A provenance-based approach to manage long term preservation of scientific data

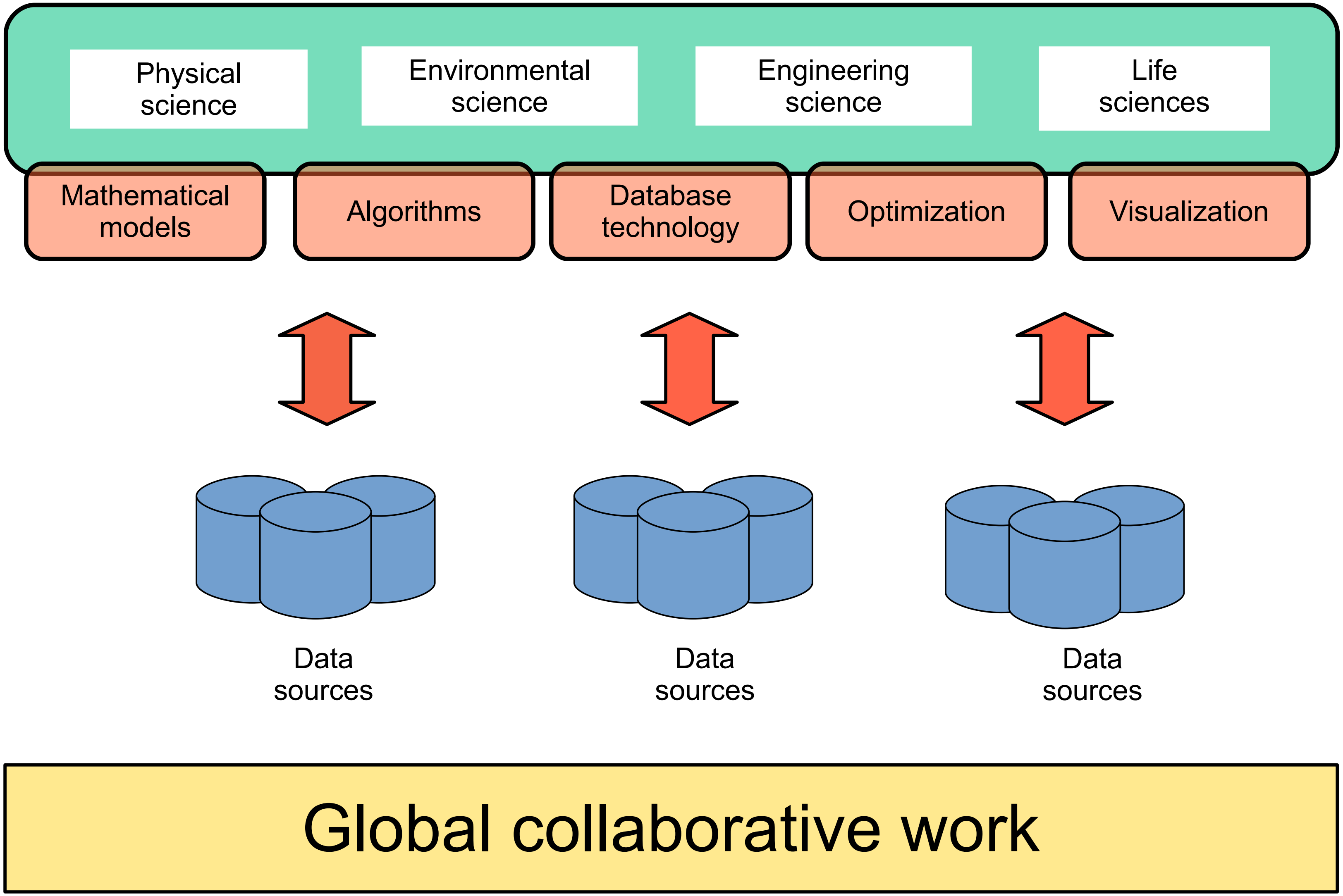
Renato Sousa, Daniel Cugler, Joana Malaverri,
Claudia Bauzer Medeiros,
Institute of Computing – University of Campinas

BRAZIL

Outline

- Motivation and related work
- Our approach
- Case Study
- Conclusions and Ongoing Work

eScience environments



Physical science

Environmental science

Engineering science

Life sciences

Mathematical models

Algorithms

Database technology

Optimization

Visualization

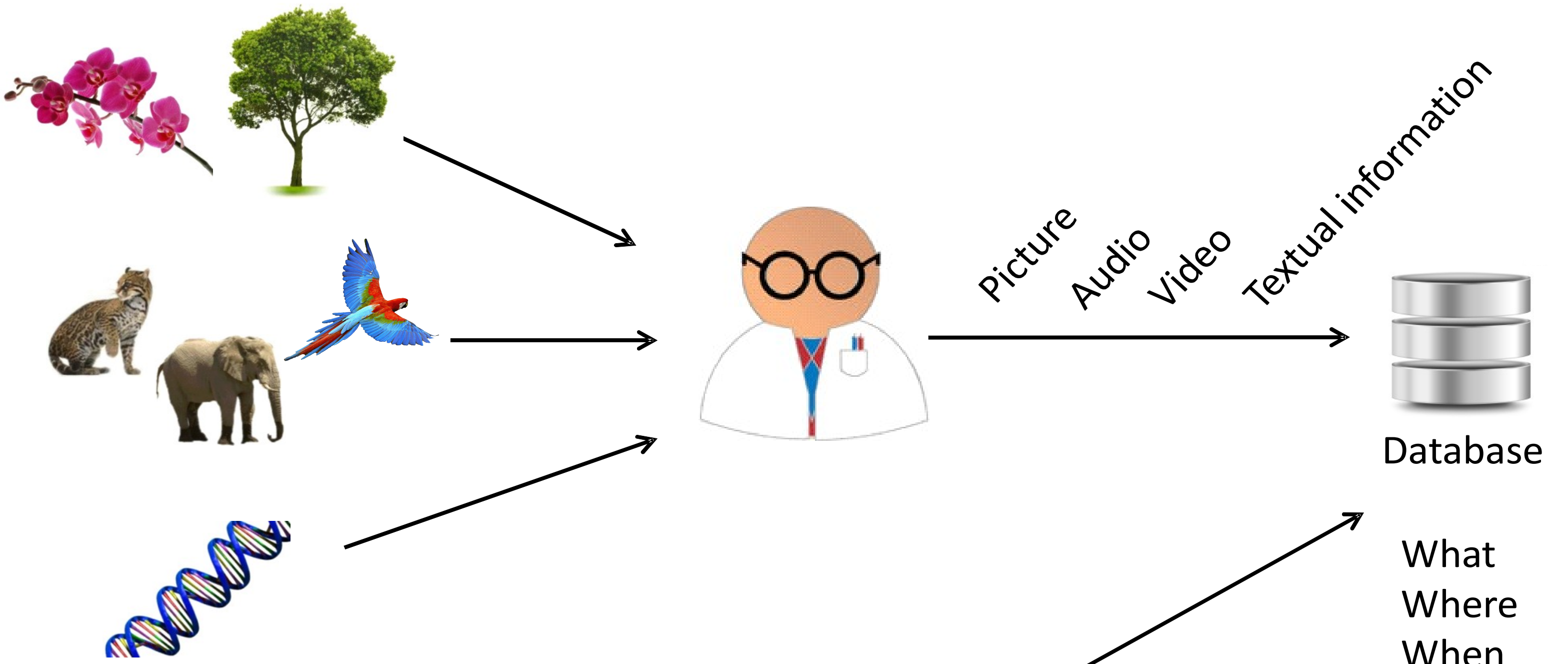
Data sources

Data sources

Data sources

Global collaborative work

Motivation – Biological Observation Databases



Picture

Audio

Video

Textual information

Database

What
Where
When
How

METADATA

Figures extracted from: flores.lojavirtualfc.com.br, www.trees4life.ca, www.jeratechsys.com, www.fanpop.com, www.arthursclipart.org, www.birdclubsva.org, www.geneticliteracyproject.org, www.psdgraphics.com

What is data?

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

- (Ideally, consider software, workflows, documentation, intermediate artifacts...)

Long term curation and preservation?

BASIC TENETS

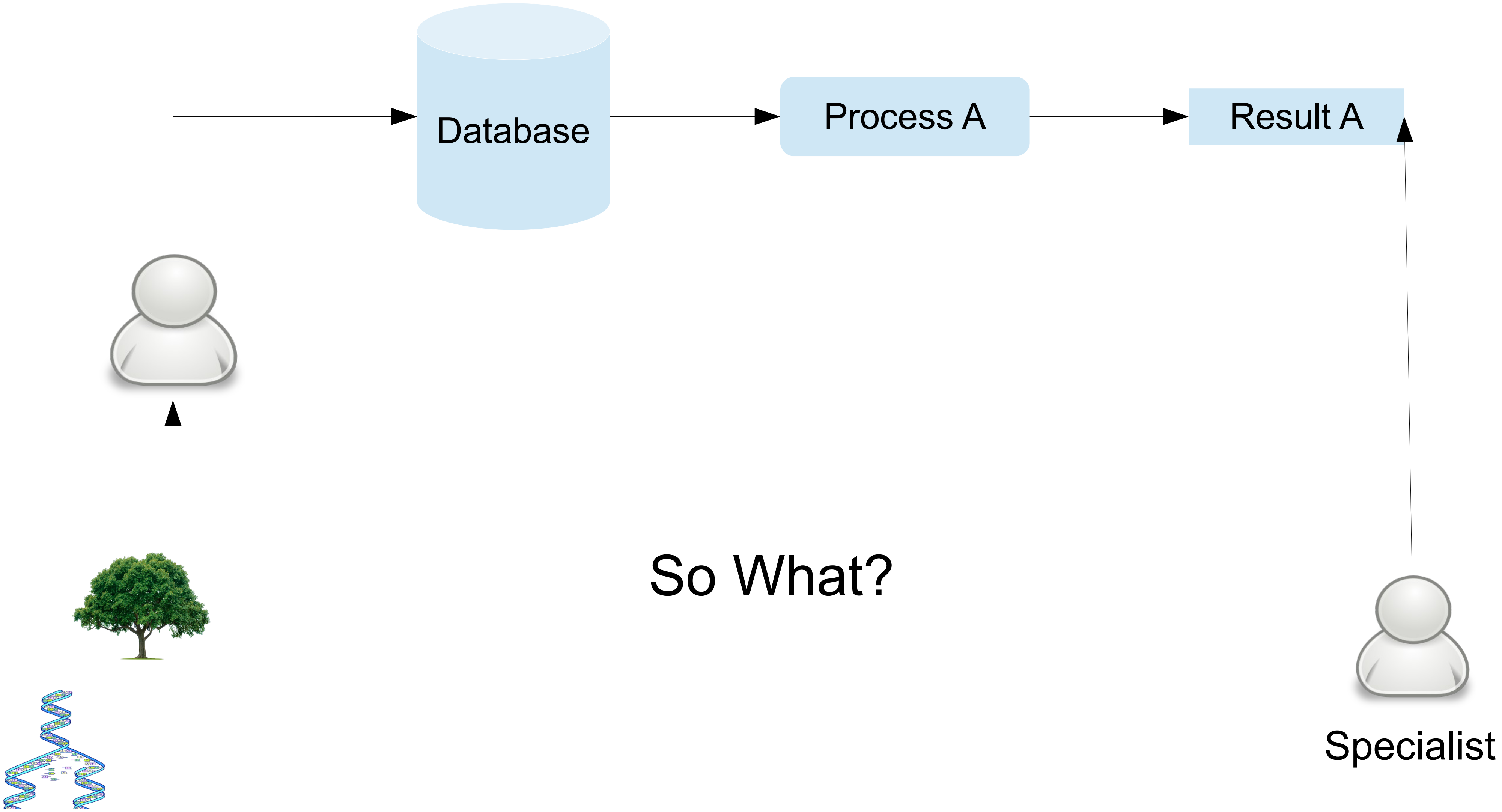
- As long as of interest to scholarship
- Curation and preservation are indissociate

Long term data preservation

REQUIRES

Ensuring (Meta)Data Quality

Motivating Example



Database

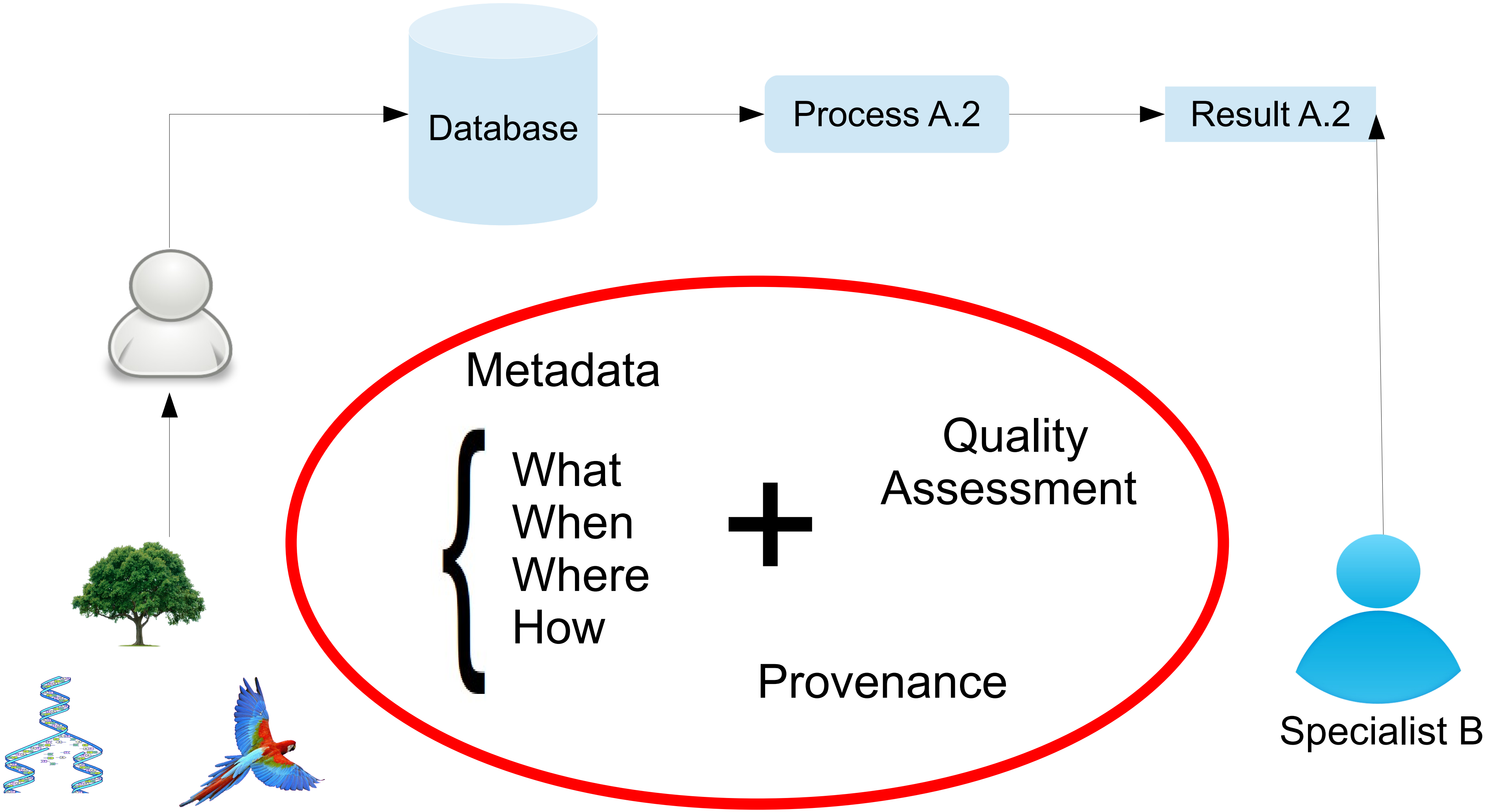
Process A

Result A

Specialist

So What?

Curating metadata



Database

Process A.2

Result A.2

Metadata

What
When
Where
How

+

Quality Assessment

Provenance

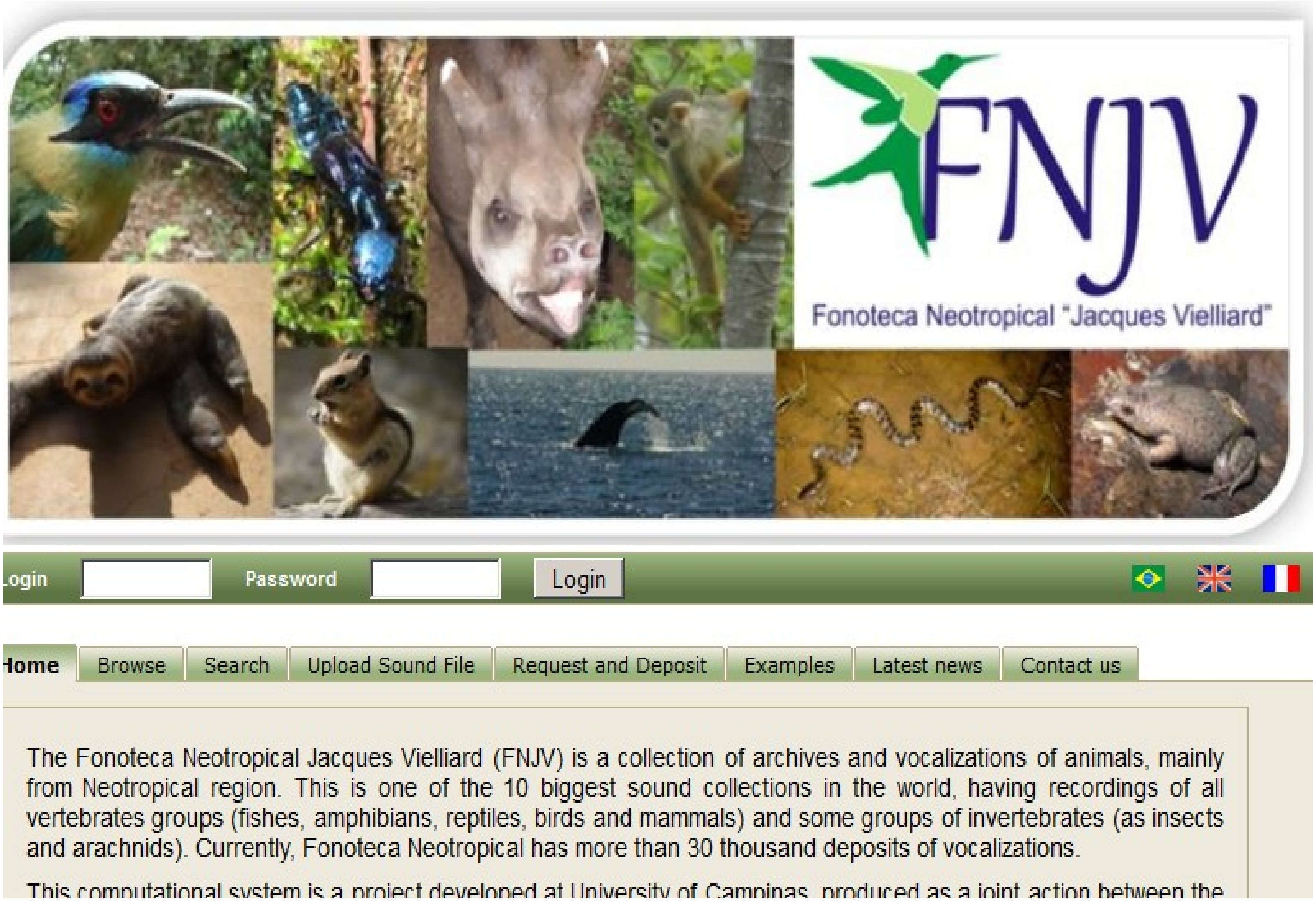
Specialist B

Quality assessment

- Fitness for use
- Dimensions?
  - Timeliness, accuracy, reputation...
- Model – provenance or attribute-based?
- Platform?

=> Provenance inducing quality assessment

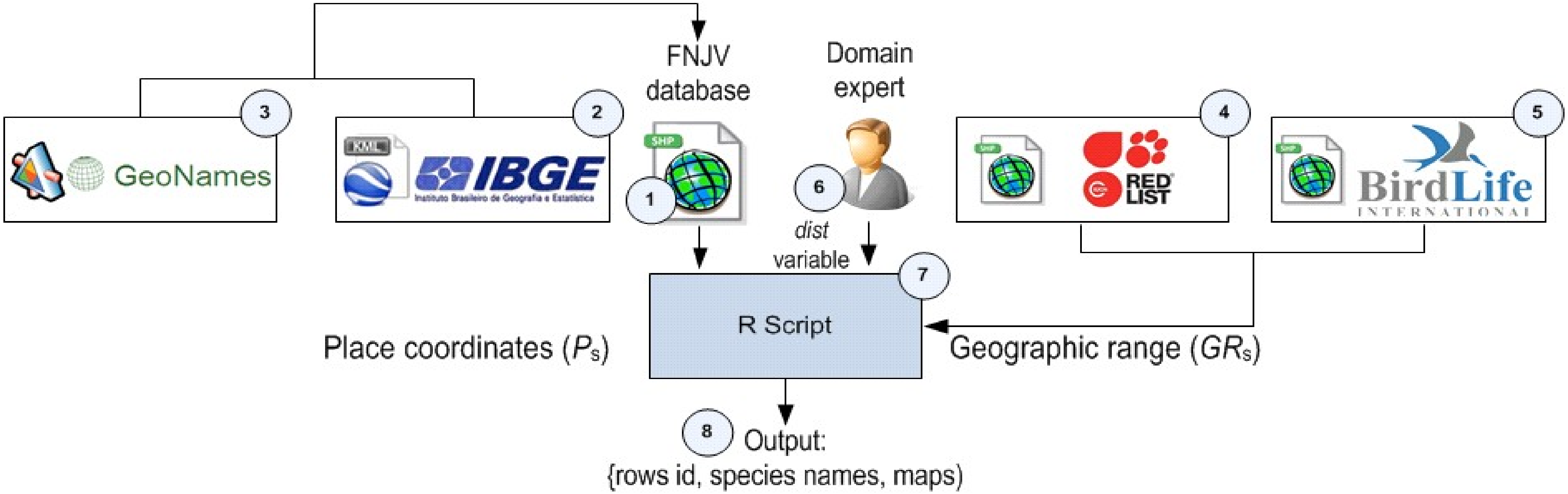
Case study - FNJV



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.

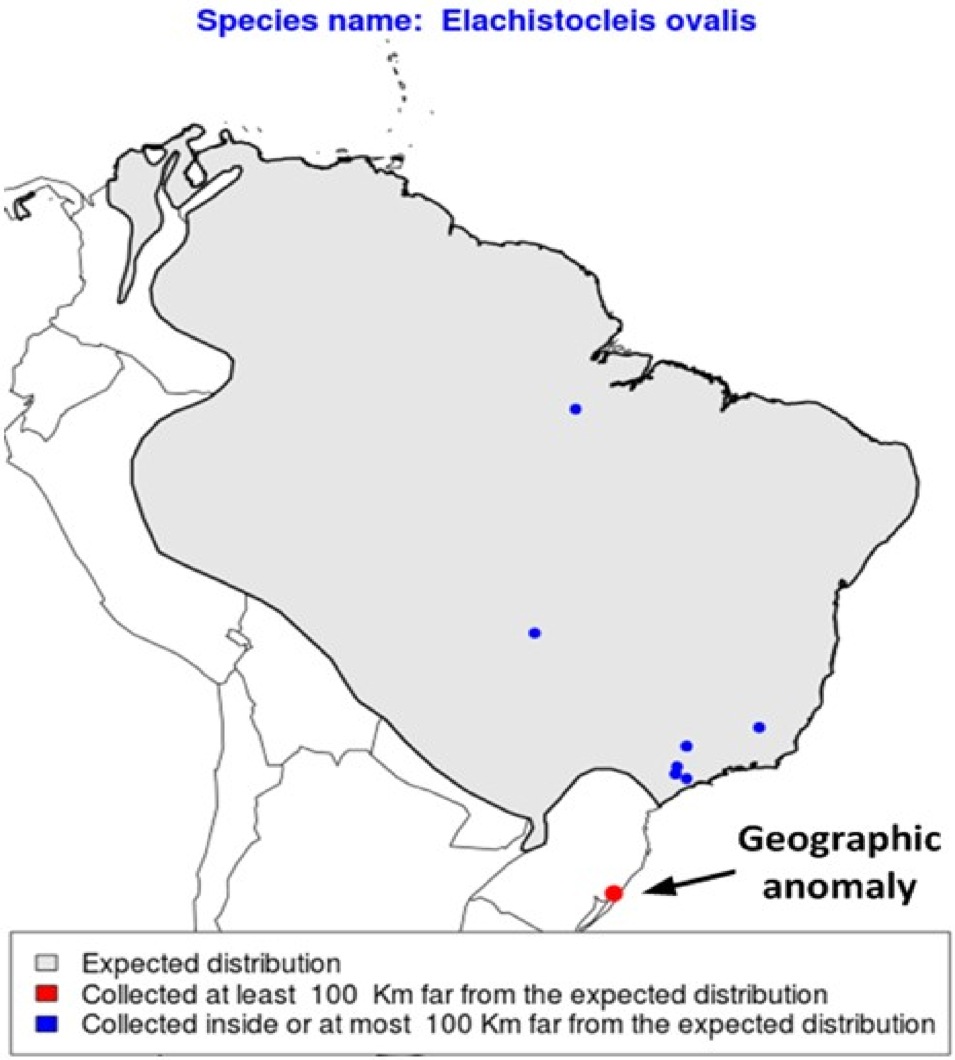
This computational system is a project developed at University of Campinas, produced as a joint action between the

Case Study - Prototype

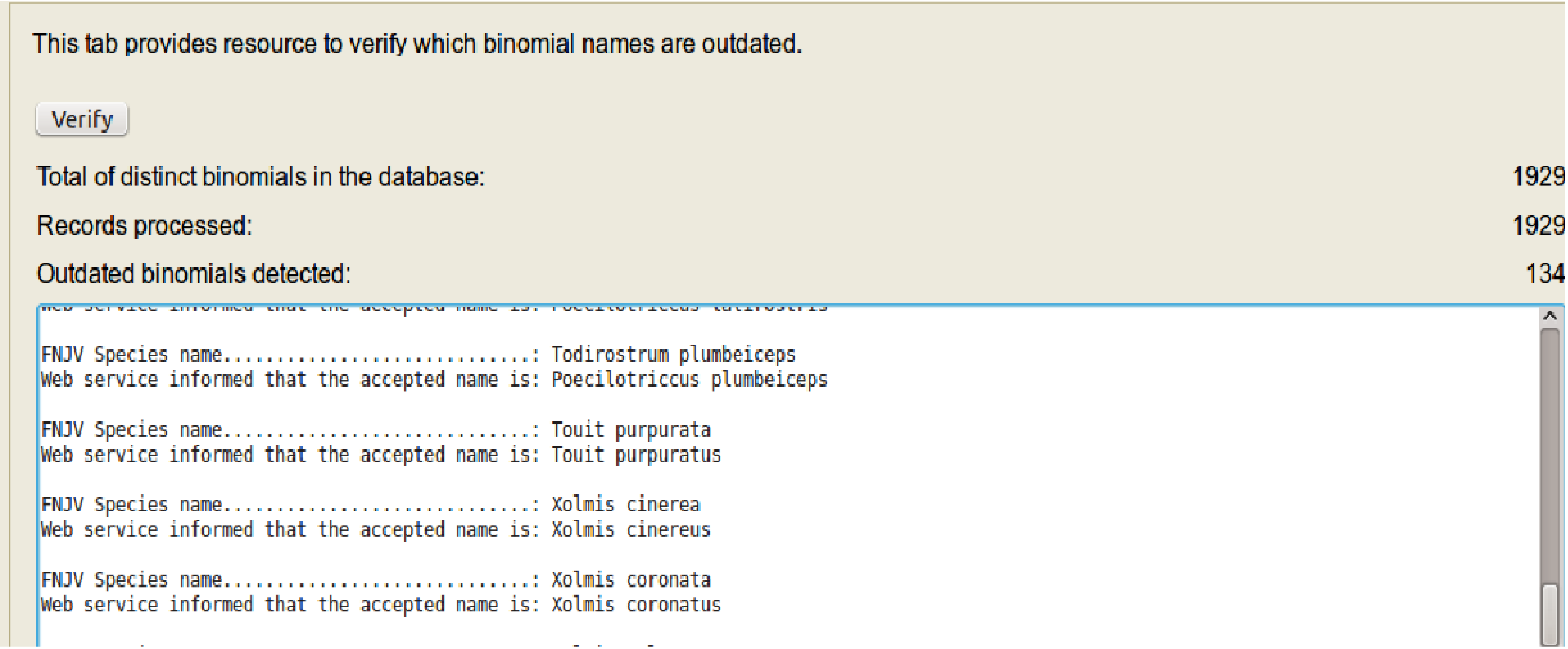

FNJV database
Domain expert
3
GeoNames
2
KML
IBGE
Instituto Brasileiro de Geografia e Estatística
SHP
1
6
4
SHP
RED LIST
5
SHP
BirdLife INTERNATIONAL
dist variable
7
R Script
Place coordinates (P_s)
Geographic range (GR_s)
8
Output:
{rows id, species names, maps)

And outdated species names

Anomalous places (32% - 12K records)


Species name: Elachistocleis ovalis
Geographic anomaly
Expected distribution
Collected at least 100 Km far from the expected distribution
Collected inside or at most 100 Km far from the expected distribution

Outdated names – 7% (134 species)



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

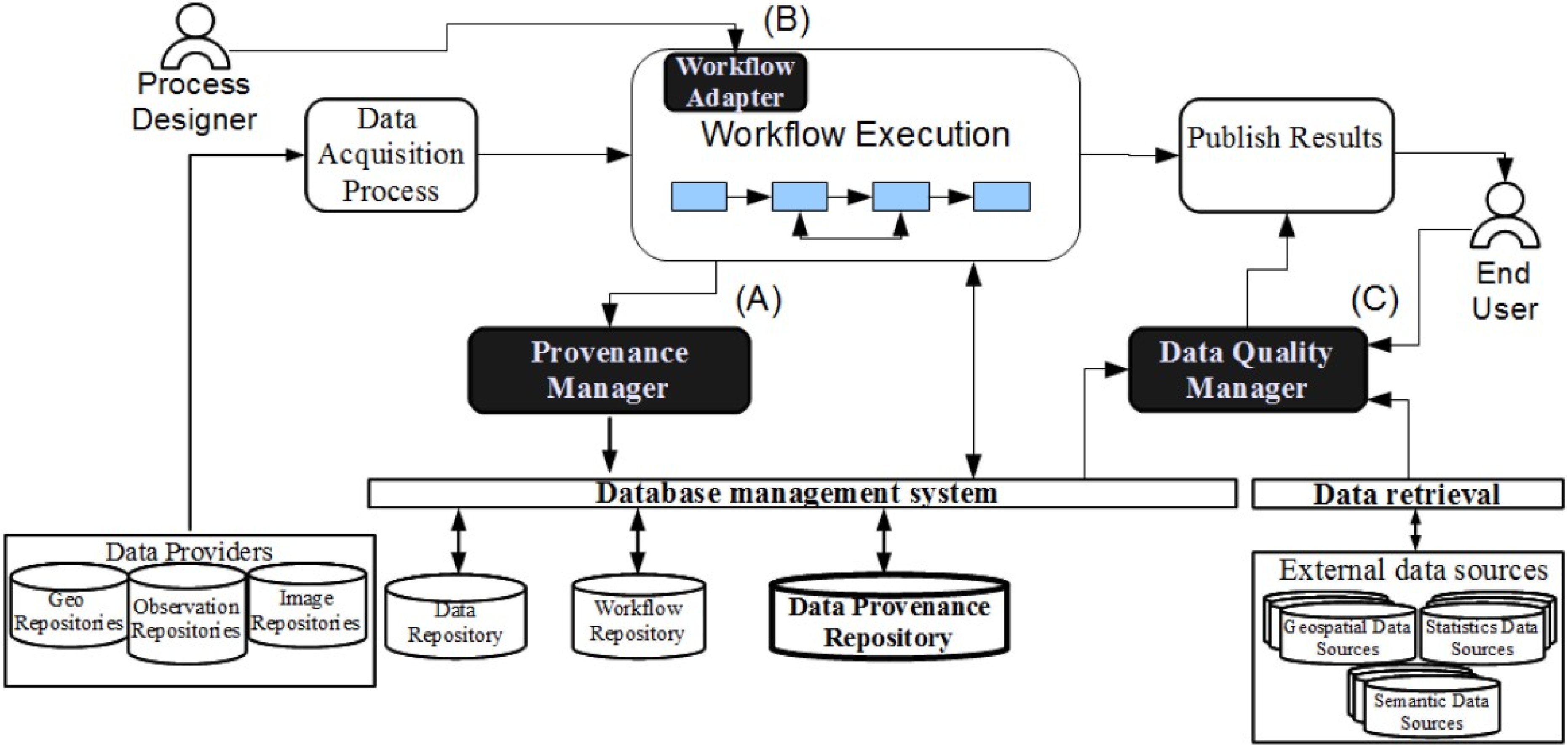
```
Web service informed that the accepted name is: Poecilotriccus latirostris

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
```

Quality-aware workflows

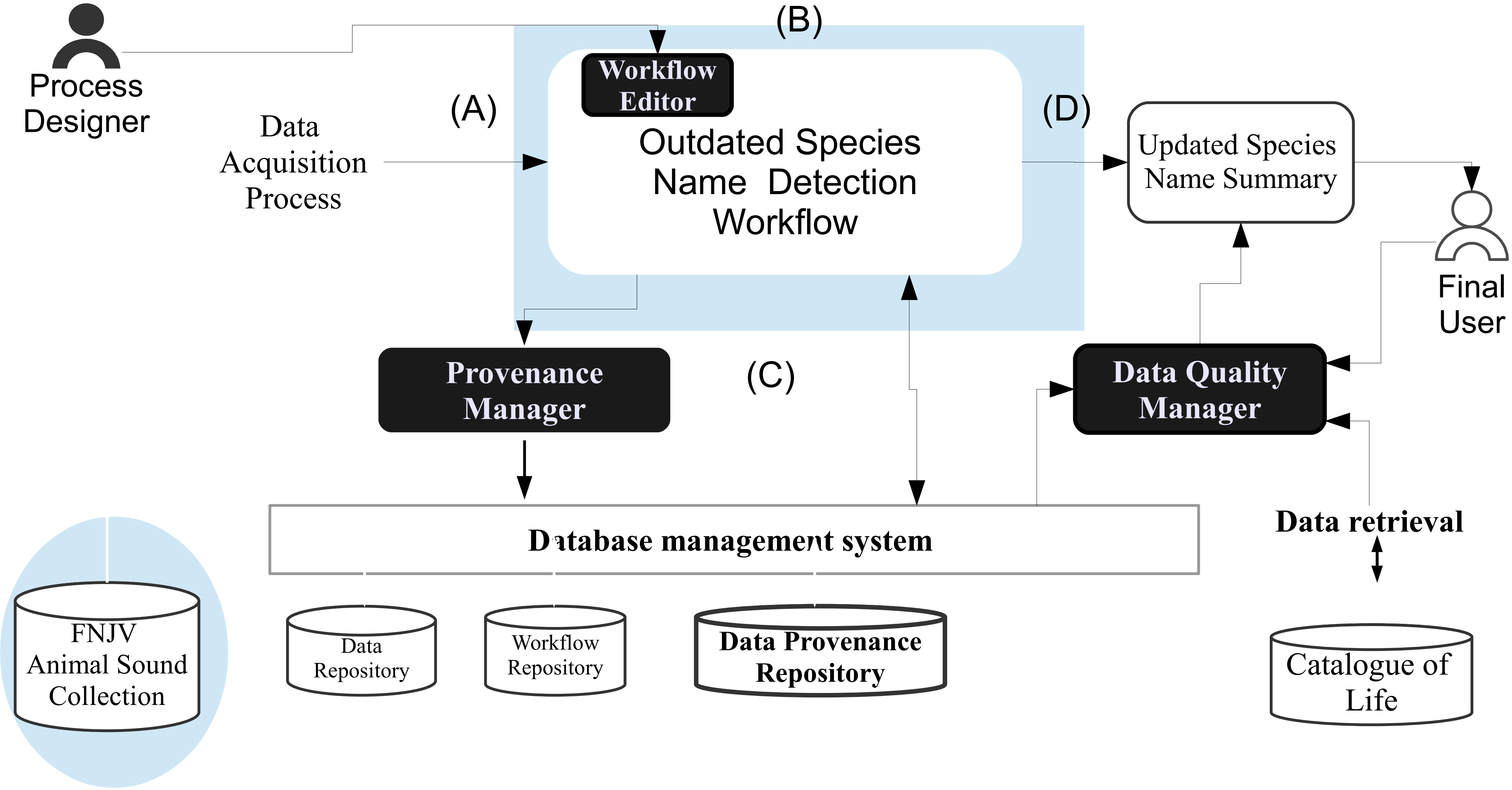
- Add quality dimensions to processes
- Add quality dimensions to data
- Provenance information from execution

- Data quality + process quality = final quality

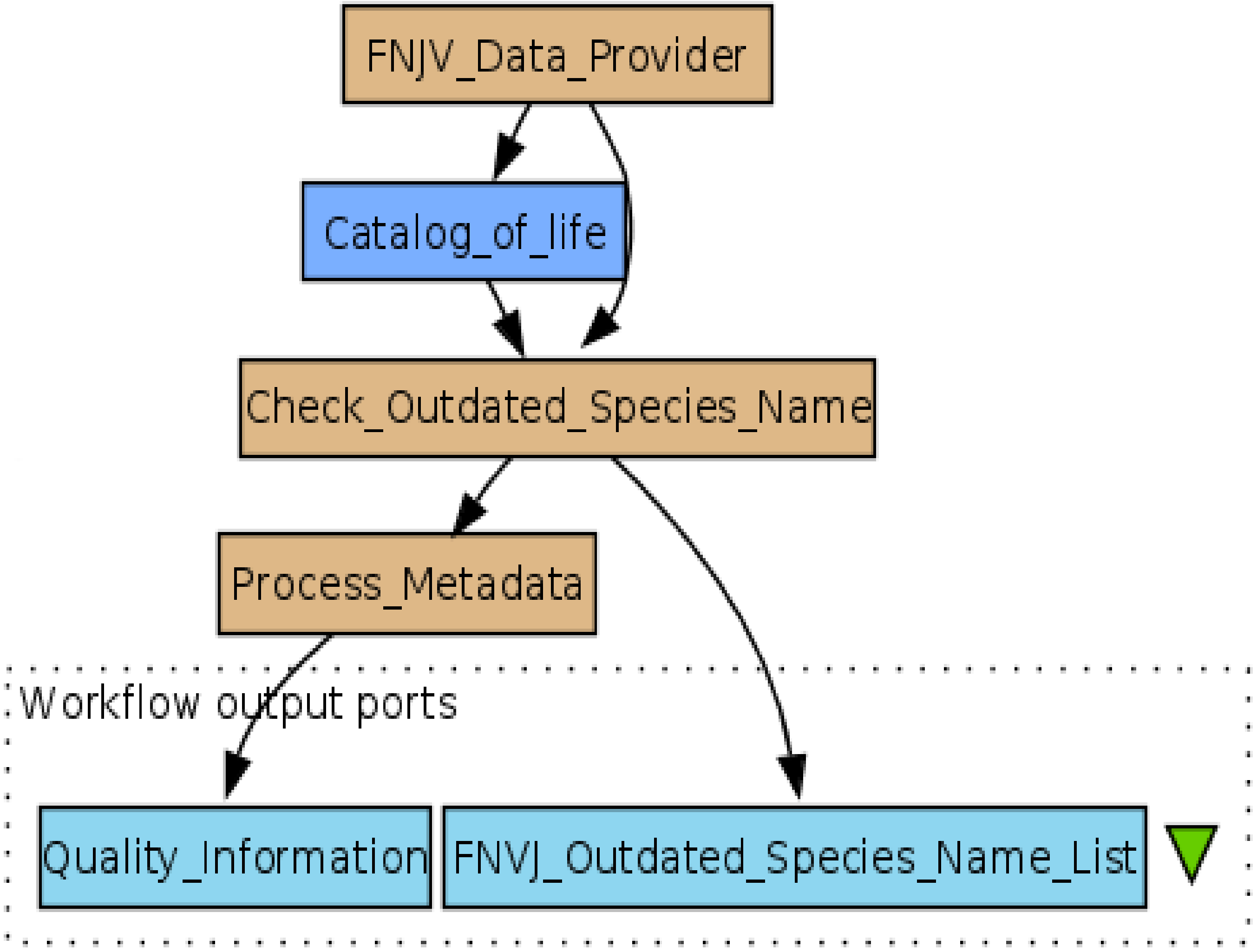
Quality aware workflows



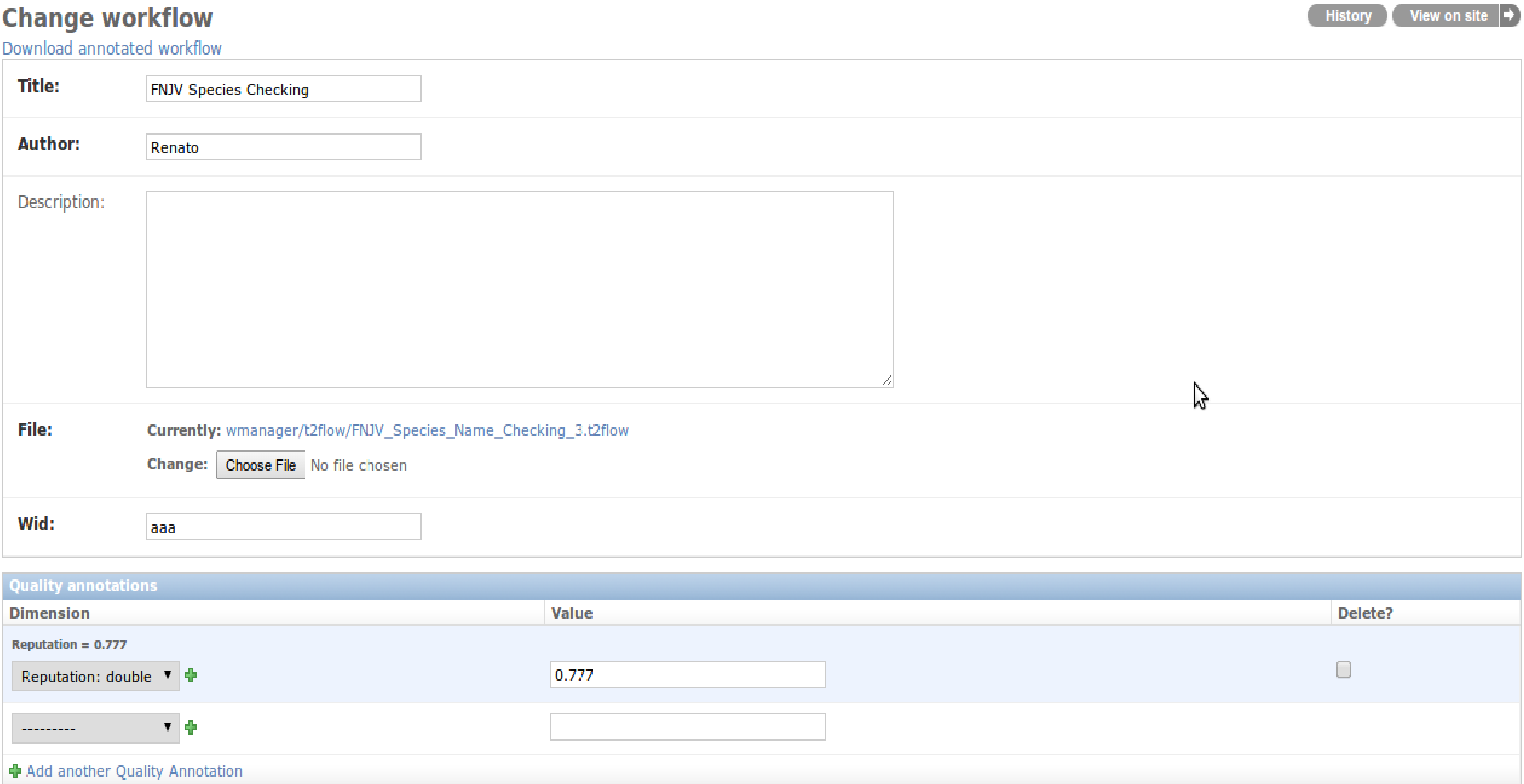
Case Study - Find out the accuracy of the original metadata



Quality processing



Workflow 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

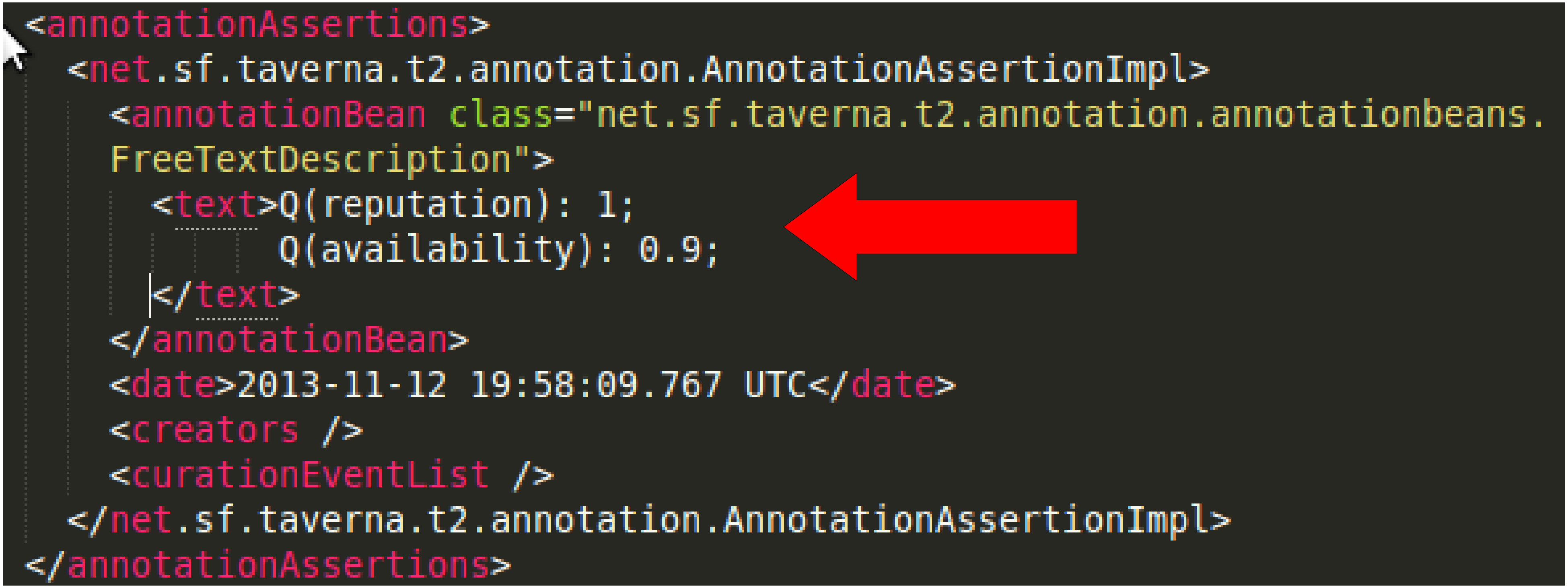
Quality annotations

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

Add another Quality Annotation

Workflow Adapter

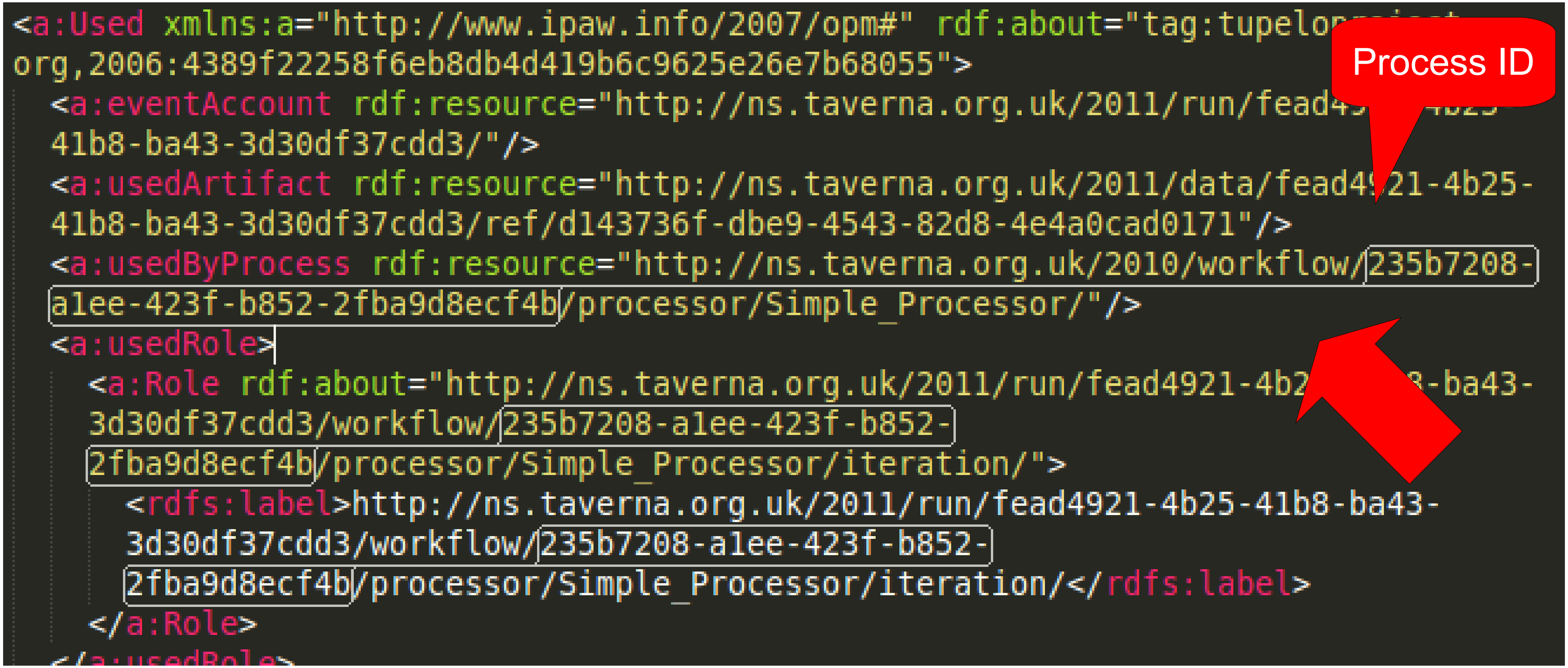
- Adds quality information to a workflow specification
- No changes to workflow model



```
<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

- Extract provenance information from metadata and workflow specification



```
<a:Used xmlns:a="http://www.ipaw.info/2007/opm#" rdf:about="tag:tupeloproject.
org,2006:4389f22258f6eb8db4d419b6c9625e26e7b68055">
  <a:eventAccount rdf:resource="http://ns.taverna.org.uk/2011/run/fead4921-4b25-
  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/fead4921-4b25-41b8-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>
```

Process ID

Quality Manager

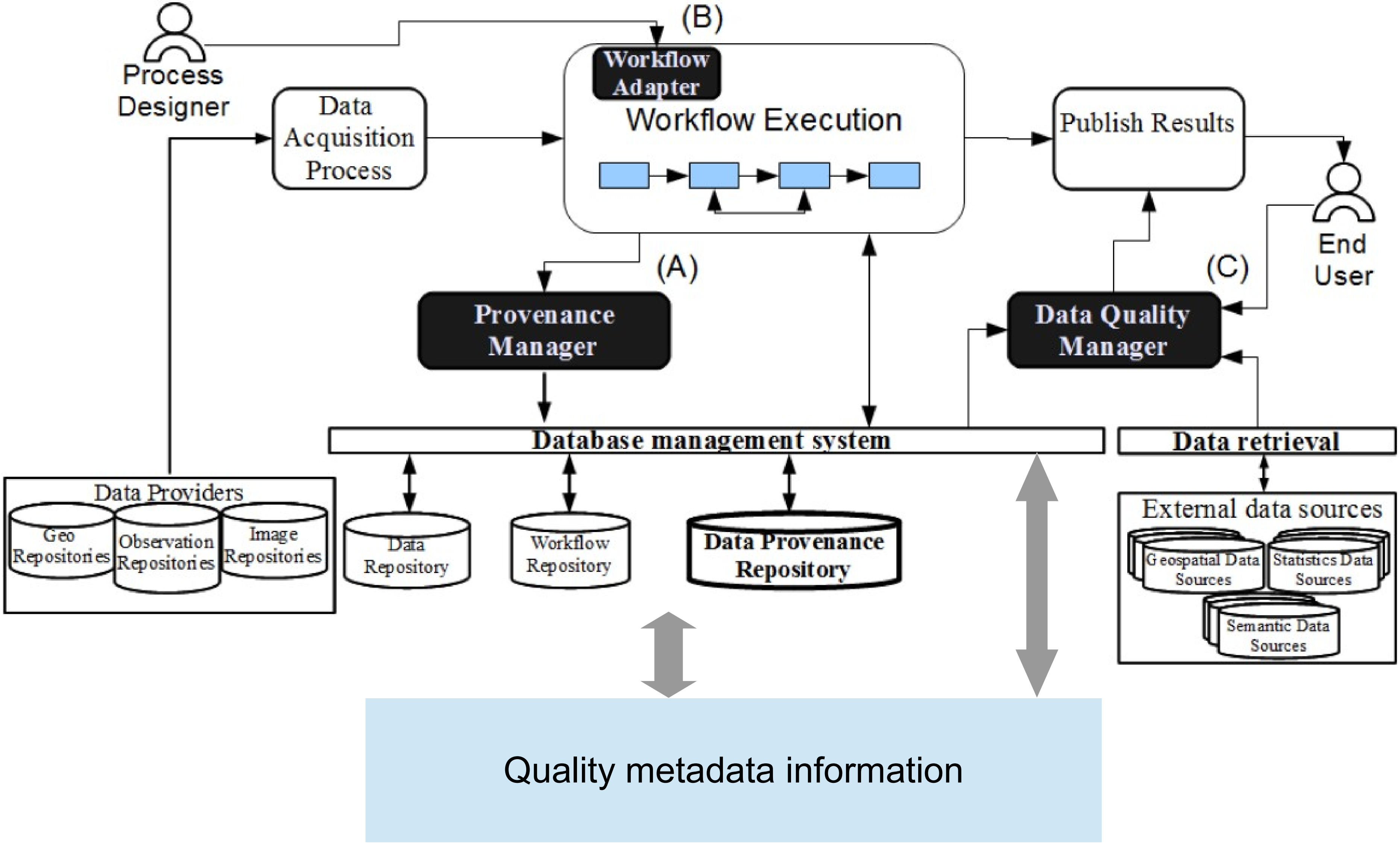
- Data quality assessment
  - From provenance
  - From annotations – quality attributes generated by Workflow Adapter
  - External data sources

Conclusions

- Preservation requires curation
- Quality-aware workflows
  - Induce curation
  - Provide user-dependent quality assessment
  -

    BUT
- Quality = eye of the beholder --- how to account for varying quality ?

Ongoing – fusion of quality metadata



(B)

Process Designer

Workflow Adapter

Workflow Execution

Data Acquisition Process

Publish Results

End User

(A)

(C)

Provenance Manager

Data Quality Manager

Database management system

Data retrieval

Data Providers

Geo Repositories

Observation Repositories

Image Repositories

Data Repository

Workflow Repository

Data Provenance Repository

External data sources

Geospatial Data Sources

Statistics Data Sources

Semantic Data Sources

Quality metadata information

Ongoing Work

- Design of quality repository (“meta”quality)

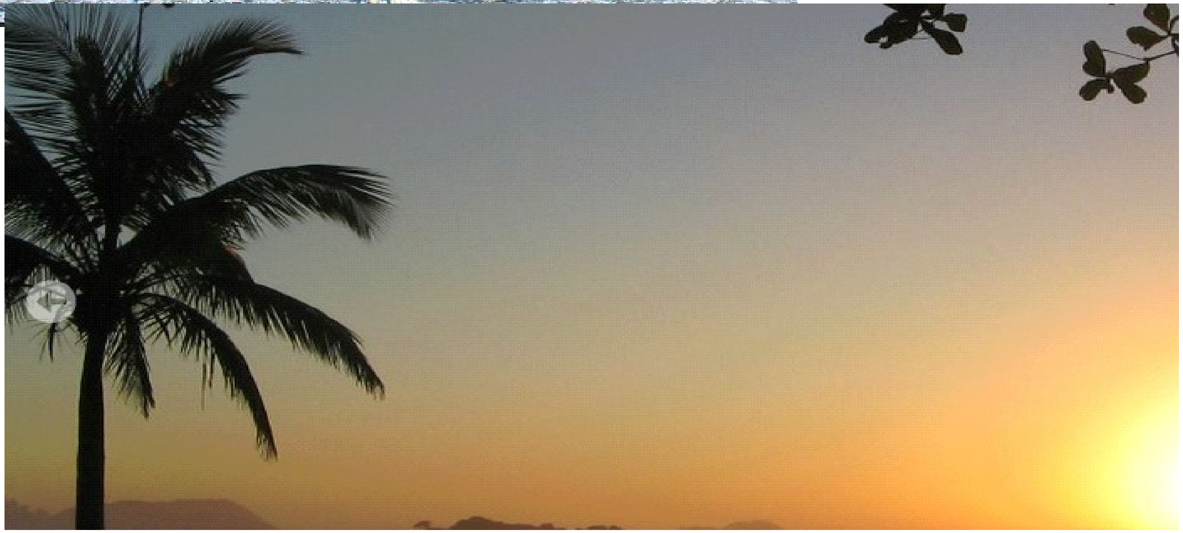
- Implementation as Web tool

- Integration with Linked Data

Acknowledgements

- FAPESP/Cepid in Computational Engineering and Sciences
- FAPESP grants (2011/19284-3),(2013/08293-7)
- Microsoft Research FAPESP Virtual Institute (NavScales project)
- CNPq (MuZOO Project)
- FAPESP-PRONEX (eScience project)
- INCT in Web Science
- CNPq
- We thank Prof. Omar Boucelma from Univ. Aix-Marseille for his valuable suggestions.

IEEE eScience 2014



IEEE eScience 2014

GUARUJA,

BRAZIL