Grant agreement no.: 27092

Workflows for Methodology and Science Preservation

Juan de Dios Santander Vela

On behalf of L. Verdes-Montenegro, J.E. Ruiz, S. Sánchez, and the Wf4Ever collaboration

European Southern Observatory, ALMA Archive Subsystem

Workflow 4Ever

Grant agreement no.: 27092

Workflows for Methodology and Science Preservation

Juan de Dios Santander Vela

On behalf of L. Verdes-Montenegro, J.E. Ruiz, S. Sánchez, and the Wf4Ever collaboration

Instituto de Astrofísica de Andalucía-CSIC, AMIGA Group (January 2012)

Who am I?

- Ph.D. within AMIGA group on making radio astronomical archives and tools work with the Virtual Observatory
- Applied Scientist at ESO VLT Archive, specialised in metadata management
- Currently working on the ALMA Science Archive, from the backend to the web GUI.
- From January 2012, working for the Wf4Ever project in bringing radio astronomical workflows to life.

- AMIGA: Analysis of the Interstellar Medium of isolated GAlaxies
  - Multi-wavelength, multi-object study on isolated galaxies with strict isolation criteria
  - Careful curation of data
  - Very careful processing of new parameters from
    - Group's own observation programs and data reduction
    - Literature table scanning
    - Virtual Observatory table harvesting and parsing
  - Emphasis on marrying astronomy and computer science, and buy-in of the VO

e-Science believers!

What is Wf4Ever?

EU funded FP7 STREP Project
December 2010 – December 2013

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 (PSNC, Poland)
5. University of Oxford (OXF, UK)
6. Instituto de Astrofísica de Andalucía (IAA, Spain)
7. Leiden University Medical Centre (LUMC, NL)

What is Wf4Ever?

Technological infrastructure for the **preservation** and **efficient retrieval** and **reuse** of scientific workflows **in a range of disciplines**

Partners

- One SME
- Six public organisations

Core Competencies (Tech)

- Digital Libraries
- Workflow Management
- Semantic Web
- Integrity & Authenticity
- Provenance
- Information Quality

Case Studies

- Astronomy (IAA)
- Genome-wide Analysis and Biobanking

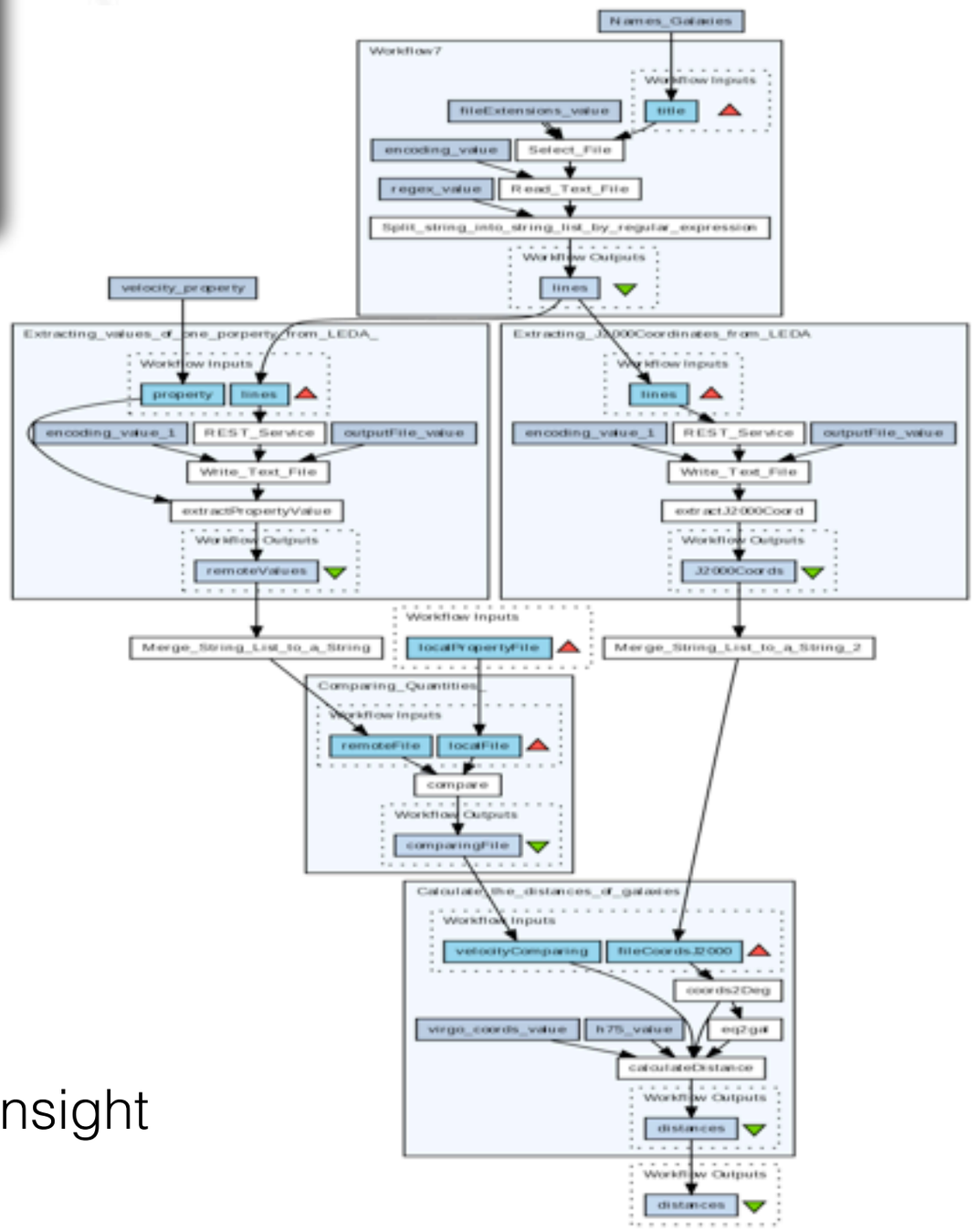
Goals

Archival, classification, and indexing of **scientific workflows** and their associated materials in scalable semantic repositories, providing advanced access and recommendation capabilities

Creation of scientific communities to collaboratively share, reuse, and evolve workflows and their parts, stimulating the development of new scientific knowledge

What are workflows?

Combination of **data and processes** into a configurable and **structured set of steps** that implement semi-automated, **problem solving**, computational solutions



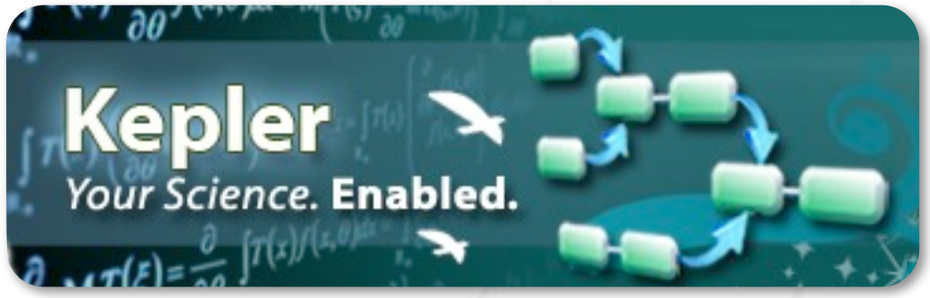
- Types of workflows in Astronomy
  - Personal script-based recipes
  - Internal group developments*
  - Multi-archive VO experiments
  - The classical processing pipeline*
  - Driving pipelines from VO services (TBD)

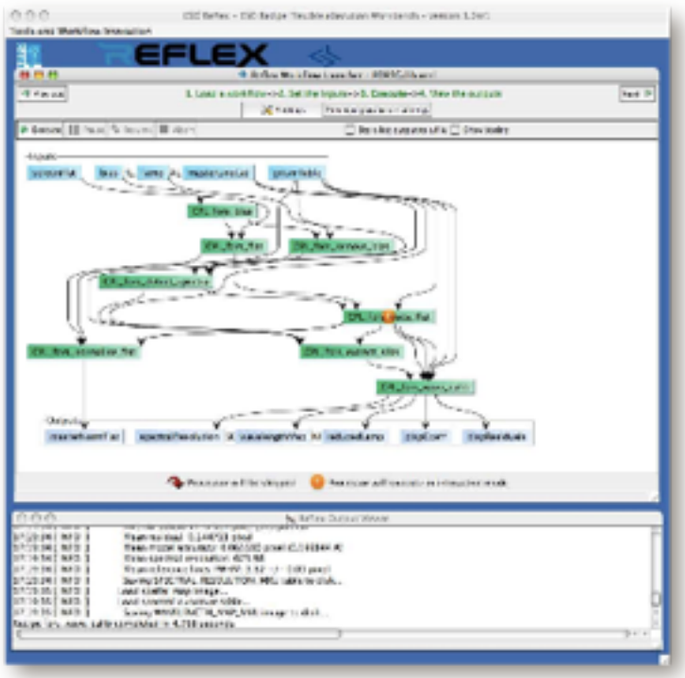
* Scientifically exploitable results vs. scientific insight

Easily accessible and reproducible

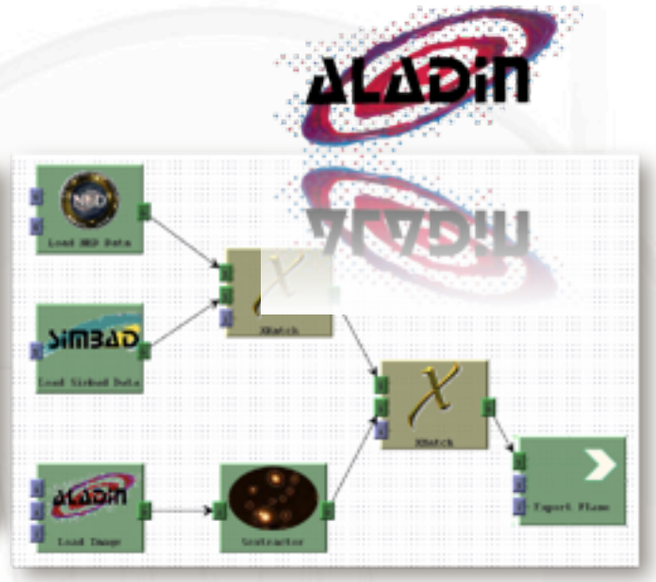
Workflow 4Ever

What tools are available?





```
#AJS
get NED m13 14'
get Simbad m13 14'
sync
xmatch NED Simbad 4
sync
get Aladin(2MASS,J) m13
sync
get SExtractor(2MASS.J)
sync
xmatch "XMatch*1" "SExtractor*"
sync
export -votable "XMatch*2"
```



What tools are available?

Combination of **data and processes** into a configurable and **structured set of steps** that implement semi-automated, **problem solving**, computational solutions

Astronomy research is entirely digital: time to go "beyond the PDF"

- Preserved experiments
  - Methodology "in action"
  - All data are exposed
  - Reproducible
  - Repeatable
  - Re-usable
  - Re-purposeable
  - Participatory
  - Collaborative
  - Formative

Trust assessment

Astronomy research is entirely digital: time to go "beyond the PDF"

- Preserved experiments
  - Methodology "in action"
  - All data are exposed
  - Reproducible
  - Repeatable
  - Re-usable
  - Re-purposeable
  - Participatory
  - Collaborative
  - Formative

Social aspect of science

Astronomy research is entirely digital: time to go "beyond the PDF"

- Preserved experiments
  - Methodology "in action"
  - All data are exposed

New kind of publication?

  - Reproducible
  - Repeatable
  - Re-usable
  - Re-purposeable
  - Participatory
  - Collaborative
  - Formative

Astronomy research is entirely digital: time to go "beyond the PDF"

Discoverable!

- Preserved experiments
  - Methodology "in action"
  - All data are exposed
  - Reproducible
  - Repeatable
  - Re-usable
  - Re-purposeable
  - Participatory
  - Collaborative
  - Formative

Workflow, not data preservation

- Workflows are interpreted through their execution
  - Complex models are required to describe them
- Severely vulnerable to obsolescence
  - Applications
  - Libraries
  - Operating environment
- Provenance is a complex issue in a cloud of services
- Resources are often beyond control of scientists
- Alleviate decay of external resources via alternates
- Ensure trustworthiness and authenticity

Workflow, not data preservation

- Versioning of the whole workflow, or its components
- Access control policies on data and processes
- Permissions, licenses, platform, costs, etc.
- Semantic discovery (WFs, processes, web services)
- QA: usage, logs, uptime…

Workflows and Processes should benefit of the same privileges acquired by Data

Preserve, Retrieve, Reconstruct, Replay

- Retrieve
  - Functionality of the WF and/or its modules
  - What are the inputs and outputs
  - Metadata: Authority, Complexity, Keywords…
- Reconstruct
  - Understand dependencies and components
  - Technical specificities
- Replay
  - Check the success of the preservation method
- Referenced and acknowledged

Preserve, Retrieve, Reconstruct, Replay

- Retrieve
  - Functionality of the WF and/or its modules
  - What are the inputs and outputs — **Characterisation**
  - Metadata: Authority, Complexity, Keywords…
- Reconstruct — **Tools**
  - Understand dependencies and components
  - Technical specificities — **Semantics & Modelling**
- Replay
  - Check the success of the preservation method

- Referenced and acknowledged — **Long term IDs**

More than a WF: The Research Object (RO)

- All components related to the research lifecycle of an experiment should be available.

- Preserved and easily retrievable
  - Proposals
  - Data
  - Processes
  - Workflows
  - Publications

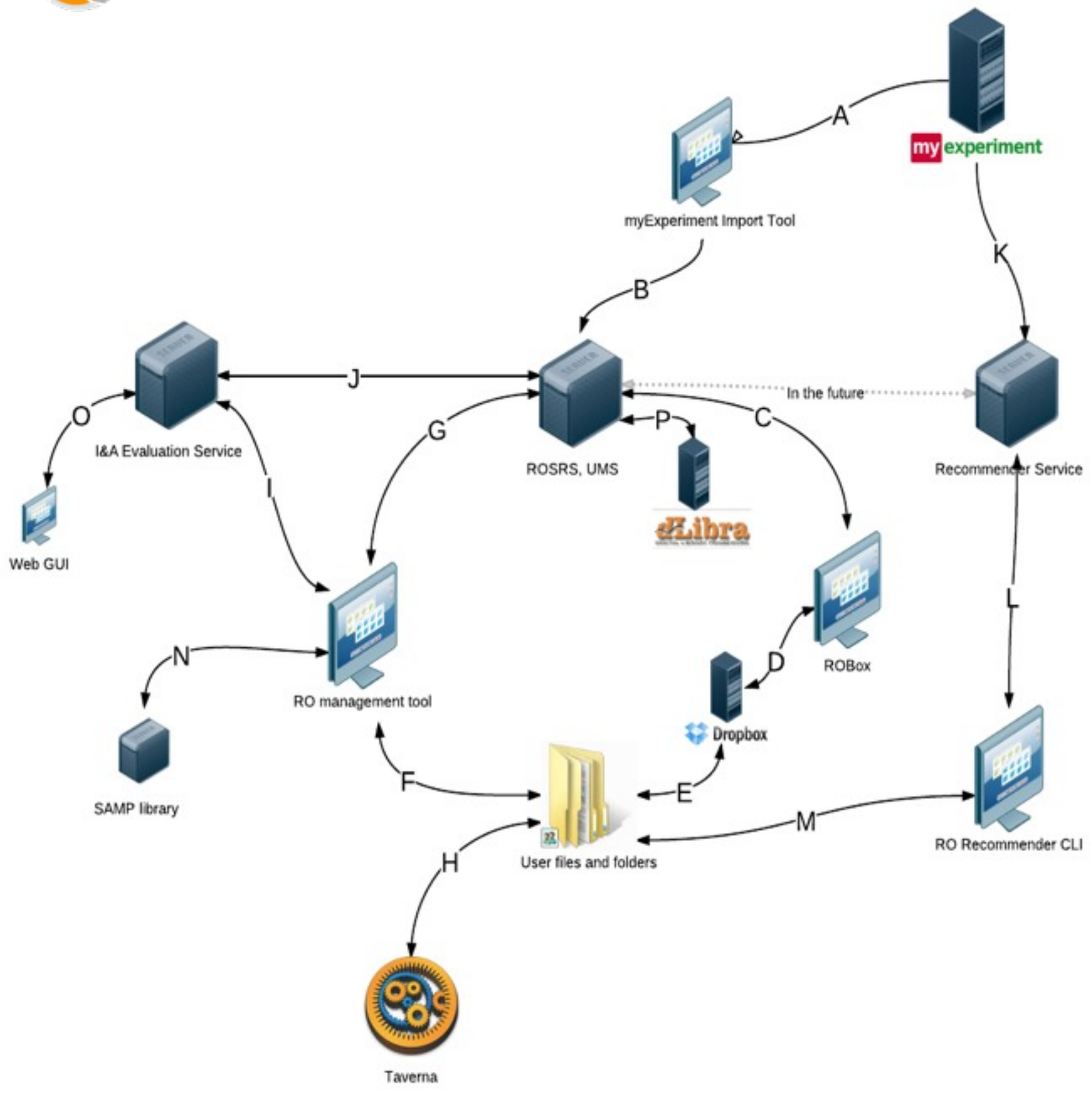
All linked by persistent IDs

- User Requirements
  - Functional requirements for Wf4Ever “working” platform
  - Focused on improving collaboration and reuse
  - Interoperability in exchanging scientific methodology
  - Expose experiment in a structured way to be understood by others

We need to build what we want to preserve!

- RO Modeling
  - Model for interlinked components in a Research Object
  - Strategies for assessing integrity and authenticity
  - Attempts in metrics for Information Quality

Wf4Ever Update



- Architecture
  - Search & Retrieval Service
  - Recommender Service
  - I&A Evaluation Service
  - Notification Service

- User-Tools Prototypes
  - RO Command Line Tool
  - RO Annotator
  - RO Box

New Workflows in myExperiment

Search results for "virtual observatory"

Search filter terms

Sort by: Rank

Showing 5 results. Use the filters on the left and the search box below to refine the results.

virtual observatory | Search

Filter by category

- Workflow 3
- Group 1
- User 1

Filter by type

- Taverna 2 3

Filter by tag

- virtual observa… 4
- astronomy 3
- votable 3
- astrogrid-taver… 1
- astrophysics 1
- workflows 1

Filter by user

- Pique 3

Filter by licence

- by-nd 3

Taverna 2

AMIGA ConeSearch (v3)

View

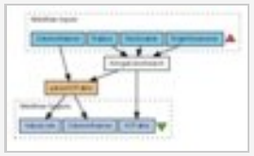
Download (v3)

Original Uploader

Pique

Created: 11/07/11 @ 22:08:06 | **Last updated:** 11/07/11 @ 23:34:13

License: BSD License



This workflow provides a VOTable response from the AMIGA ConeSearch service and extract values from VOTable columns.

Rating: 0.0 / 5 (0 ratings) | **Versions:** 3 | **Reviews:** 0 | **Comments:** 0 | **Citations:** 0

Viewed: 4 times | **Downloaded:** 1 time

Tags (3):

astronomy | virtual observatory | votable

Taverna 2

AMIGA ConeSearch from a file of targets/positions (v1)

View

Download (v1)

Original Uploader

Created: 12/07/11 @ 17:34:33 | **Last updated:** 12/07/11 @ 17:36:37

License: BSD License

This workflow takes an ASCII file of position of

New/Upload

Workflow | GO

Log in / Register

Username or Email:

Password:

Remember me:

OR

Use OpenID:

(eg: name.myopenid.com)

Log in

Need an account?
Click here to register

Forgot Password?

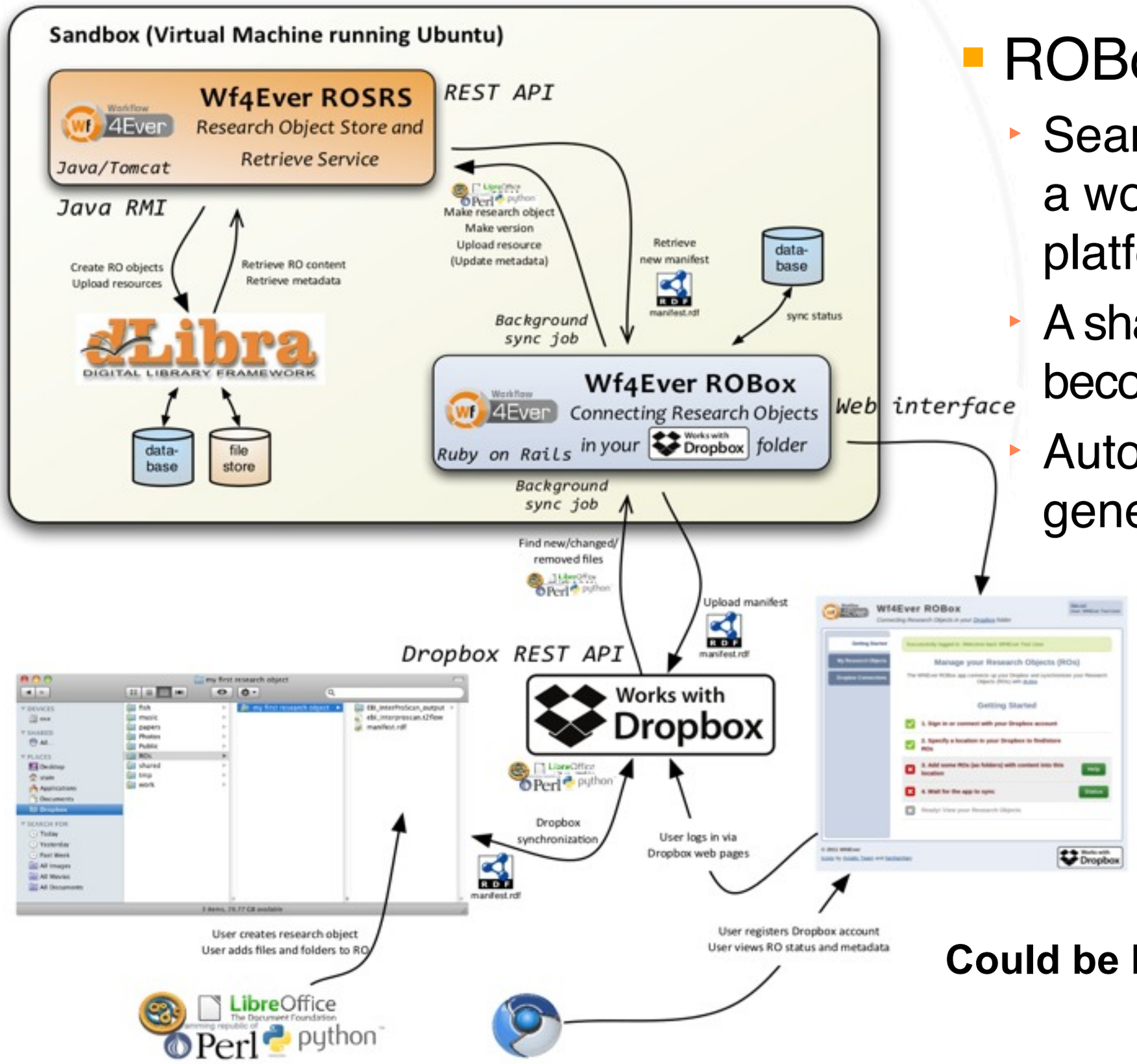
Popular Tags

25 tags

[All Tags]

benchmarks | bio2rdf | bioinformatics | BLAST | cheminformatics | data integration | ebi | example | gene |

Architecture of Prototype 1



- ROBox
  - Seamless contribution to a working collaborative platform
  - A shared folder in Dropbox becomes a Working RO
  - Automatic metadata generation

Could be based on VOSpace!

Wf4Ever - RO Annotator MOCKUP

Research Object: Epigenius_experiment1

- Datasets
  - HD_dataset1 (GEO series datafile)
  - HD_dataset2 (GEO series datafile)
- Scripts
- Web Services
- Workflows
- Docs

Structure in Dropbox

Annotating "HD_dataset1 (GEO series datafile)"

Type	GEO series datafile
Keywords	human, brain, datas...
Description	Human brain data...
Role	To be used as input...
Created At	2011-09-06 11:00...

Metadata for selected item

What kind of annotation is this?

Description

Value for the annotation

B I S U | X₂ X² | H1 H2 H3

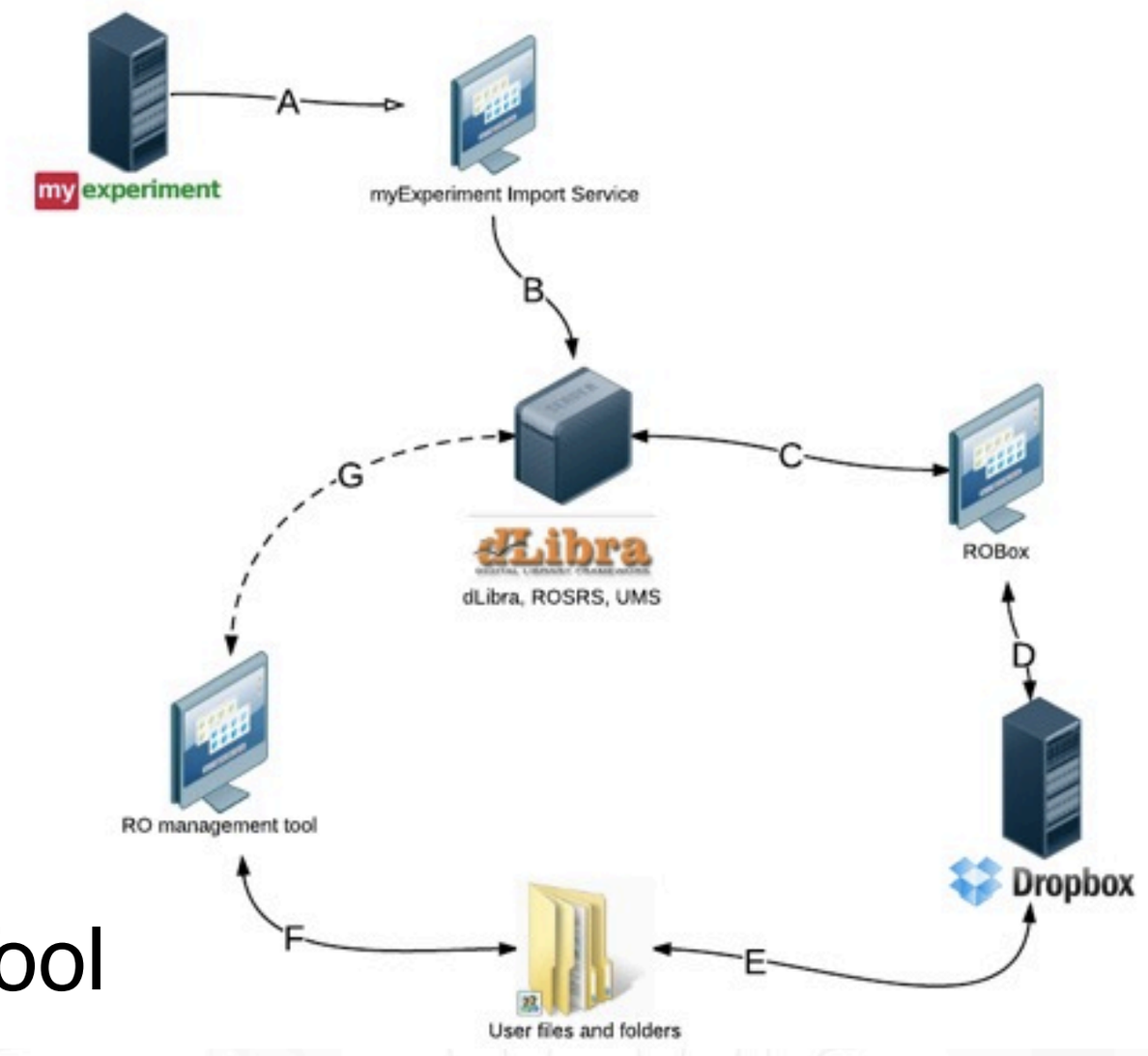
Human brain dataset. 44 HD samples, 36 Controls age and sex matched. Brain areas:caudate nucleus, frontal cortex and cerebellum. Affymetrix platform. Rows correspond to probe ids and columns to samples.

Save Changes | Cancel

Unstructured, rich-text metadata editor

Notification Service for Authors

- What should be notified?
  - Fails
  - Downloads
  - Annotations
  - Linked/Similarity
  - Modifications on Working RO
  - Acknowledgements
- Notification Management Tool
  - Avoid spam



Conclusions

- Workflows are a powerful, semantically rich way of describing astronomical knowledge discovery methods
  - Provide both glue and structure to the method
  - Also allow for metadata encapsulation

- Preserving workflows allows for method reuse, experiment replay, dissemination, attribution, trust building

- Wf4Ever is providing a framework for allowing astronomers to start using workflows without leaving their tools
  - But with the idea of nudging them toward more structured workflow descriptions