



Workflow preservation

Julian Garrido
Instituto de Astrofísica de Andalucía – CSIC
& WF4EVER team

SKA-LINK (Granada)
4th April 2017

Outline

- » Main concepts and objectives
- » Workflow-based Science
- » Technology for Scientific Workflow Preservation
- » Workflow Preservation in Astronomy

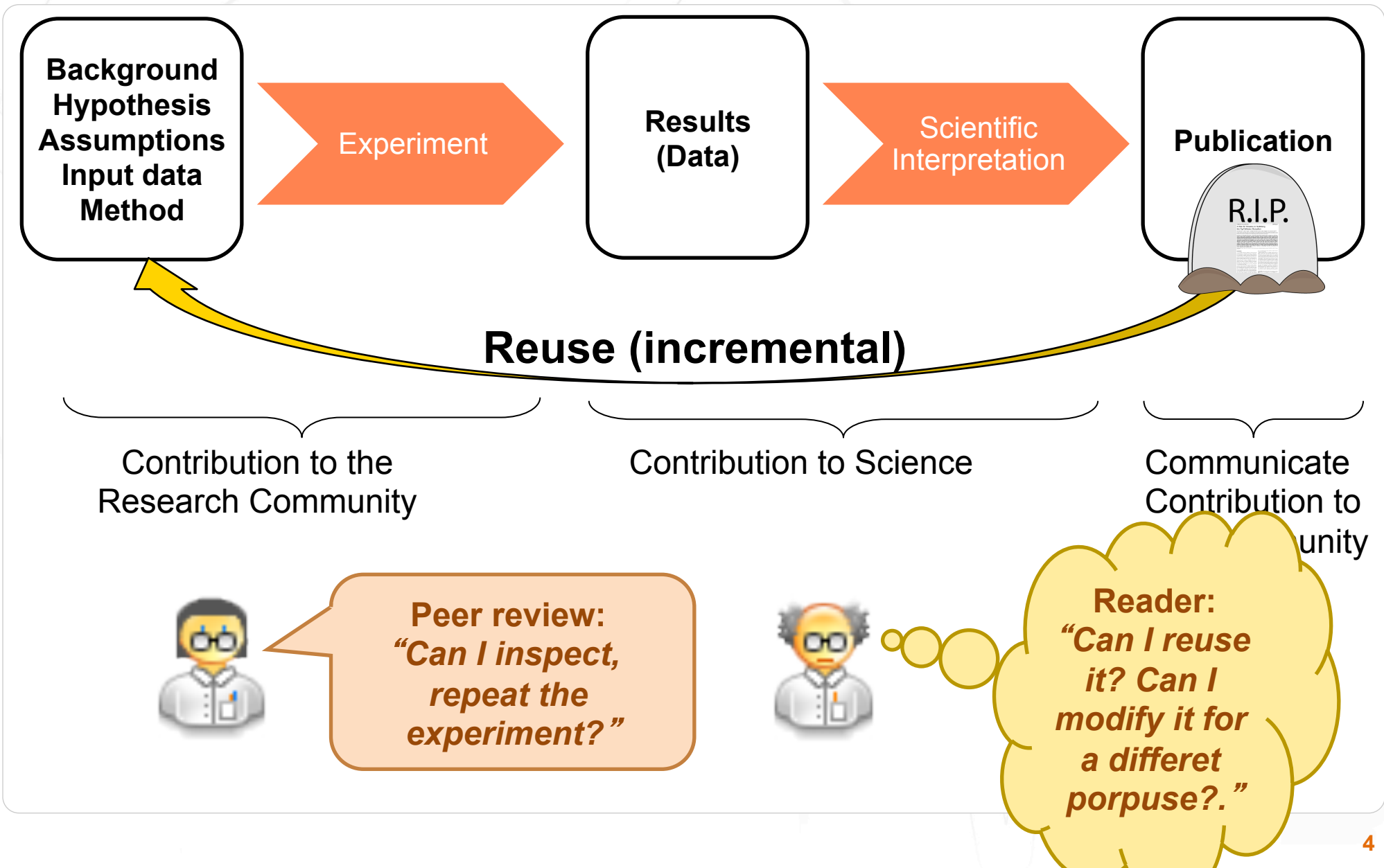
Wf4Ever

Advanced Workflow Preservation Technologies for Enhanced Science

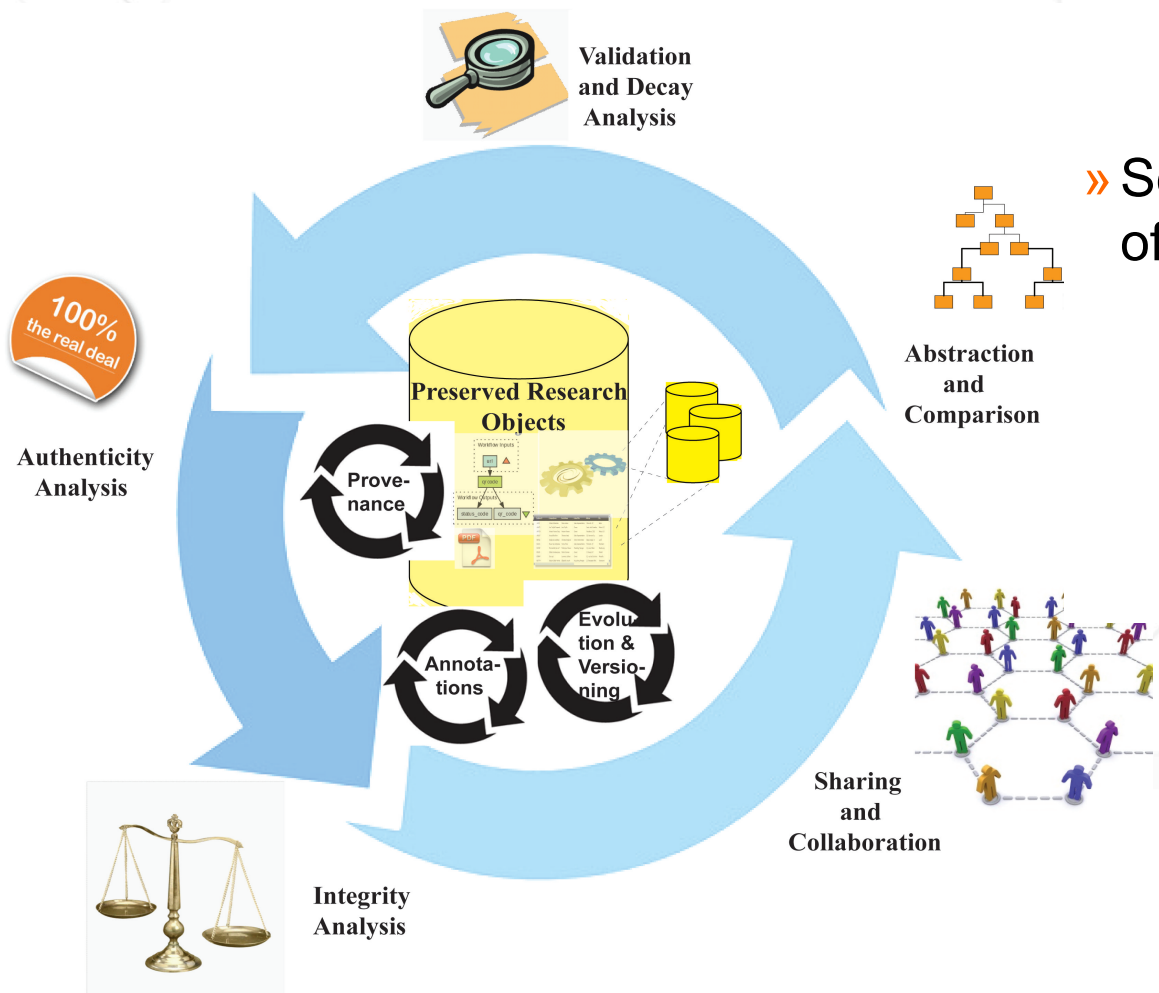


1. Intelligent Software Components (ISOCO, Spain)
2. University of Manchester (UNIMAN, UK)
3. Universidad Politécnica de Madrid (UPM, Spain)
4. Poznan Supercomputing and Networking Centre (Poland)
5. University of Oxford and OeRC (OXF, UK)
6. Instituto Astrofísica Andalucía (IAA-CSIC, Spain)
7. Leiden University Medical Centre (LUMC, NL)





Preservation of scientific workflows in data-intensive science



» What is a workflow?

- » A mechanism for coordinating the execution of services and codes, and linking together resources.

» Scientific workflows are at the heart of experimental science

- » Enable automation of scientific methods
- » Encourage best practices
- » End user oriented
- » Workflows as means to describe, re-run and reuse scientific methods
- » Support experimental reproducibility

Questions for Workflows	Issues
Who are you ? Where and when were you born ? Who were your parents (creators) ?	Description Authenticity Uniqueness
For which purpose were you conceived and you have been used ?	Re-use, re-purpose
What do you have inside ?	Inspection Visualization Annotations
How is your content linked ?	Graphical Representation
May I access all your parts ?	Access Rights
Which parts can I replace ?	Adaptability
What have they done to you ? Who and When ? Why did they do that ?	RO Provenance Versioning Annotations
Why have you been recommended to me ? Can I believe what you are saying or trust your results ?	Information Quality Data Provenance

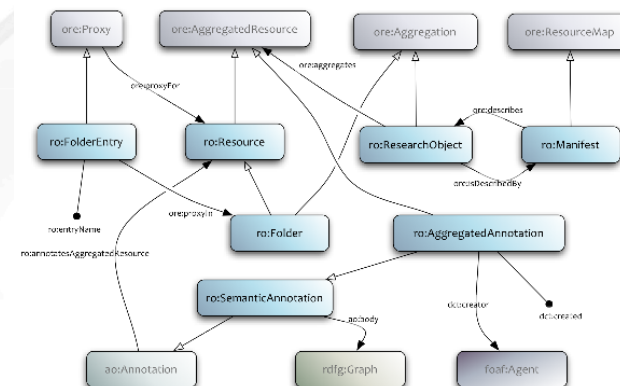
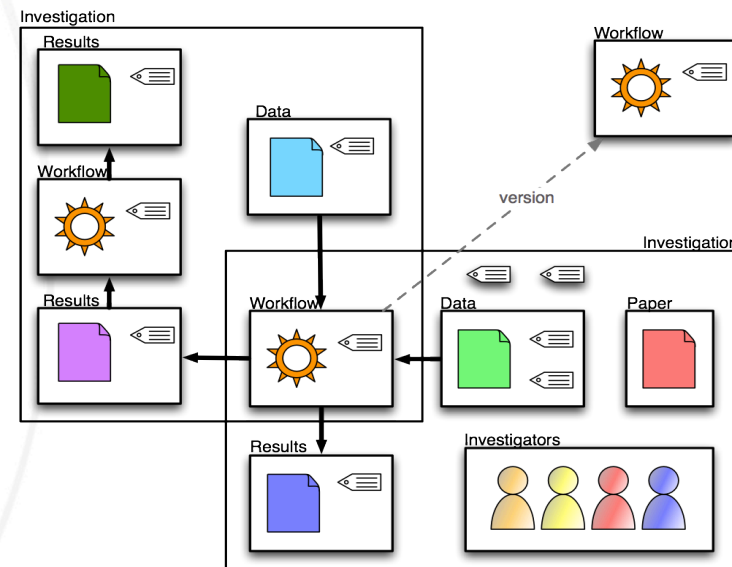
**Artifact
Instrument
to Curate
Preserve,
Conserve
Reuse**

**Communication
Record
to Package,
Exchange, Share,
Publish, Find,
Explore, Inspect,
Review**

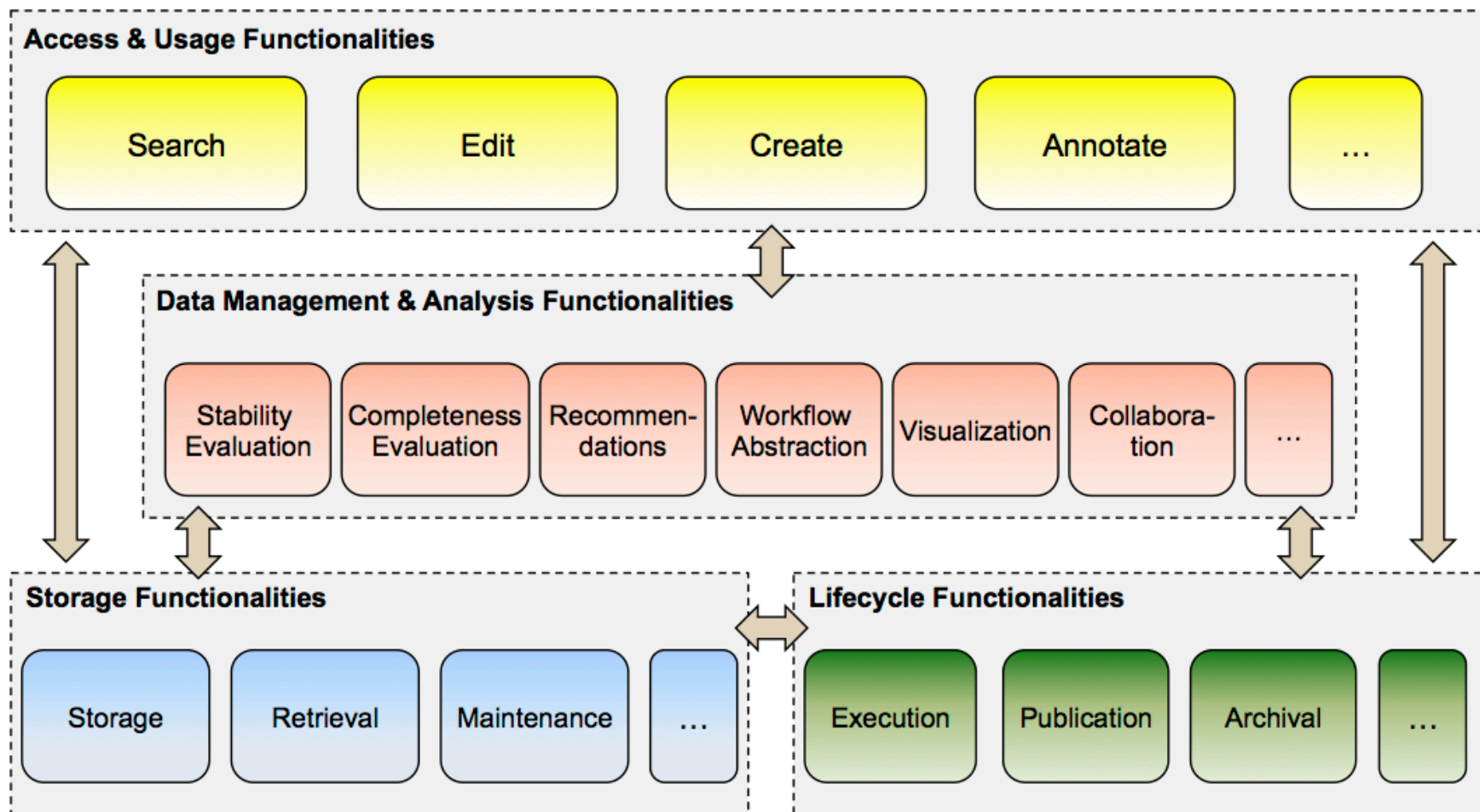
Research Objects (ROs)

- Semantically rich aggregations of resources that bring together data, methods and people in scientific investigations. Their goal is to create a class of artifacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge.
- Workflow-centric RO can be viewed as an aggregation of resources that bundles a workflow specification and additional auxiliary resources, including documents, input and output data, annotations, provenance traces of past executions of the workflow, etc.
- Target: reusability, reproducibility and better understanding

<http://www.researchobject.org/specifications/>

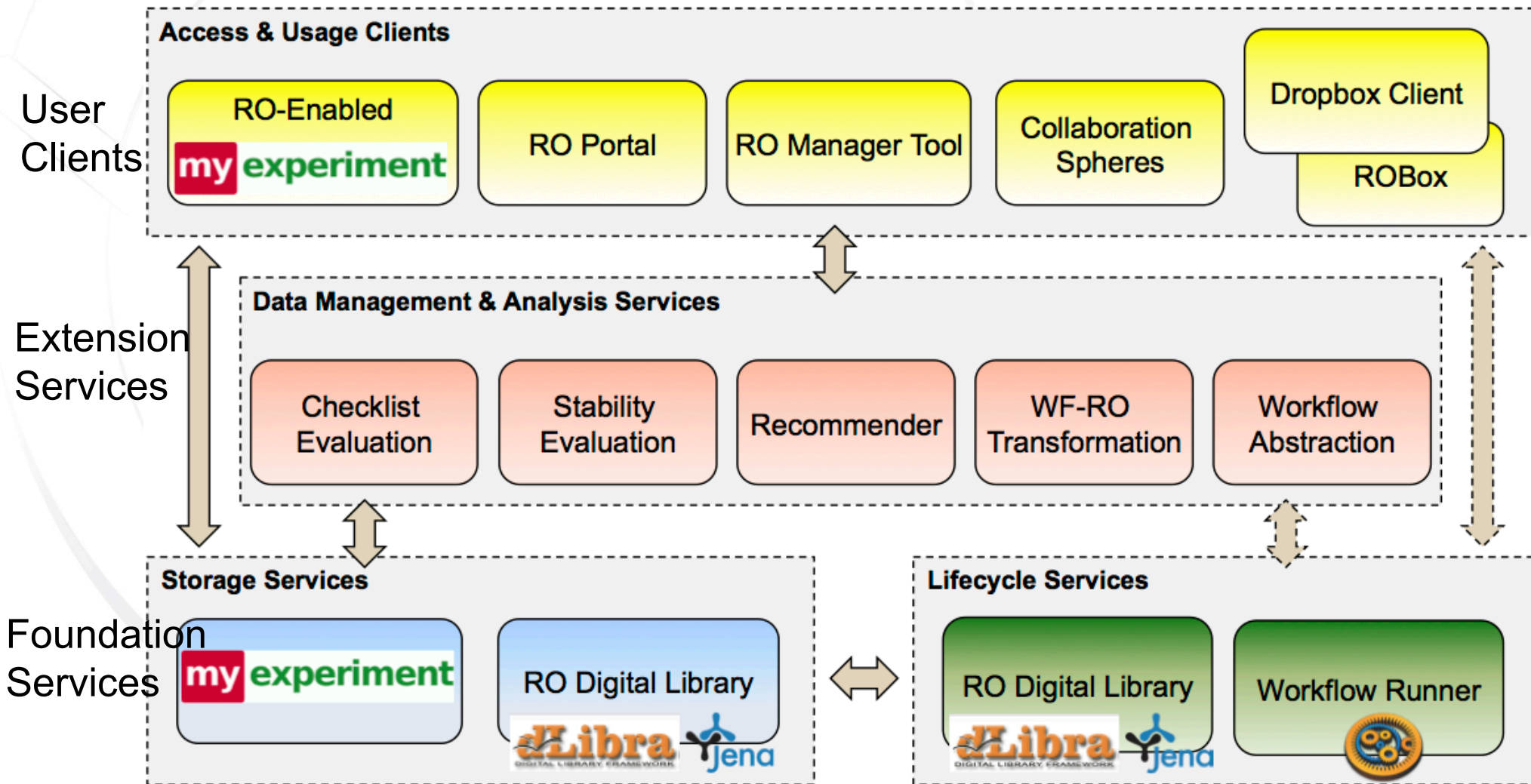


How to enable reproducibility?



Page, Palma, et al. **From workflows to Research Objects: an architecture for preserving the semantics of science**. Linked Science 2012. Boston, USA.



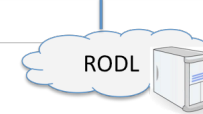
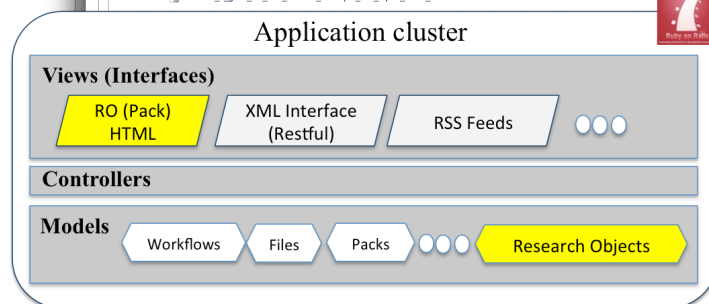
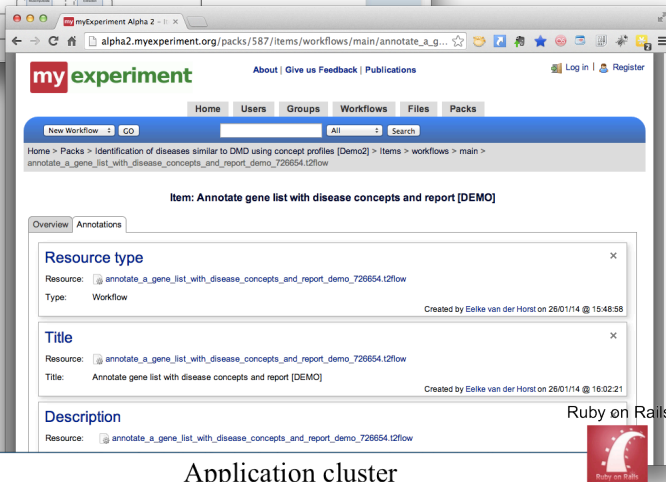
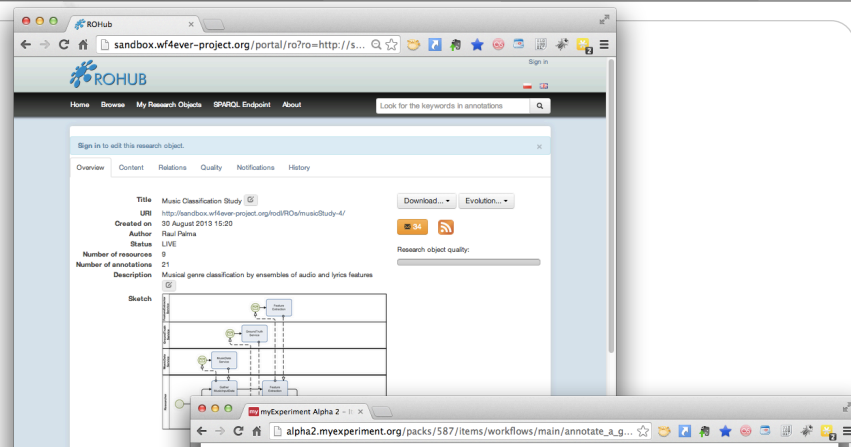


» ROHub

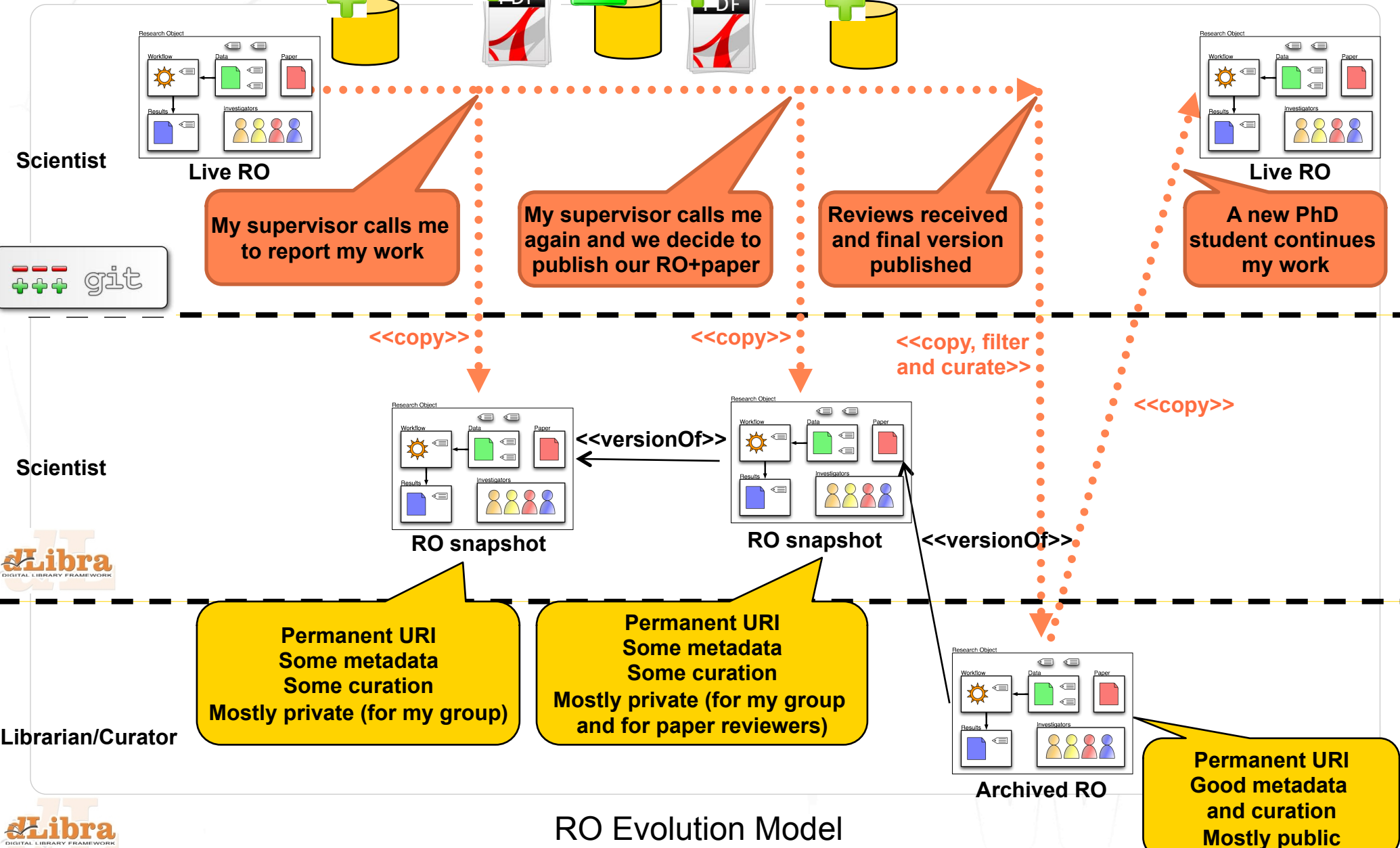
- » RO API
- » RO Evolution API
- » SPARQL Query of metadata
- » Notifications
- » Long term preservation via dArceo backend
 - Decay alerts, fixity checking
 - Checklist monitoring

» RO Enabled myExperiment

- » RO model for descriptions of content
- » ROHub/myExperiment communication through ROs and RO Bundles
- » RO presentation/browsing
- » Checklist Service Integration



Workflow Evolution, Sharing and Collaboration



RO evolution

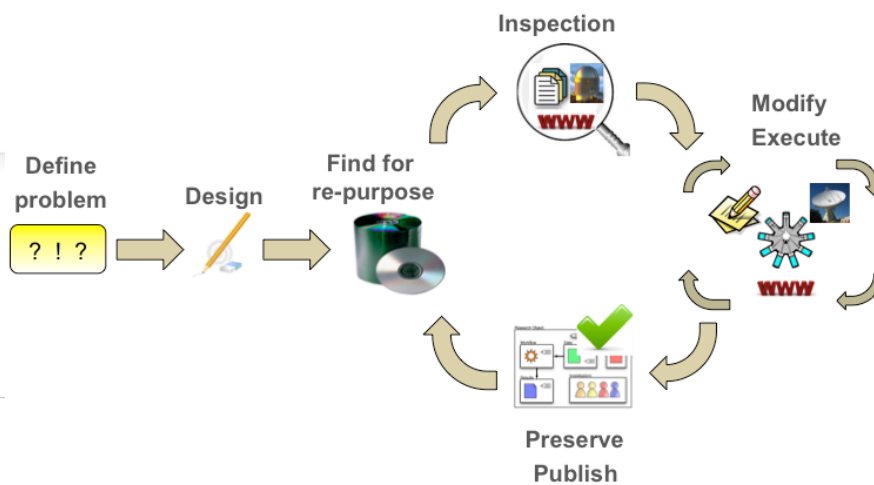
Live



Snapshots

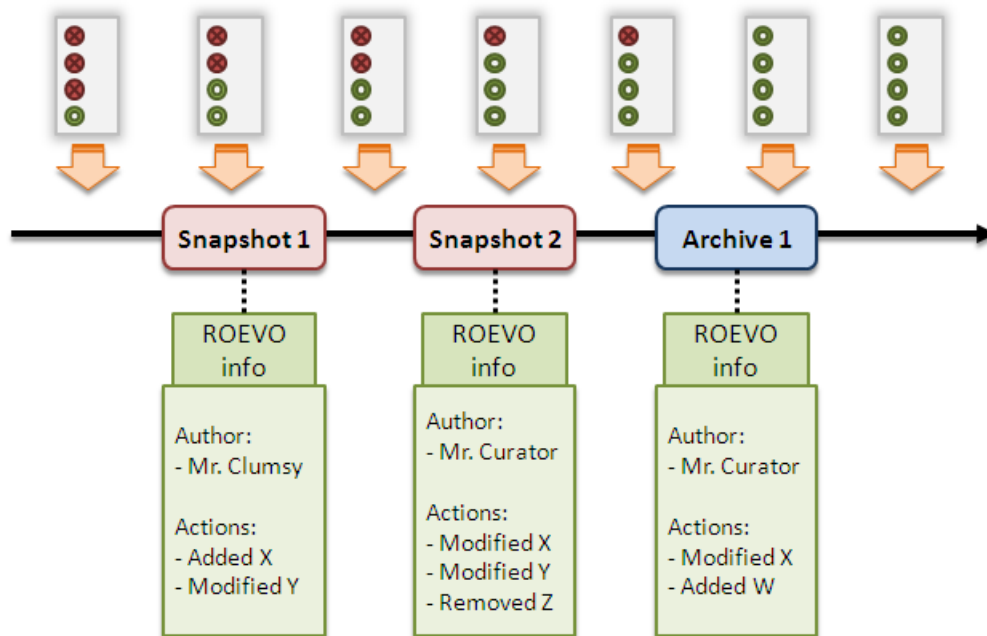


Archived



Stability Assessment:

- » Computes how RO quality evolves in time based on existing checklist
- » Allows comparing the quality of different RO snapshots based on ROEVO traces
- » Analytics and explanations of quality evolution
- » Services for gathering evaluations and evolution of ROs in order to provide information to end-users (e.g. for curation purposes)



» Checklist service:

- › detecting observed causes of decay in workflows
- › SPARQL endpoint
- › New Minim Info designs adapted to applications:
 - Detection of workflow decay
 - Completeness assessment for workflow decay prevention
 - Completeness assessment of resource descriptions
 - Supports stability/reliability assessment

» Completeness:

- › check fulfilment of requirements specified in checklists for evaluating the quality of a RO a purpose
- › Minim model to meet additional requirements for validation
- › Traffic light display of checklist results

GWAS to pathway

This pack is for a workflow that finds KEGG pathways for genes from a GWAS.

✗ Target [Pack384](#) does not satisfy checklist for ready-to-release.

- ✓ Experiment hypothesis is present
- ✓ Workflow design sketch is present
- ✓ All workflow definitions are accessible
- ✗ One or more web services used by one of the workflows are inaccessible, including <http://rest.kegg.jp/get/{query}>
- ✓ Input data is present
- ✓ Experiment conclusions are present

[Wf4Ever project](#)

Sign in to edit this research object.

Overview | Content | Relations | Quality | Notifications | History

Title	Not set
URI	http://sandbox.wf4ever-project.org/rodl/ROs/Pack559/
Created on	08 January 2014 16:19
Author	Unknown
Status	LIVE
Number of resources	15
Number of annotations	27
Description	Not set
Sketch	No image available

Basic view | Import | Annotate

The advanced annotations view

type	http://purl.org/wf4ever/roevo#
type	http://purl.org/wf4ever/ro#ResearchObject

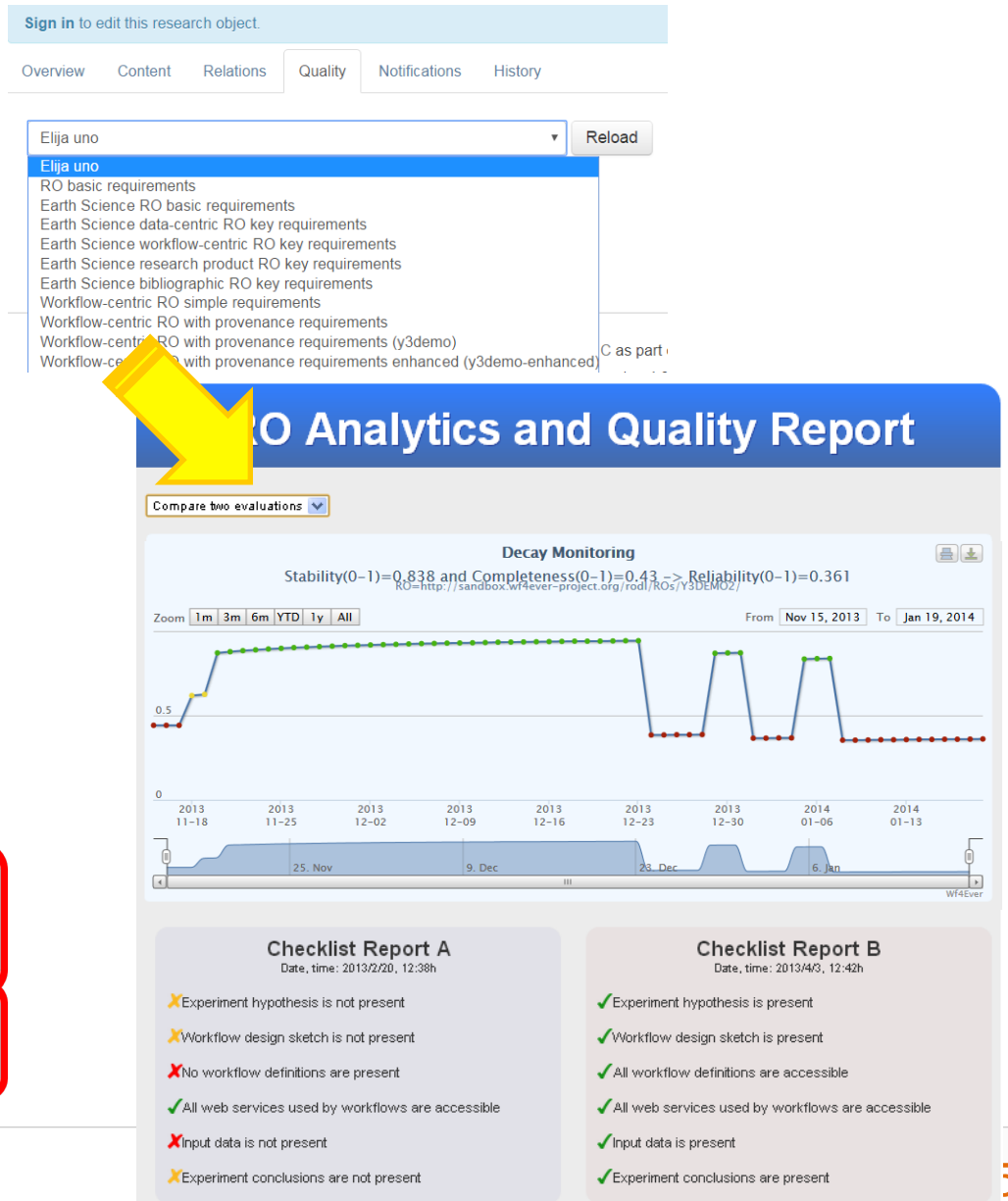
Research object quality:

Details

- ✓ Research Object title is present
- ✗ Research Object description is not present
- ✓ Experiment hypothesis or research question is present
- ✓ Experiment design sketch is present
- ✓ Experiment conclusions are present
- ✓ Annotations bodies are all accessible

Quality Service:

- » Measurement of the quality and decay of ROs by the completeness and stability assessments
- » Completeness provides a measure of the overall status of a research object at an specific time for a purpose
- » Stability provides a measure of the overall status of a research object throughout its whole lifecycle



The screenshot shows a web interface for a research object named 'Eljja uno'. It features a navigation menu with 'Quality' selected. A dropdown menu lists various requirements like 'RO basic requirements' and 'Earth Science RO basic requirements'. A yellow arrow points to the 'Quality Analytics and Quality Report' section.

Quality Analytics and Quality Report

Stability(0-1)=0.838 and Completeness(0-1)=0.43 -> Reliability(0-1)=0.361

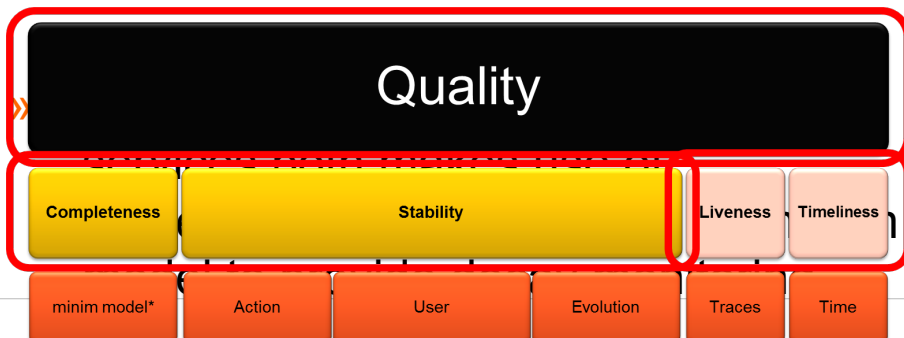
Decay Monitoring chart showing stability over time from 2013 to 2014. The chart shows a high stability level (around 0.8) until late 2013, followed by a sharp drop to near zero, with some subsequent fluctuations.

Checklist Report A (Date: 2013/2/20, 12:38h):

- ✗ Experiment hypothesis is not present
- ✗ Workflow design sketch is not present
- ✗ No workflow definitions are present
- ✓ All web services used by workflows are accessible
- ✗ Input data is not present
- ✗ Experiment conclusions are not present

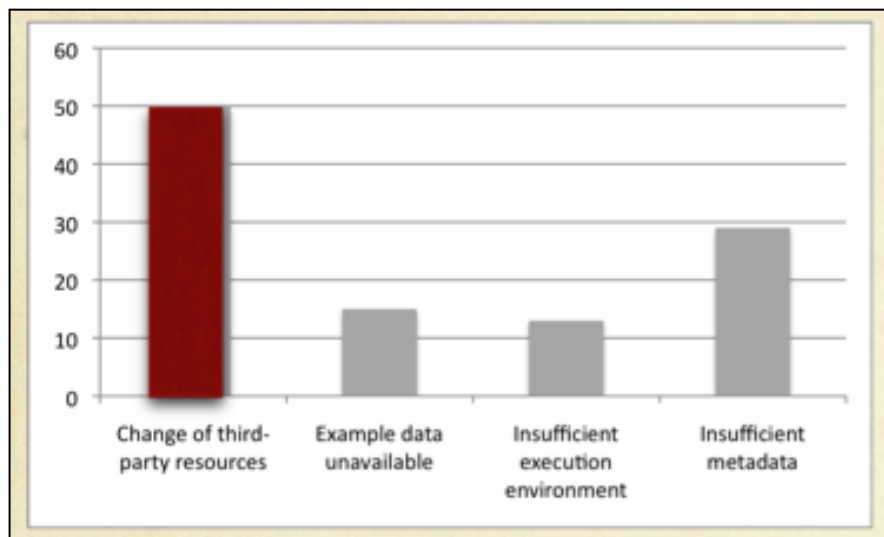
Checklist Report B (Date: 2013/4/3, 12:42h):

- ✓ Experiment hypothesis is present
- ✓ Workflow design sketch is present
- ✓ All workflow definitions are accessible
- ✓ All web services used by workflows are accessible
- ✓ Input data is present
- ✓ Experiment conclusions are present



*Adaptation of the MiM model created by Matt Gamble from University of Manchester, U.K.

- » Systematically selected a set of samples of real Taverna workflows from myExperiment to determine if they suffer from decay and the reasons that caused their decay
- » Four main categories of causes of decay:
 - » Missing example data
 - » Missing execution environment
 - » Insufficient descriptions about workflows
 - » Volatile third-party resources (most common)





Workflow Decay

- Component level
 - flux/decay/unavailability
- Data level
 - formats/ids/standards
- Infrastructure level
 - platform/resources

Experiment Decay

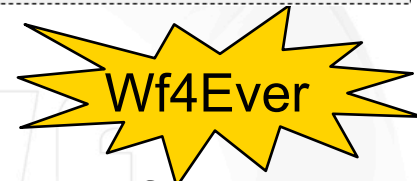
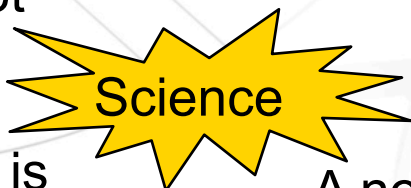
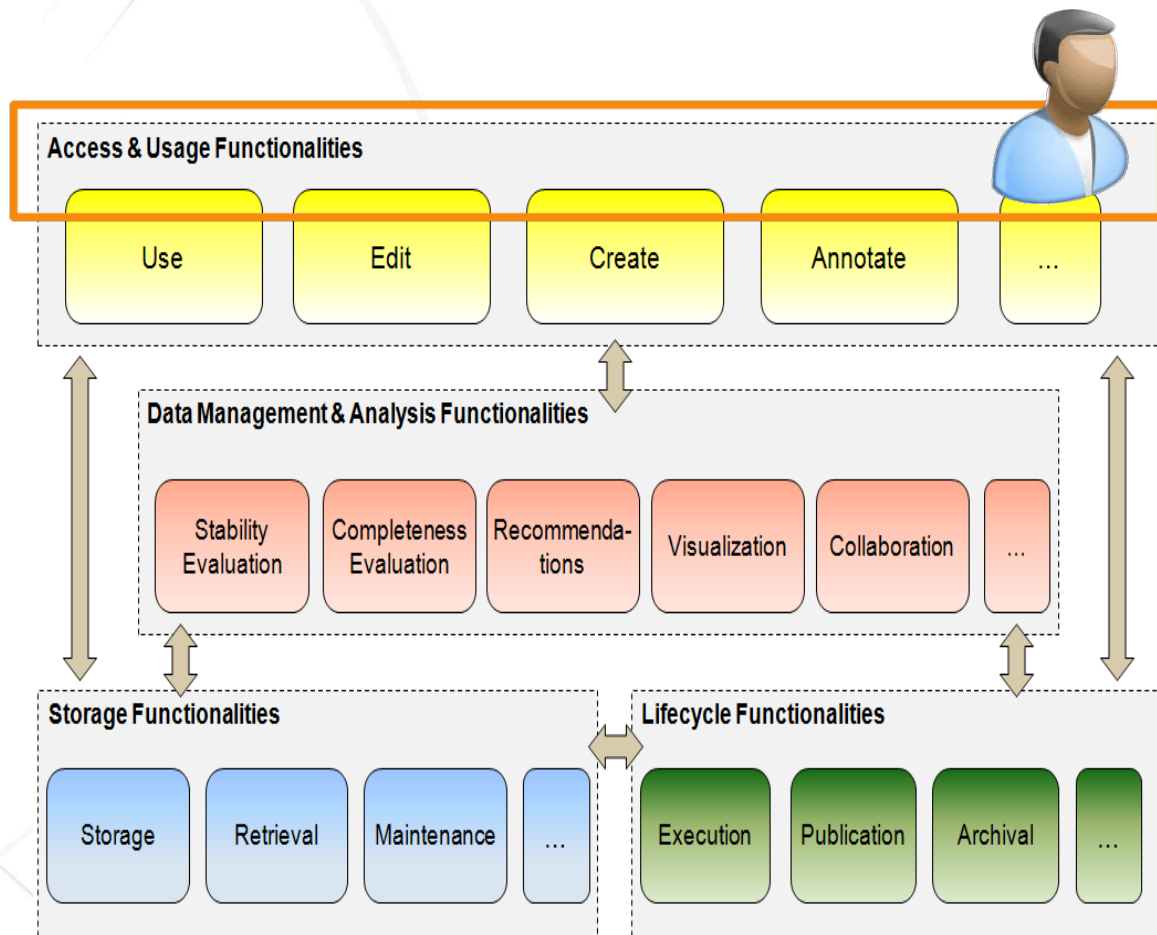
- Methodological changes
- New technologies
- New resources/components
- New data

Astronomy research lifecycle is **entirely digital**

- » Observation proposals 
- » Data reduction pipelines
- » Catalogues of objects
- » Analysis of science ready data
- » Publish process
 - › Final data results
 - › Experiment in DL 

Reproducible research is still not possible in a digital world

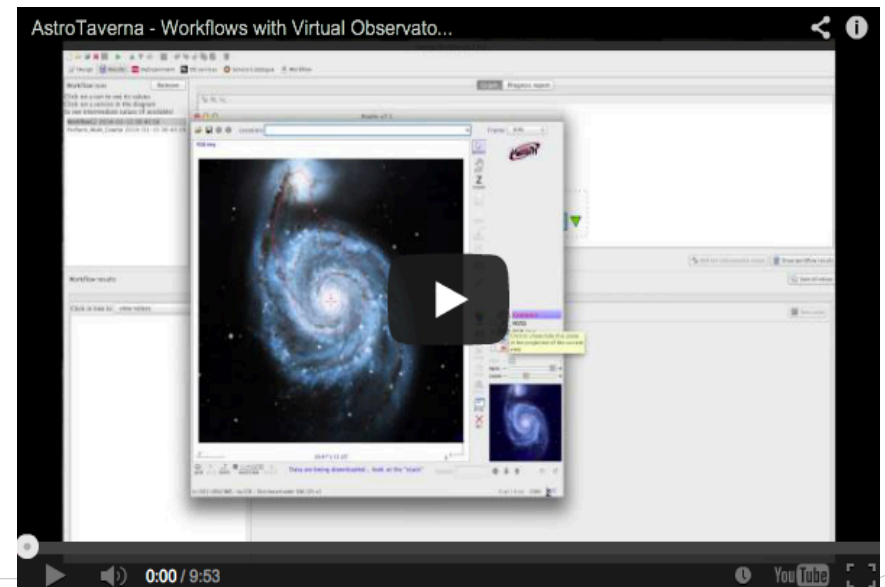
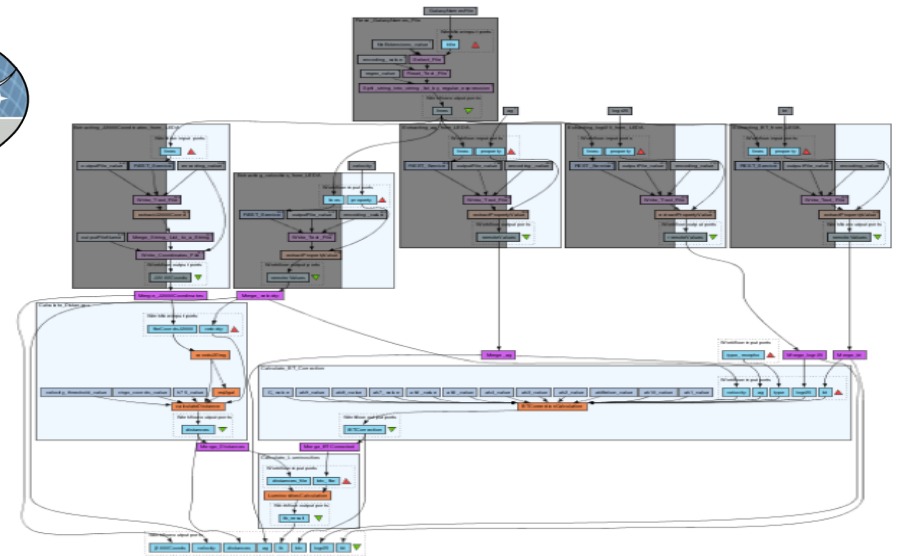
A rich infrastructure of data (VO) is not efficiently used



A normalized preservation of methodology is needed

Main Functionalities

- » Integration with VO Software
 - › Message exchanging (SAMP)
 - › Aladin user interactive execution
 - › Seamless data exploration
- » VO Registry Search
- » VO Services Orchestration
- » VOTable Rendering
- » VOTable Manipulation
- » Coordinates conversion
- » Resolve object names into coordinates
- » Access to PDL Services
- » Access to TAP Services
- » Access to Starter Pack



- › Best Practices – How to prevent workflow decay
 - › Make an abstract workflow
 - › Use modules
 - › Think about the output
 - › Provide input and output examples
 - › Annotate
 - › Make it executable from outside the local environment
 - › Choose services carefully
 - › Reuse existing workflows
 - › Test and validate
 - › Advertise and maintain

• Hettne *et al.* Best Practices for workflow design: how to prevent workflow decay

- › Astronomy research is entirely digital. Time has come to go “Beyond the PDF”
- › Don't publish. Release!
- › Digital Libraries, repository, SGs are key to use existing/coming infrastructure of data and computation
- › Methods evolve. Data changes. Metadata changes. Services get replaced. Platforms break. Stuff gets versioned. Things need repair.
- › The next generation of archives: service providers

Thanks for your attention!

- Carol Goble
- Sean Bechhofer
- Stian Soiland-Reyes
- Jose Enrique Ruiz del Mazo
- Marco Roos
- David De Roure
- Raul Palma
- Kristina Hettne
- Khalid Belhajjame
- Daniel Garijo
- José Manuel Gómez
- Lourdes Verdes-Montenegro
-

- Wf4ever team

Julián Garrido
IAA-CSIC
jgarrido@iaa.es

