# @out dmrs_plots @URI dmrs_plots.pdf Document containing graphic charts of DMRs.

# -----------------------------------------------------------------------------
# @begin get_data @desc Download data from Gene Expression Omnibus (GEO).
# @param geo_id @desc GEO accession number.
# @out eset @URI eset.rds @desc RDS file containing serialized ExpressionSet object.
library(GEOquery)
eset <- getGEO("GSE37579", GSEMatrix=TRUE)[[1]]
saveRDS(eset, "eset.rds")
# @end get_data

# -----------------------------------------------------------------------------
# @begin sub_data @desc Subset ESet object by selected sample groups.
# @param groups @desc List of sample groups. Example: control, treatent
# @in eset @URI eset.rds @desc RDS file containing serialized ExpressionSet object.
# @out eset_sub @URI eset.sub.rds @desc RDS file containg serialized ESet object (subset).
library(Biobase)
eset <- readRDS("eset.rds")
groups <- c("white blood cells, control", "prefrontal cortex, control")
idx = which(fData(eset)$SPOT_ID == "CONTROL")
eset.sub <- eset[-idx, which(eset$source_name_ch1 %in% groups)]
eset.sub$source_name_ch1 <- droplevels(eset.sub$source_name_ch1)
eset.sub <- eset.sub[fData(eset.sub)[["Genome Location"]] != "",]
saveRDS(eset.sub, "eset.sub.rds")
# @end sub_data

## ----------------------------------------------------------------------------
# @begin eda @desc Exploratory data analysis
# @param groups @desc List of sample groups.
# @in eset_sub @URI eset.sub.rds @desc RDS file containg serialized ESet object (subset).
# @out eda_plots @URI eda_plots.pdf @desc Document with graphic charts from EDA.
library(Biobase)
library(quantro)
library(ggplot2)
eset.sub <- readRDS("eset.sub.rds")
groups <- c("white blood cells, control", "prefrontal cortex, control")
meth <- exprs(eset.sub)
pdf("eda_plots.pdf")
hist(meth)

matboxplot(exprs(eset.sub), groupFactor = eset.sub$source_name_ch1,
    xaxt = "n", main = "Beta Values")

matdensity(exprs(eset.sub), groupFactor = eset.sub$source_name_ch1,
    xlab = " ", ylab = "density", main = "Beta Values")
legend('top', groups, col = c(1,2), lty= 1, lwd = 3)

pc <- prcomp(t(meth))$x
df <- data.frame(PC1 = pc[,1], PC2 = pc[,2],
    Tissue = sub(pattern = ",.+$", "", eset.sub$source_name_ch1),
    Origin= sub("\\s.+", "", pData(eset.sub)$title))
qplot(data = df, PC1, PC2, colour = Tissue, shape = Origin, geom = "point") +
    theme_bw()
dev.off()
# @end eda

```

w2Share: Generating the abstract workflow

Script annotations

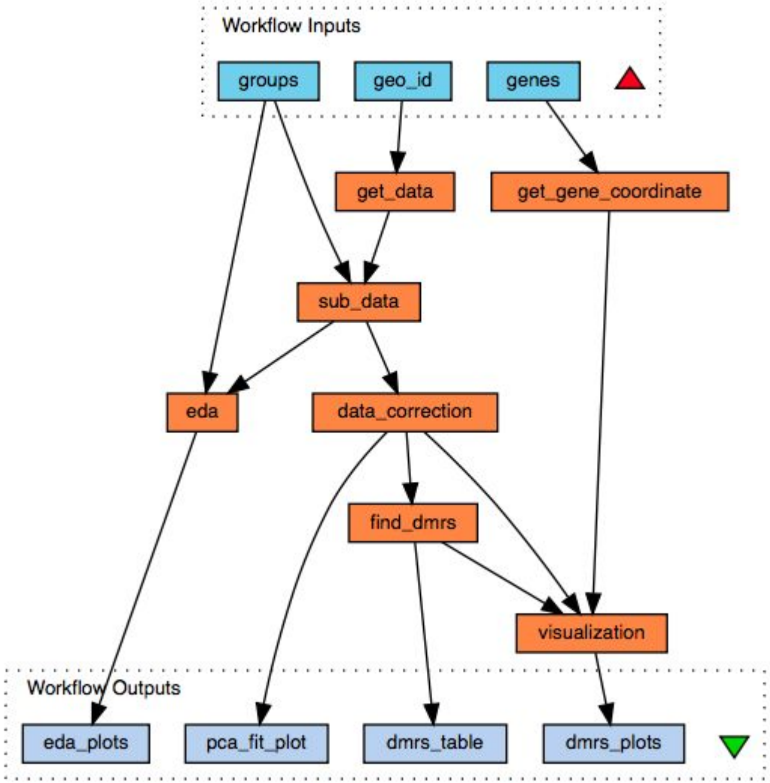


```
# @begin GSE37579_analysis @desc Identification of differentially methylated genes potentially ass
# @param geo_id @desc GEO accession number.
# @param groups @desc List of sample groups.
# @in genes @URI genes.Rda @desc Names of genes related to epilepsy and stroke diseases
# @out eda_plots @URI eda_plots.pdf @desc Document with graphic charts from EDA.
# @out pca_fit_plot @URI pca_fit.pdf @desc Document containg PCA grachic chart of corrected data.
# @out dmrs_table @URI dmrs.csv @desc CSV file containing DMRs found.
# @out dmrs_plots @URI dmrs_plots.pdf Document containing graphic charts of DMRs.

# -----------------------------------------------------------------------------
# @begin get_data @desc Download data from Gene Expression Omnibus (GEO).
# @param geo_id @desc GEO accession number.
# @out eset @URI eset.rds @desc RDS file containing serialized ExpressionSet object.
library(GEOquery)
eset <- getGEO("GSE37579", GSEMatrix=TRUE)[[1]]
saveRDS(eset, "eset.rds")
# @end get_data

# -----------------------------------------------------------------------------
# @begin sub_data @desc Subset ESet object by selected sample groups.
# @param groups @desc List of sample groups. Example: control, treatent
# @in eset @URI eset.rds @desc RDS file containing serialized ExpressionSet object.
# @out eset_sub @URI eset.sub.rds @desc RDS file containg serialized ESet object (subset).
library(Biobase)
eset <- readRDS("eset.rds")
groups <- c("white blood cells, control", "prefrontal cortex, control")
idx = which(fData(eset)$SPOT_ID == "CONTROL")
eset.sub <- eset[-idx, which(eset$source_name_ch1 %in% groups)]
eset.sub$source_name_ch1 <- droplevels(eset.sub$source_name_ch1)
eset.sub <- eset.sub[fData(eset.sub)[["Genome Location"]] != "",]
saveRDS(eset.sub, "eset.sub.rds")
# @end sub_data

## ----------------------------------------------------------------------------
# @begin eda @desc Exploratory data analysis
# @param groups @desc List of sample groups.
# @in eset_sub @URI eset.sub.rds @desc RDS file containg serialized ESet object (subset).
# @out eda_plots @URI eda_plots.pdf @desc Document with graphic charts from EDA.
library(Biobase)
library(quantro)
library(ggplot2)
eset.sub <- readRDS("eset.sub.rds")
groups <- c("white blood cells, control", "prefrontal cortex, control")
meth <- exprs(eset.sub)
pdf("eda_plots.pdf")
hist(meth)

matboxplot(exprs(eset.sub), groupFactor = eset.sub$source_name_ch1,
    xaxt = "n", main = "Beta Values")

matdensity(exprs(eset.sub), groupFactor = eset.sub$source_name_ch1,
    xlab = " ", ylab = "density", main = "Beta Values")
legend('top', groups, col = c(1,2), lty= 1, lwd = 3)

pc <- prcomp(t(meth))$x
df <- data.frame(PC1 = pc[,1], PC2 = pc[,2],
    Tissue = sub(pattern = ",.+$", "", eset.sub$source_name_ch1),
    Origin= sub("\\s.+", "", pData(eset.sub)$title))
qplot(data = df, PC1, PC2, colour = Tissue, shape = Origin, geom = "point") +
    theme_bw()
dev.off()
# @end eda
```

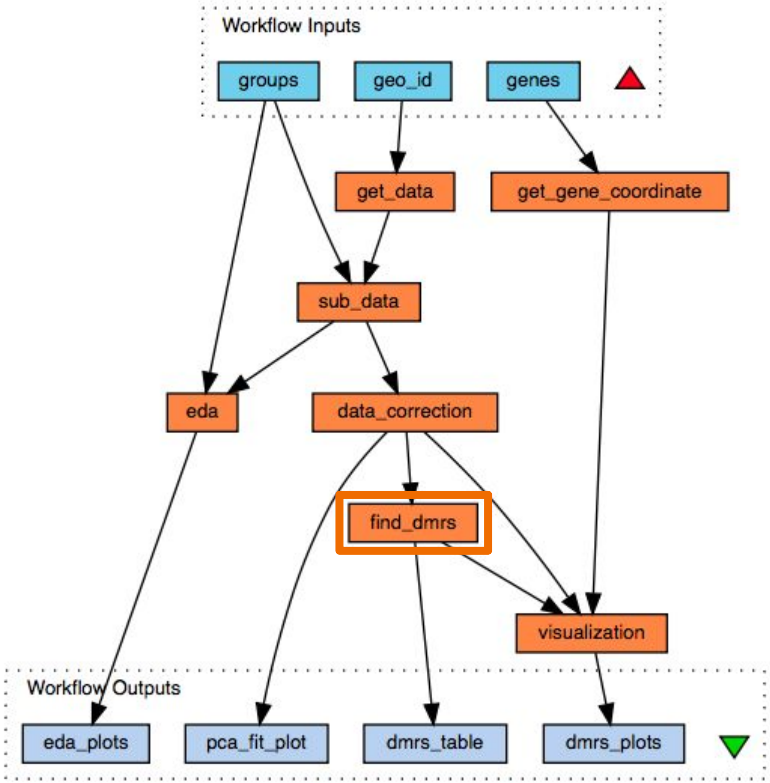

GSE37579_analysis
geo_id
get_data
Download data from Gene Expression Omnibus (GEO).
eset
eset.rds
groups
sub_data
Subset ESet object by selected sample groups.
eset_sub
eset.sub.rds
data_correction
Correct expression data using surrogate variables method.
eda
Exploratory data analysis
genes
genes.Rda
eset_fit
eset.fit.rds
pca_fit_plot
pca_fit.pdf
eda_plots
eda_plots.pdf
get_gene_coordinate
find_dmrs
Identify differentially methylated regions using bump hunting method.
gen_pos
gen_pos.rds
dmrs
dmrs.rds
dmrs_table
dmrs.csv
visualization
Plot DMRs that are found within selected genes.
dmrs_plots
dmrs_plots.pdf

Abstract workflow

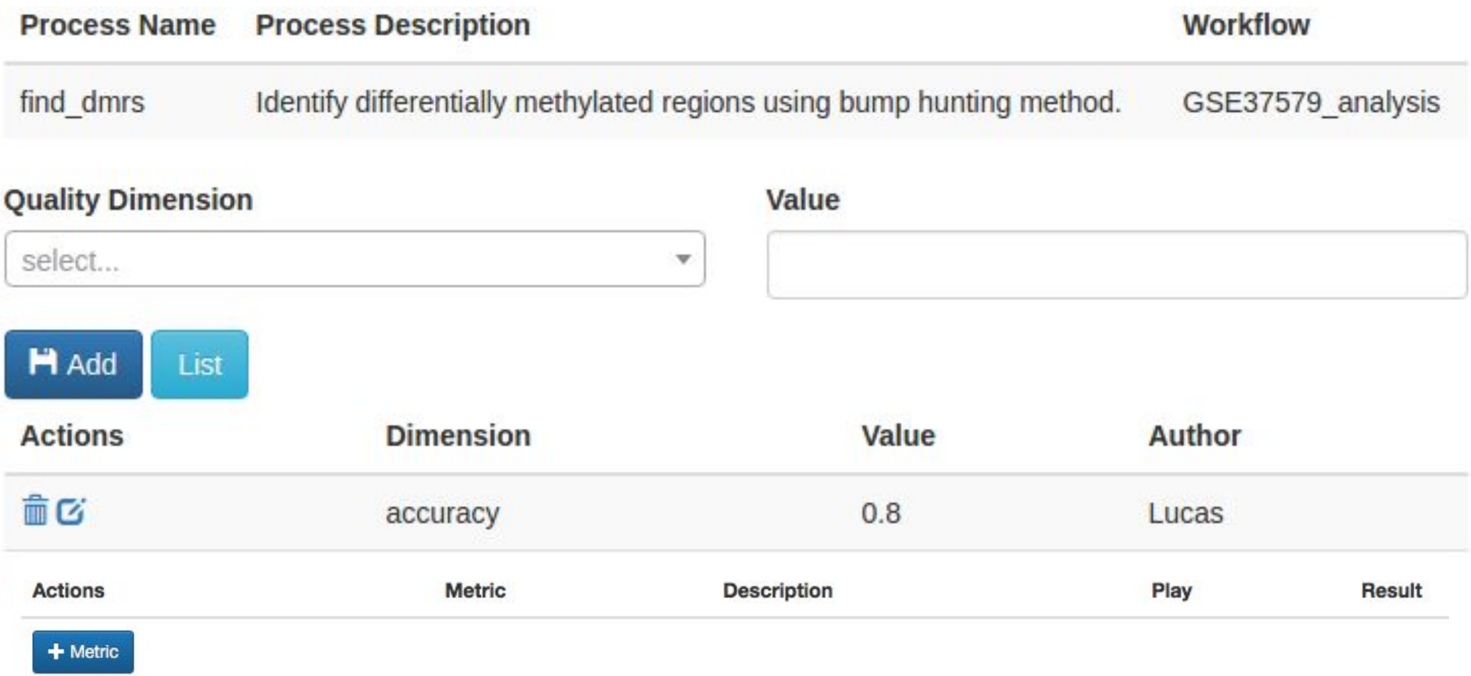
w2Share: Generating the executable workflow



w2Share: Annotating with quality information



Annotating with quality information



Process Name	Process Description	Workflow
find_dmrs	Identify differentially methylated regions using bump hunting method.	GSE37579_analysis

Quality Dimension

select... ▾

Value

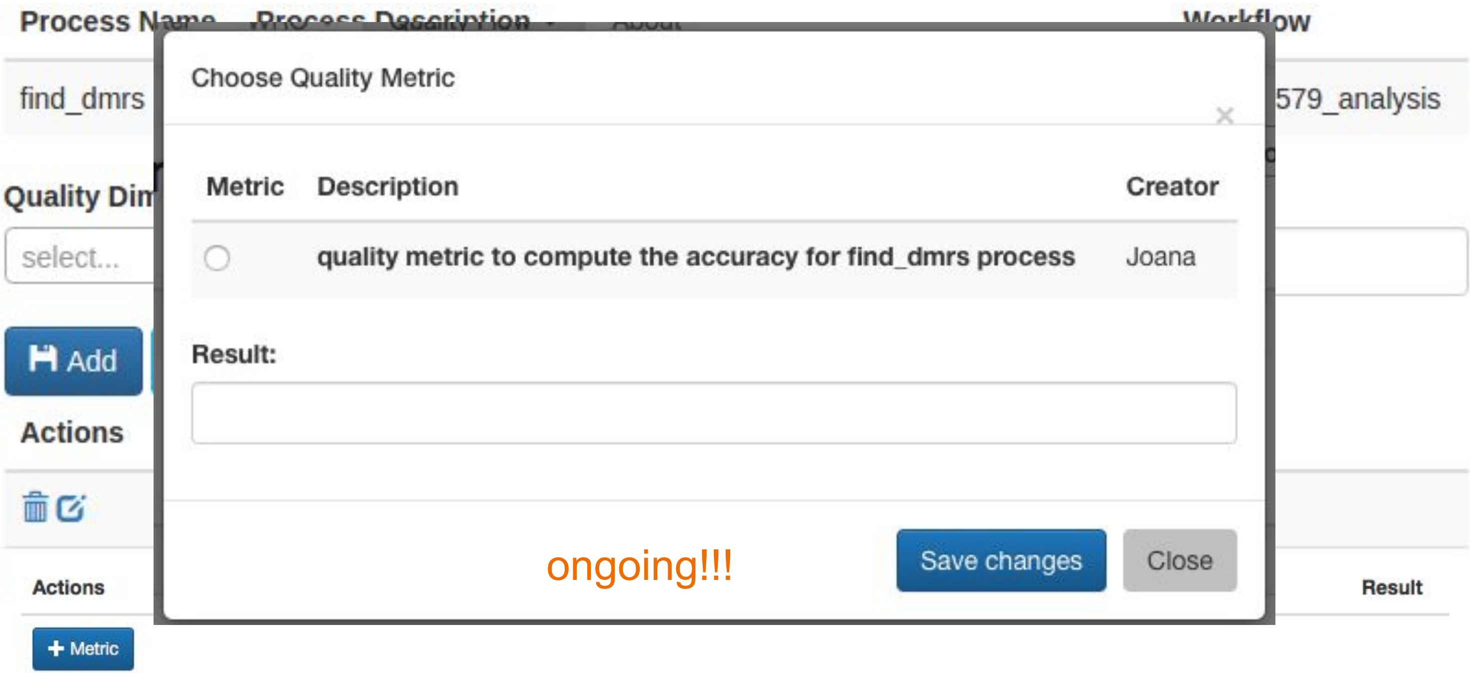
Add

List

Actions	Dimension	Value	Author
🗑 ✎	accuracy	0.8	Lucas

Actions	Metric	Description	Play	Result
+ Metric				

Annotating with quality information



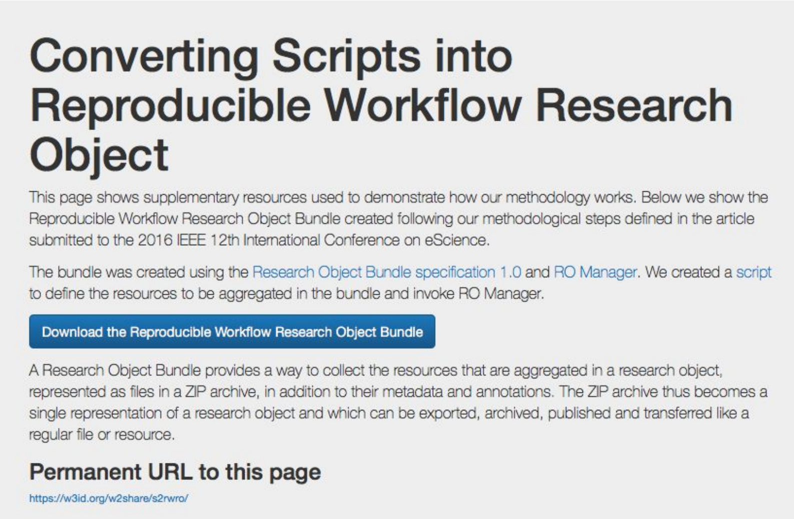
Choose Quality Metric

Metric	Description	Creator
○	quality metric to compute the accuracy for find_dmrs process	Joana

Result:

ongoing!!!

w2Share: Generating the WRO



Converting Scripts into Reproducible Workflow Research Object

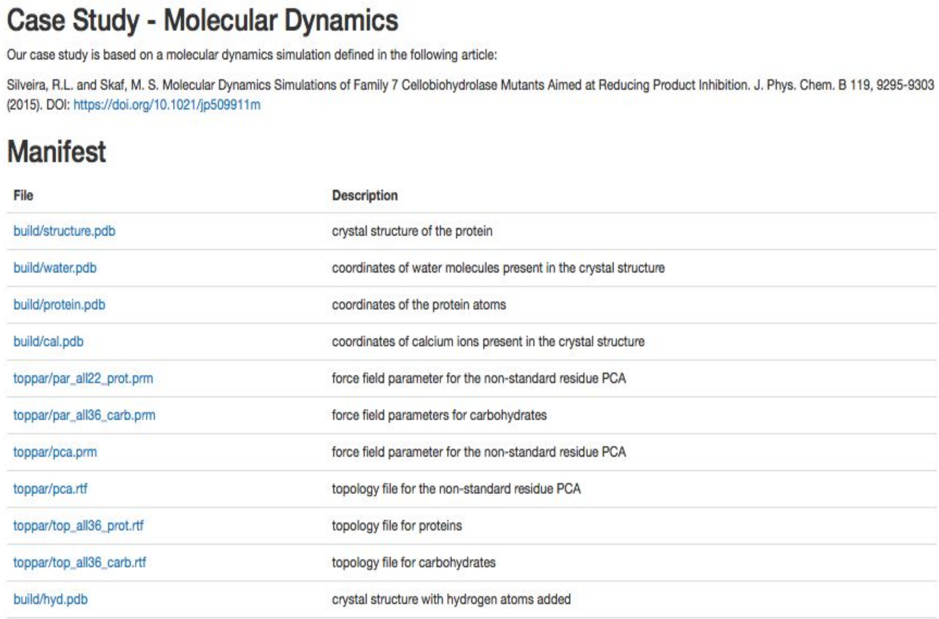
This page shows supplementary resources used to demonstrate how our methodology works. Below we show the Reproducible Workflow Research Object Bundle created following our methodological steps defined in the article submitted to the 2016 IEEE 12th International Conference on eScience.

The bundle was created using the Research Object Bundle specification 1.0 and RO Manager. We created a script to define the resources to be aggregated in the bundle and invoke RO Manager.

A Research Object Bundle provides a way to collect the resources that are aggregated in a research object, represented as files in a ZIP archive, in addition to their metadata and annotations. The ZIP archive thus becomes a single representation of a research object and which can be exported, archived, published and transferred like a regular file or resource.

Permanent URL to this page

https://w3id.org/w2share/s2rwro/



Case Study - Molecular Dynamics

Our case study is based on a molecular dynamics simulation defined in the following article:

Silveira, R.L. and Skaf, M. S. Molecular Dynamics Simulations of Family 7 Cellobiohydrolase Mutants Aimed at Reducing Product Inhibition. J. Phys. Chem. B 119, 9295-9303 (2015). DOI: https://doi.org/10.1021/jp509911m

Manifest

File	Description
build/structure.pdb	crystal structure of the protein
build/water.pdb	coordinates of water molecules present in the crystal structure
build/protein.pdb	coordinates of the protein atoms
build/cal.pdb	coordinates of calcium ions present in the crystal structure
toppar/par_all22_prot.prm	force field parameter for the non-standard residue PCA
toppar/par_all36_carb.prm	force field parameters for carbohydrates
toppar/pca.prm	force field parameter for the non-standard residue PCA
toppar/pca.rtf	topology file for the non-standard residue PCA
toppar/top_all36_prot.rtf	topology file for proteins
toppar/top_all36_carb.rtf	topology file for carbohydrates
build/hyd.pdb	crystal structure with hydrogen atoms added

Available at
https://w3id.org/w2share/s2rwro/

w2Share

Papers:

- 2017. Implementing W2Share: Supporting Reproducibility and Quality Assessment in eScience. XI Brazilian e-Science Workshop (BreSci). (under review)

Contributions

- A solution to explore the assessment of quality in the context of scientific experiments:
    - scientific workflows systems + provenance + (semantic) models + standards
    - Enrichment of provenance to provide a greater amount of information
- Quality-aware workflows
    - provide user-dependent quality assessment

What’s next?

Ongoing work

- How to combine the information from workflow executions to feed and compute the quality metrics
- (Semantic) rules to derive quality
  - Minim model for defining checklists (must/should/may requirements, associated with rules)

Acknowledgements

- Laboratory of Information Systems (LIS)
  - Carvalho, Lucas (PhD student) & Sousa, Renato B. (Master in Computer Science)
- IC infrastructure and staff
- Financial support:
  - CAPES
  - Microsoft Research FAPESP Virtual Institute (NavScales Project)
  - CNPq (MuZOO Project)
  - FAPESP-PRONEX (eScience project)
  - FAPESP/CEPID
  - INCT in Web Science
  - CNPq

Thanks...

Supporting Data Quality Assessment in eScience

Postdoctoral fellow: Joana Gonzales Malaverri
Supervisor: Prof. Claudia Bauzer Medeiros

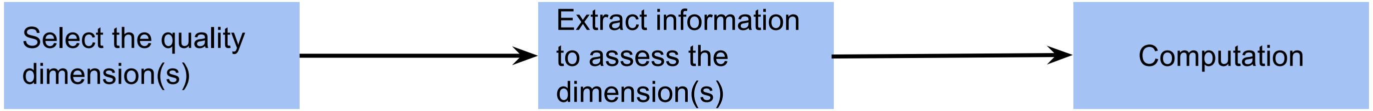
IC - Unicamp

Research on Data Quality

- 50s: researchers began to study quality issues related to products
- 90s: different definitions of data quality and division methods of quality dimensions.
    - 96 (Wang & Strong): MIT data quality group - fitness for use and data quality judgment depends on data consumers.
    - data quality dimension a set of data quality attributes.
- 2000:
    - data quality principles, ISO 8000
    - Approaches to assess quality: mathematical models, cleansing techniques, user feedback...

Extras:

Methodology

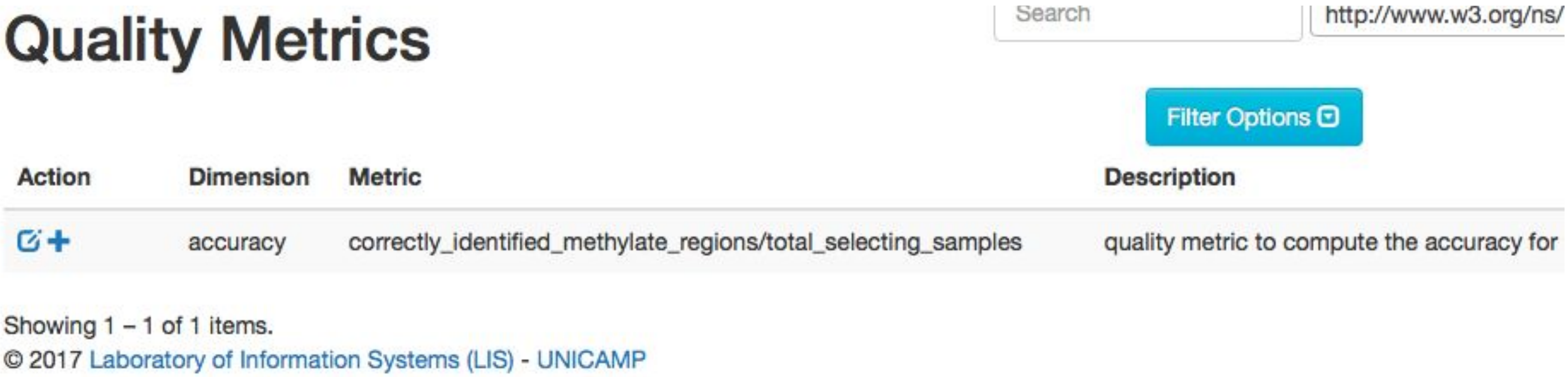


Select the quality dimension(s)

Extract information to assess the dimension(s)

Computation

w2Share: Annotating with quality information



Quality Metrics

Action	Dimension	Metric	Description
✎ +	accuracy	correctly_identified_methylate_regions/total_selecting_samples	quality metric to compute the accuracy for

Showing 1 – 1 of 1 items.