The Astronomy challenge: How can workflow preservation help?

Susana Sánchez, Jose Enrique Ruíz, Lourdes Verdes-Montenegro,
Julian Garrido, Juan de Dios Santander-Vela and Wf4Ever team.

Instituto de Astrofísica de Andalucía – CSIC

SHIWA Summer School. 3 July 2012

- Introduction to AMIGA group
- The astronomy challenge and context of Wf4Ever project
- How Wf4ever tools can help the astronomers
- Our astronomy use case

AMIGA

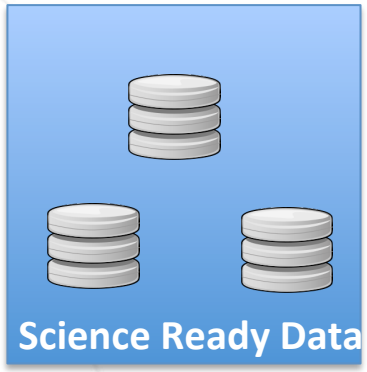
Analysis of the interstellar Medium of Isolated GAlaxies

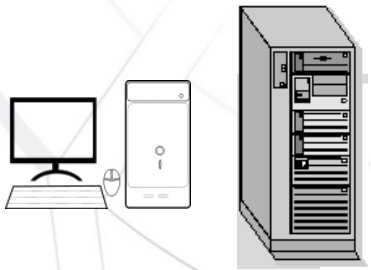
An international collaboration coordinated from the IAA-CSIC
P.I. Lourdes Verdes-Montenegro
http://amiga.iaa.es

- Statistical baseline of isolated galaxies to compare with the behavior of galaxies in denser environments
- Multi-λ study of ~1000 galaxies:
  - Need of intensive and complex analysis of multidimensional data

The astronomy challenge and context of Wf4Ever project

Past (or current?) situation



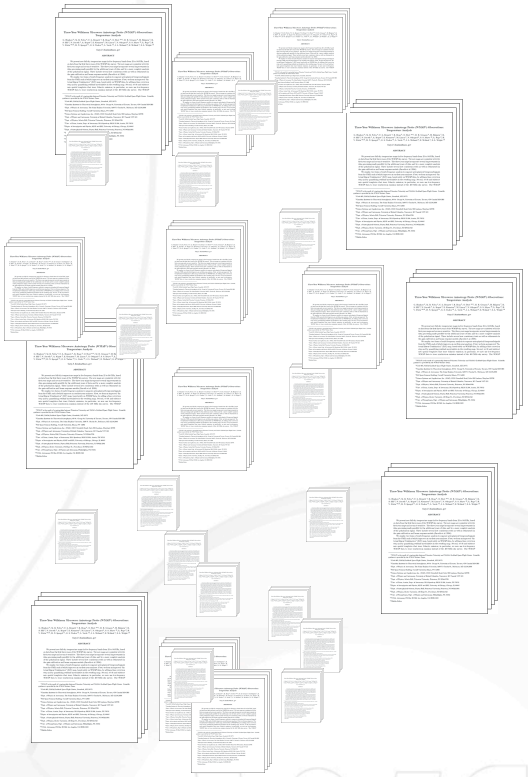


Analysis

Python

MathWorks

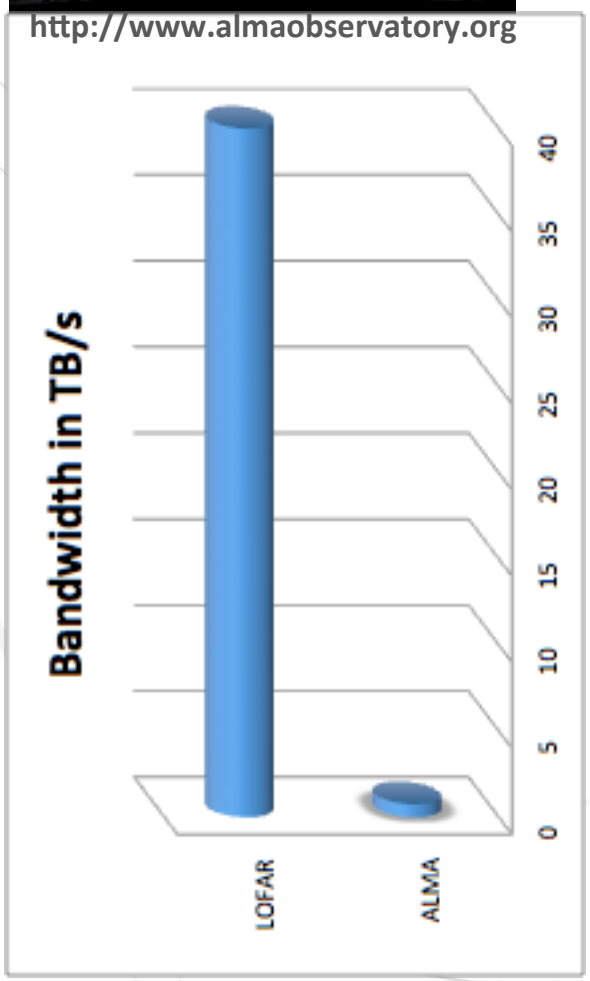
GIPSY

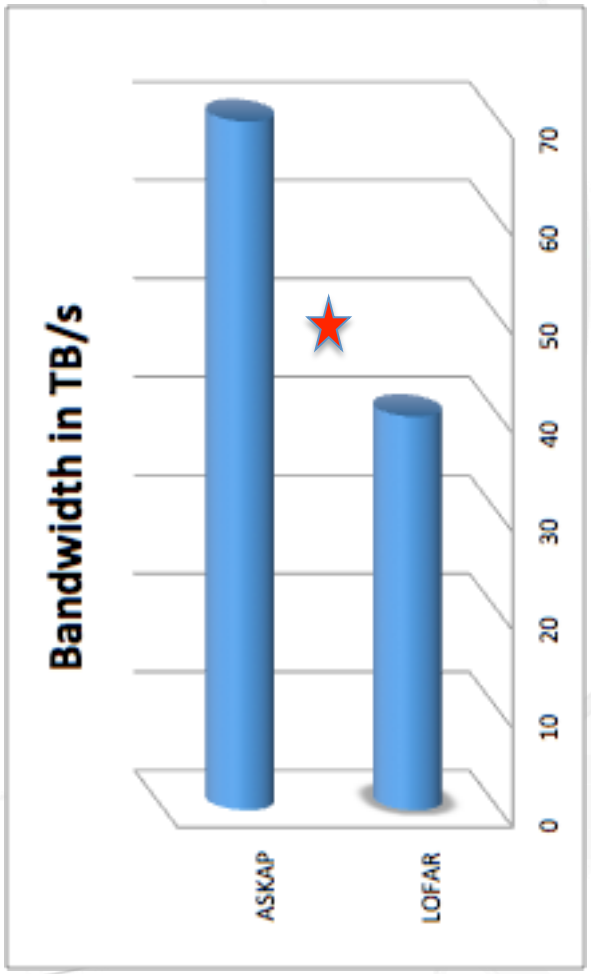


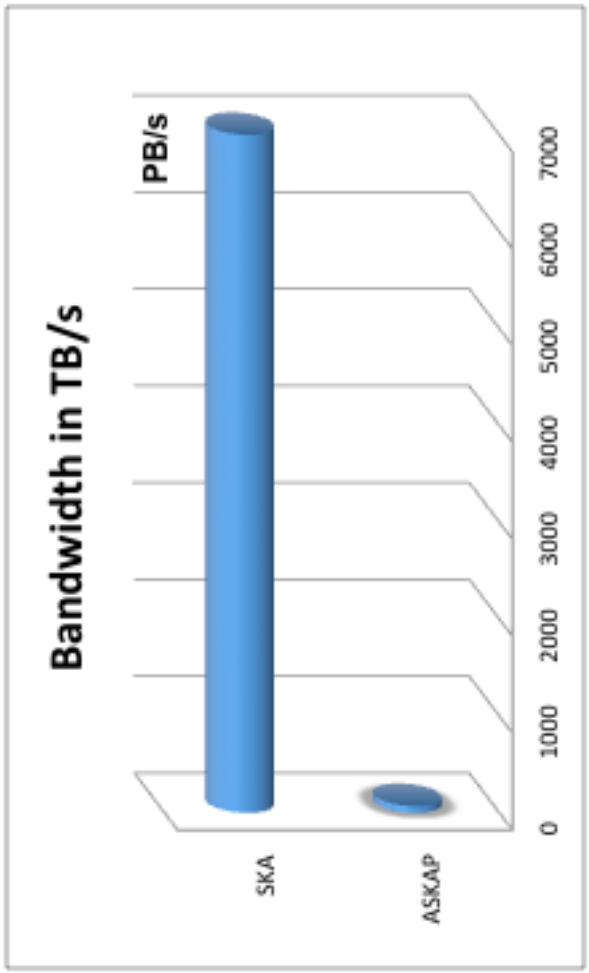
The astronomy challenge and context of Wf4Ever project

http://www.almaobservatory.org

Antenna → Correlation



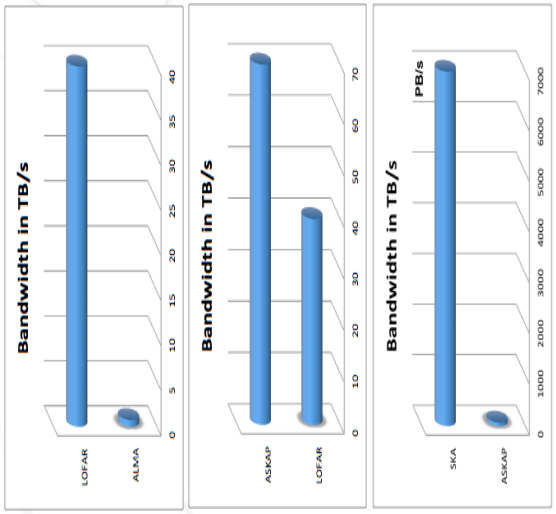




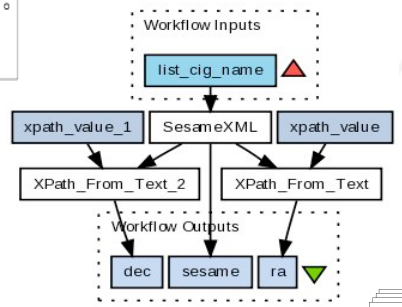
★ LHC – Tier 1

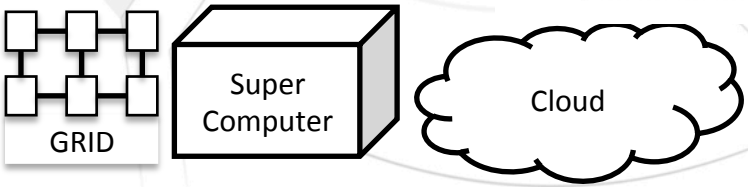
The astronomy challenge and context of Wf4Ever project

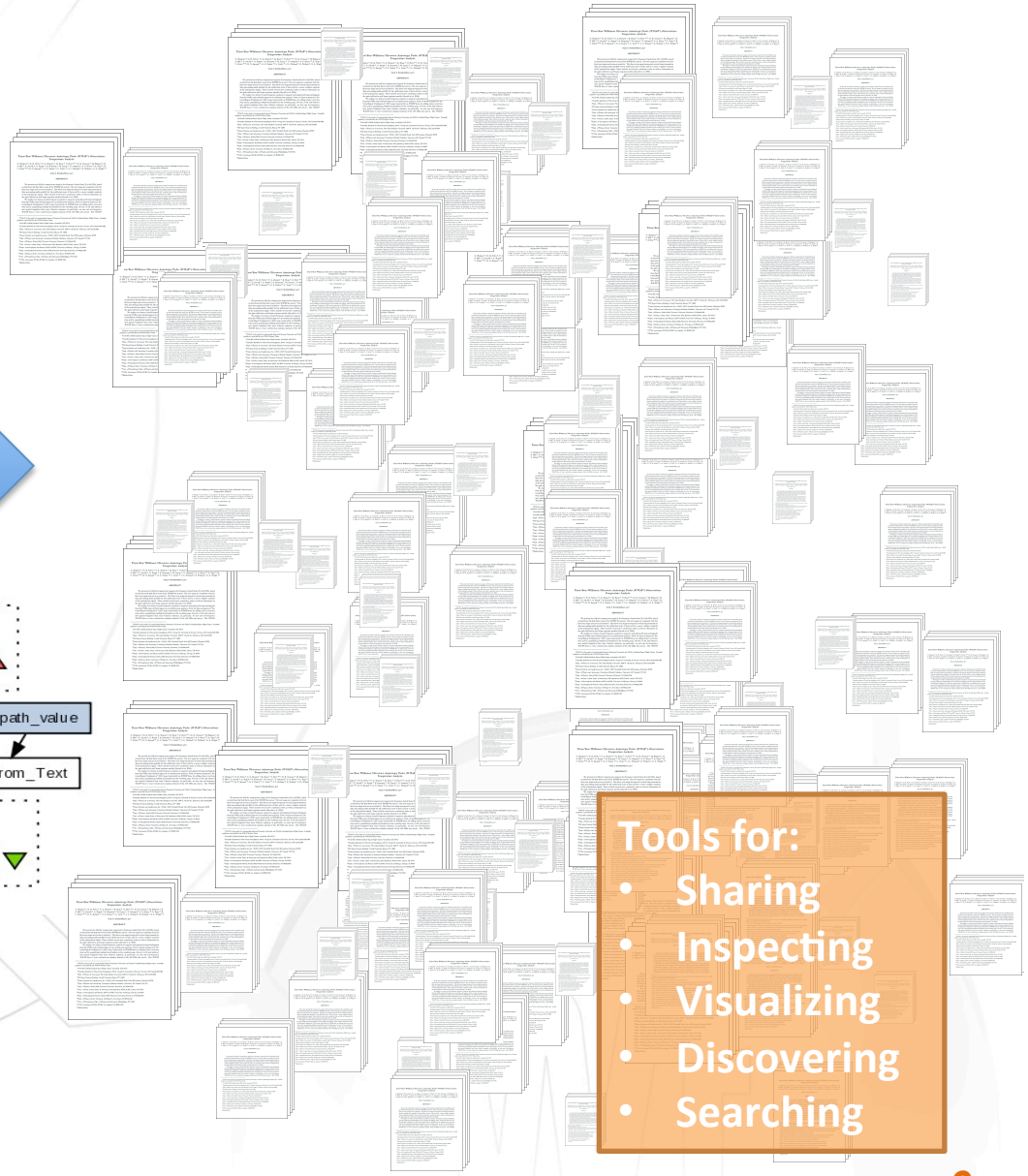
A disruptive change in the methodology is needed



Analysis







Tools for:

- Sharing
- Inspecting
- Visualizing
- Discovering
- Searching

» Keywords: efficient use of data and reliable science

› Scientific Workflows:

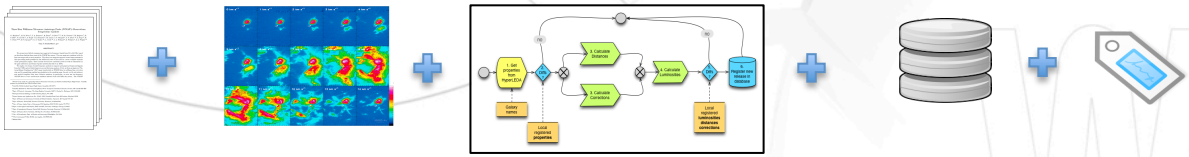
- Enable automation and expose the flow of scientific methods
- Encourage best practices in packing the experiment
- Provide a way to share the experiment

› But more is needed:

- Reusability, fundamental for incremental scientific development
- Reproducibility, key for reliable science

→ **Preserve the data and the scientific method**

Workflow preservation



Wf4Ever - Preservation of scientific workflows in data-intensive science

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)

The astronomy challenge and context of Wf4Ever project

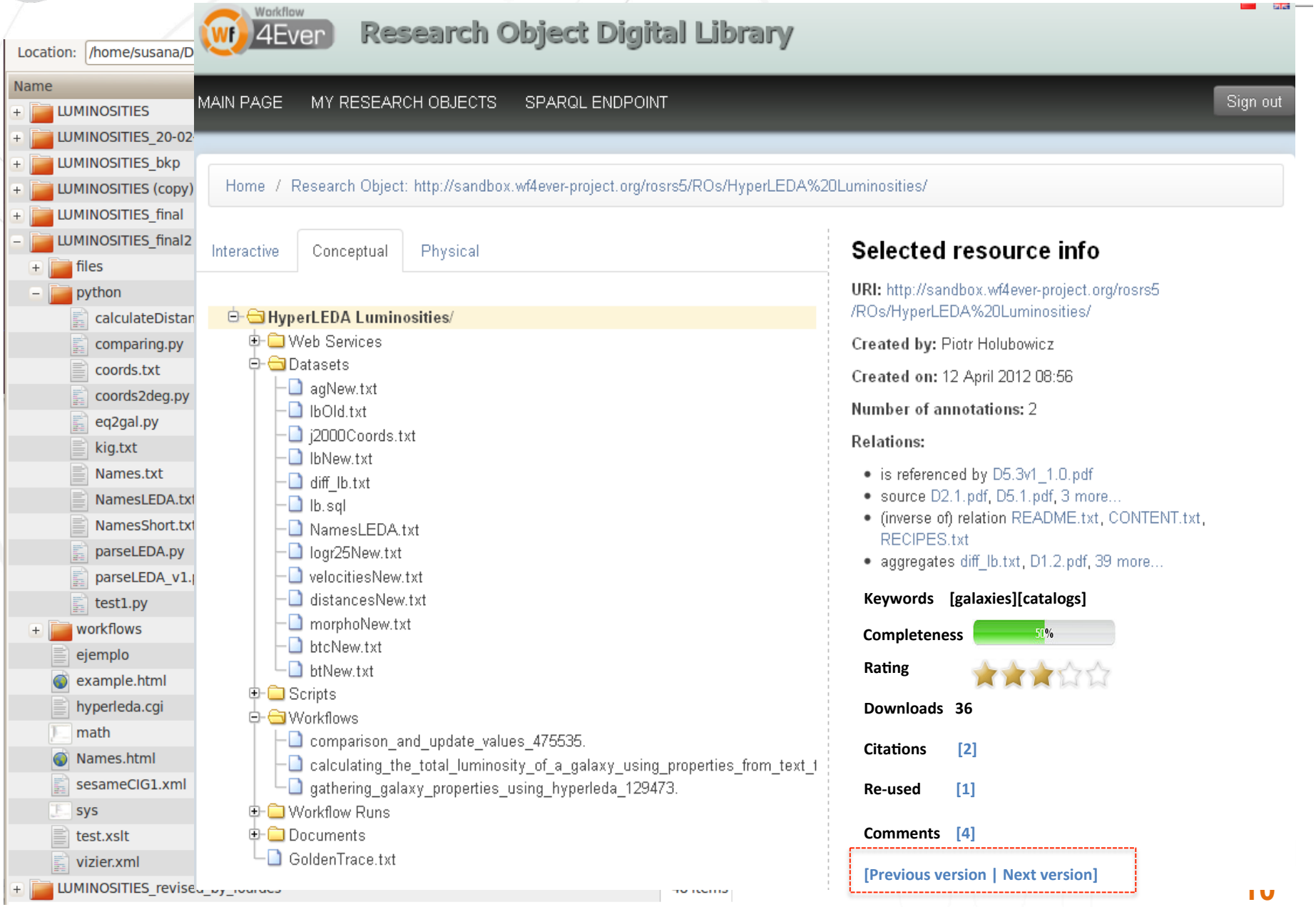
The aim of Wf4Ever

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

- Encapsulate the scientific methodology (the workflows and all the associated information) in an artefact called Research Object.
- Archival, classification and indexing of the research object in scalable semantic repositories, providing advanced access and recommendation capabilities based on monitoring and metrics to evaluate similarities, decay, quality, stability, completeness.
- Creation of scientific communities to collaboratively share, reuse and evolve Research Objects stimulating the development of new scientific knowledge
- Use Cases:
  - Astronomy (IAA)
  - Genome-wide Analysis and Biobanking (LUMC)

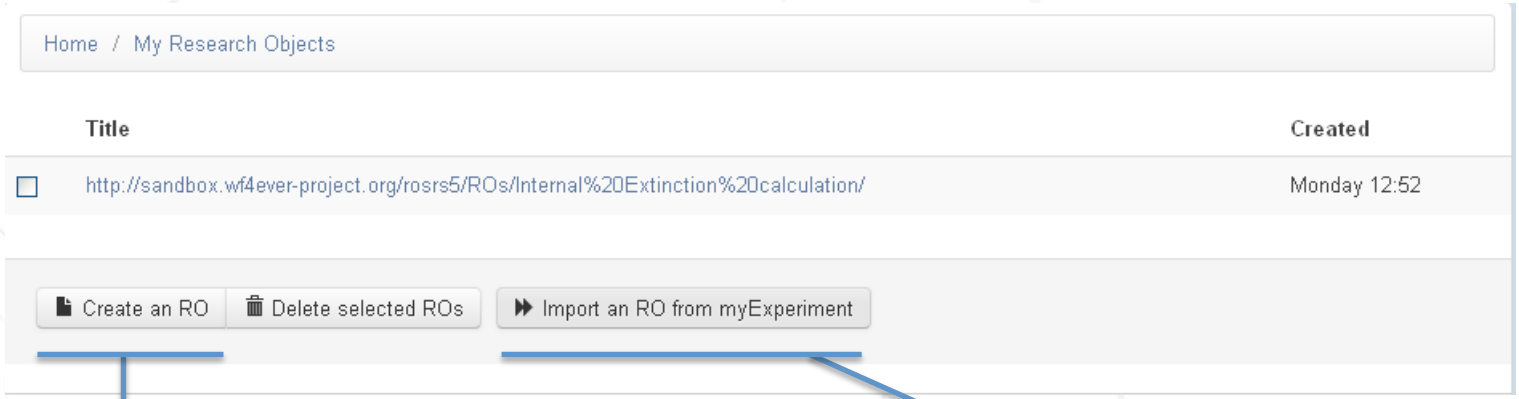
How Wf4ever tools can help the ~~astronomers~~ scientists

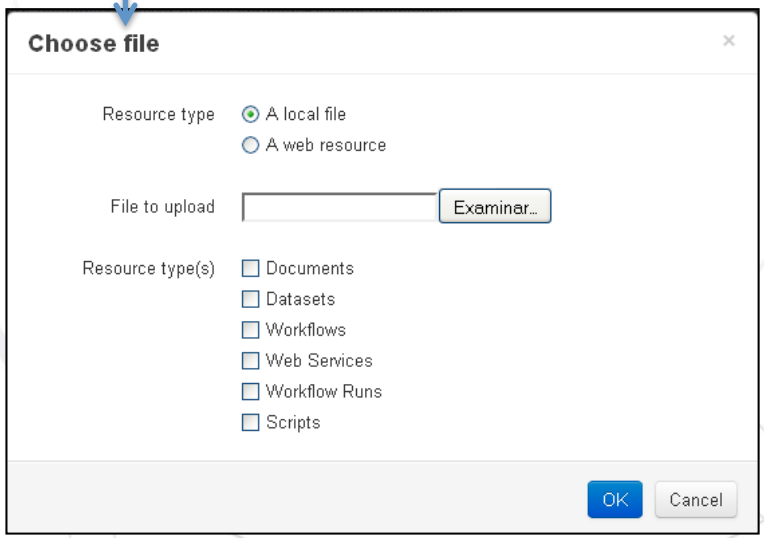
Sharing the experiment

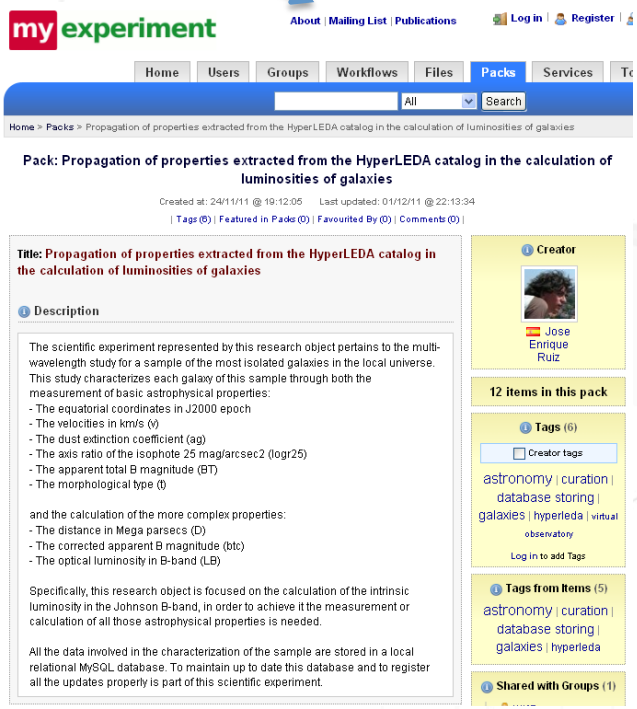


How Wf4ever tools can help the scientists

How to import the experiment







INTEROPERABILITY

How Wf4ever tools can help the scientists

How to import the experiment

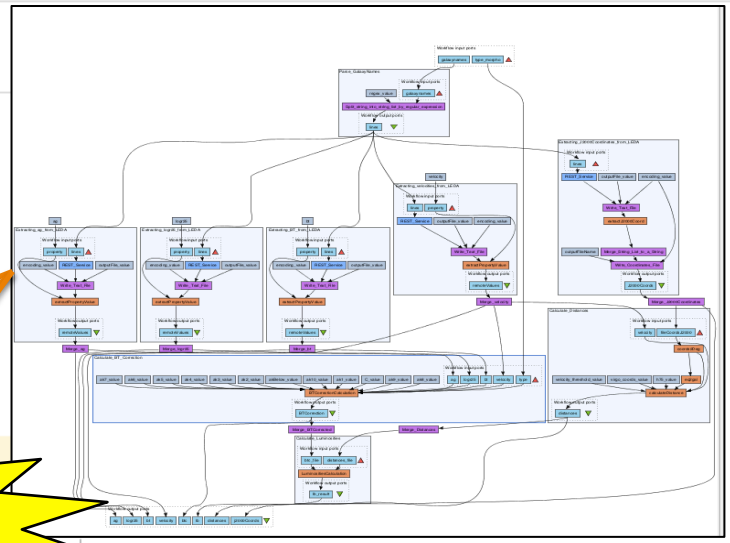
Research Object Digital Library

MAIN PAGE | MY RESEARCH OBJECTS | SPARQL ENDPOINT | Sign out

Home / Research Object: http://sandbox.wf4ever-project.org/rosrs5/ROs/HyperLEDA%20Luminosities/

Interactive | Conceptual | Physical

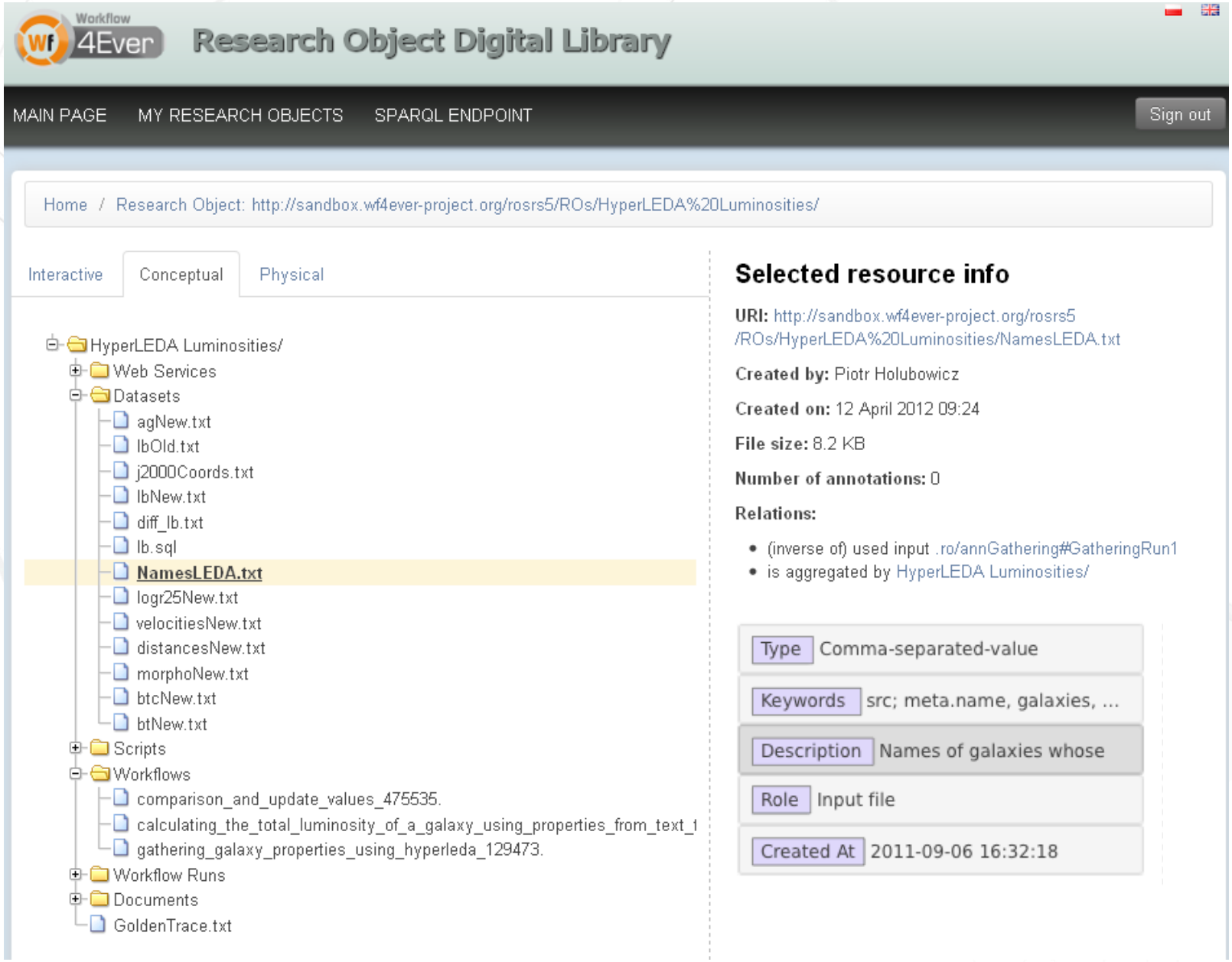
- HyperLEDA Luminosities/
  - Web Services
  - Datasets
    - agNew.txt
    - lbOld.txt
    - j2000Coords.txt
    - lbNew.txt
    - diff_lb.txt
    - lb.sql
    - **NamesLEDA.txt**
    - logr25New.txt
    - velocitiesNew.txt
    - distancesNew.txt
    - morphoNew.txt
    - btcNew.txt
    - btNew.txt
  - Scripts
  - Workflows
    - comparison_and_update_values_475535.
    - calculating_the_total_luminosity_of_a_galaxy_using_properties_from_text_1
    - gathering_galaxy_properties_using_hyperleda_129473.
  - Workflow Runs
  - Documents
  - GoldenTrace.txt



Extract !

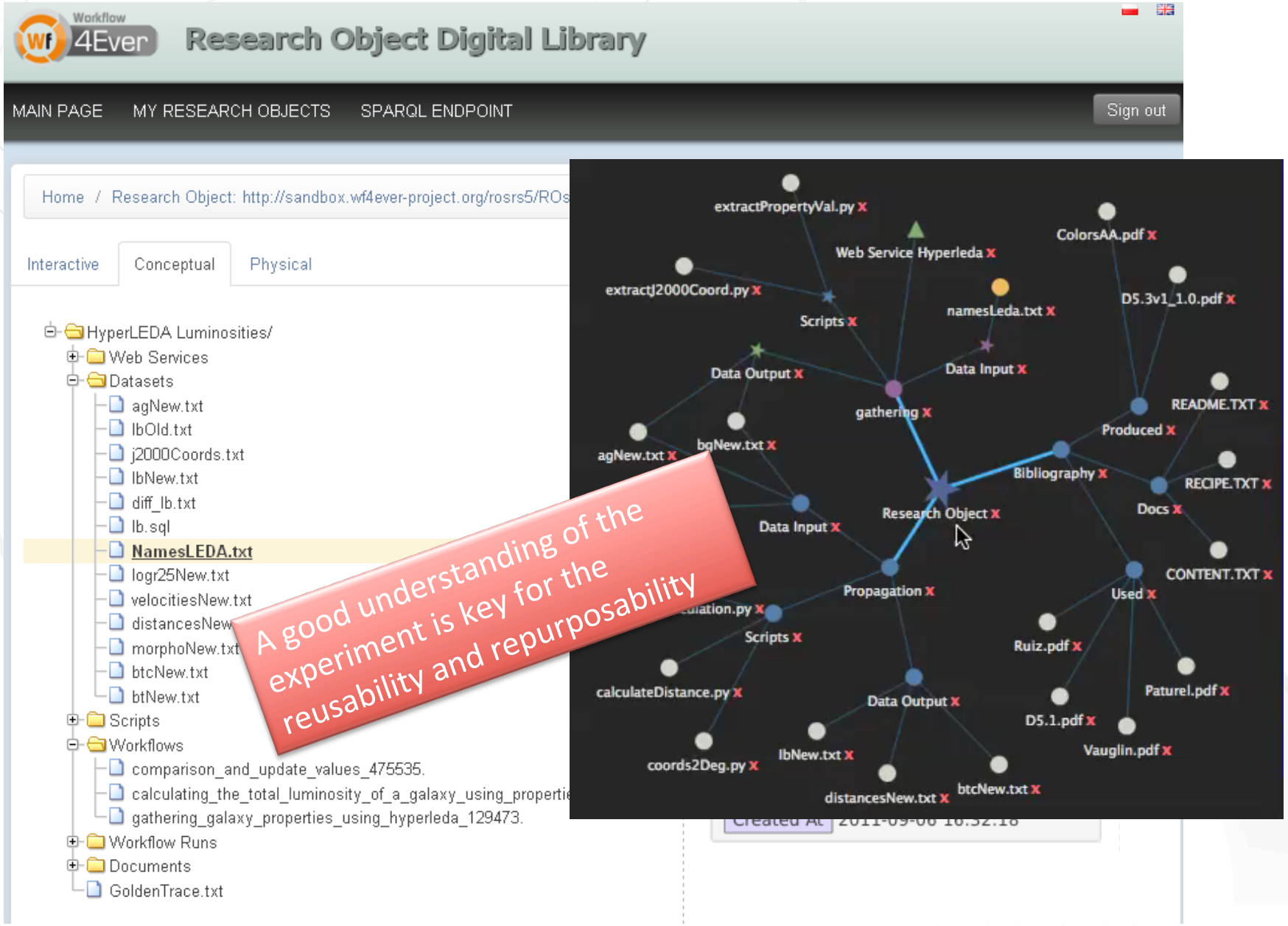
How Wf4ever tools can help the scientists

Describing the experiment



How Wf4ever tools can help the scientists

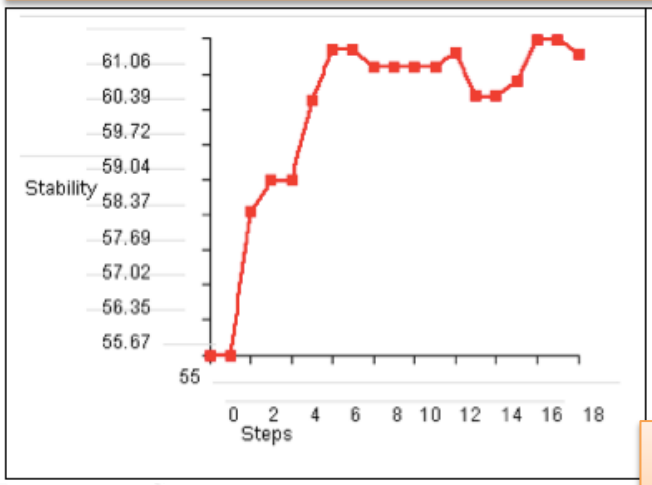
Visualizing the experiment



How Wf4ever tools can help the scientists

Checking the live experiment

Stability: Changes made by different kind of users on the RO, can improve it or make it worse



0. Alice creates RO
1. Alice adds Recipe.txt
2. Alice adds Content.txt
3. John adds Results.txt
4. Alice adds Bibliography.pdf
5. Alice annotates Bibliography.pdf
6. John removes Script.py
7. John edits annotation on Recipe.txt
8. Unknown adds Dropme.txt

Completeness: It contains all the resources needed to be run, published, shared or repeated

75%	Runnable	Service up, software working, etc.
60%	Repeatable	The output can be reproduced
20%	Shareable	Enough annotation.
50%	Publishable	Service up, software working and all well commented

Decay: The health of the RO: state of the services (up or down), of the applications (updated or deprecated), permissions to access the input data

Decay Information

Last check was performed 2 days ago and returned one error:

```
The service SDSS-DR7, needed by the workflow
Calculate_galaxy_distances is down
```

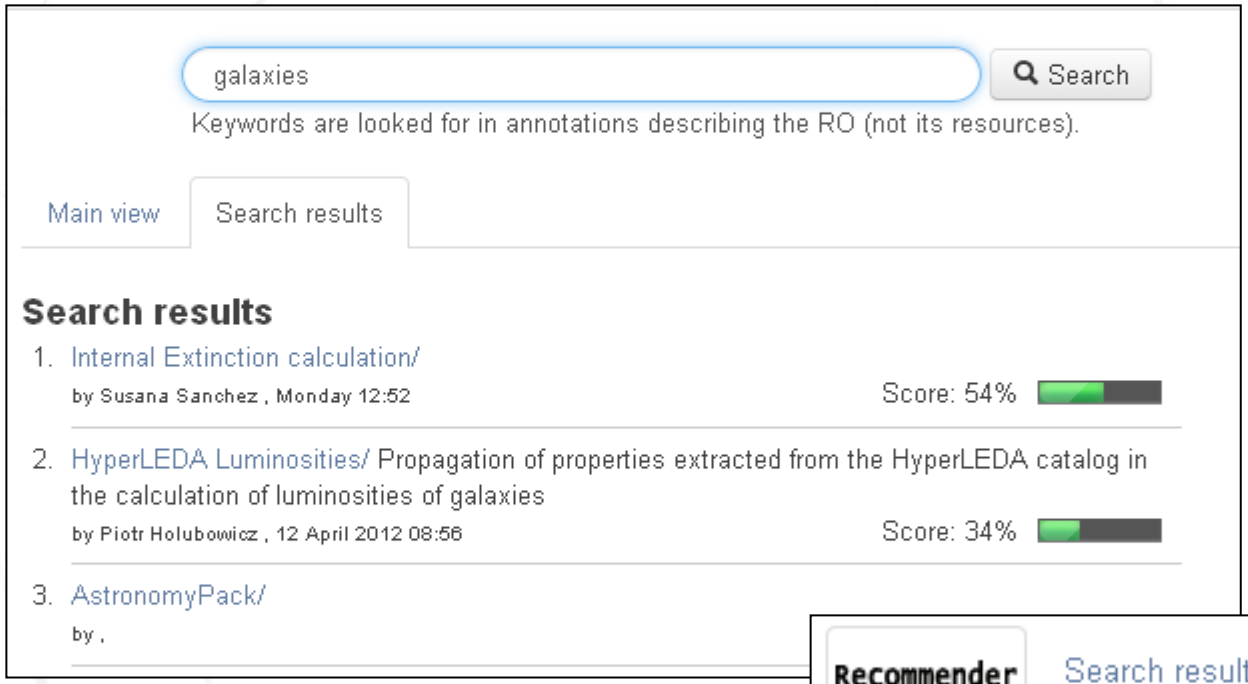
Check now Try to repair

Tracking: Rating by other users, who used the RO, comments, etc.

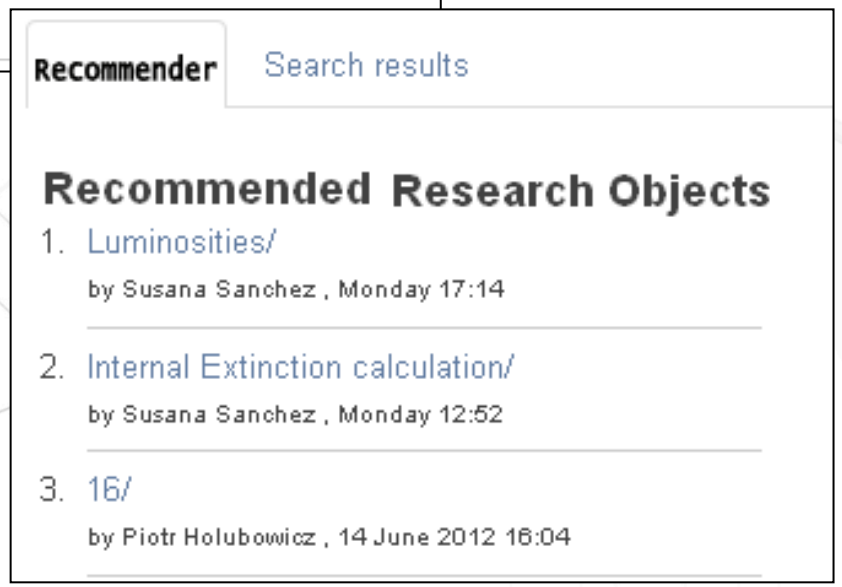
Rating	★★★☆☆
Downloads	**36**
Citations	[2]
Re-used	[1]
Comments	[4]

How Wf4ever tools can help the scientists

Discovering an experiment



Search by keywords



Past ratings
Ratings of similar users

REUSABILITY AND REPURPOSABILITY

- Annotations for the description of the whole and its components
- Visualization of the relationships between the components
- Versioning of the whole or its components

SHARING

- Restricted access on data and processes
- Ensuring authorship and allow to RO to be cited

REPRODUCIBILITY

- Completeness checking
- Decay monitoring and notification, reacting to decay
- Execution and interoperability support

DISCOVERY

- Semantic discovery of ROs, processes, web services
- Recommendation capabilities.

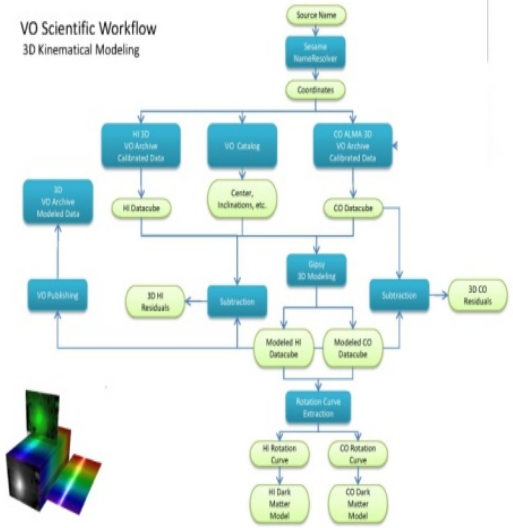
Our astronomy Use Case

Why I am here as student

Deploy **web-services-based workflows** for analysis of **multidimensional data** on heterogeneous **e-infrastructures**

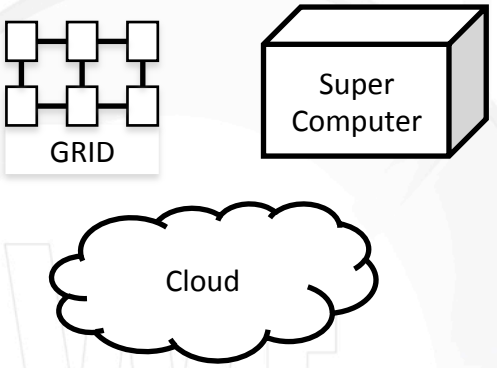
	Low Res		High Res		Extreme Res	
Number	4 Bytes	4B	4 Bytes	4B	4 Bytes	4B
Resolution	2,048 x 2,048	16MB	8,192 x 8,192	268MB	12,288 x 12,288	603MB
Channels	16,384	0.27TB	16,384	4.39TB	16,384	9.8TB
Stokes & Weighting	1	0.27TB	1	4.39TB	4 + 1	49.5TB

ASKAP Cubes Prof. Kevin Vinsen



VIRTUAL OBSERVATORY

- Standards for publishing and accessing astronomical data (Service Oriented Architecture)
- Data provider → analysis service provider



Susana Sánchez: sse@iaa.es

http://amiga.iaa.es/p/212-workflows.htm
http://www.wf4ever-project.org