

e-Science for the SKA

WF4Ever: Supporting Reuse and Reproducibility in Experimental Science

Lourdes Verdes-Montenegro*

AMIGA and Wf4Ever teams

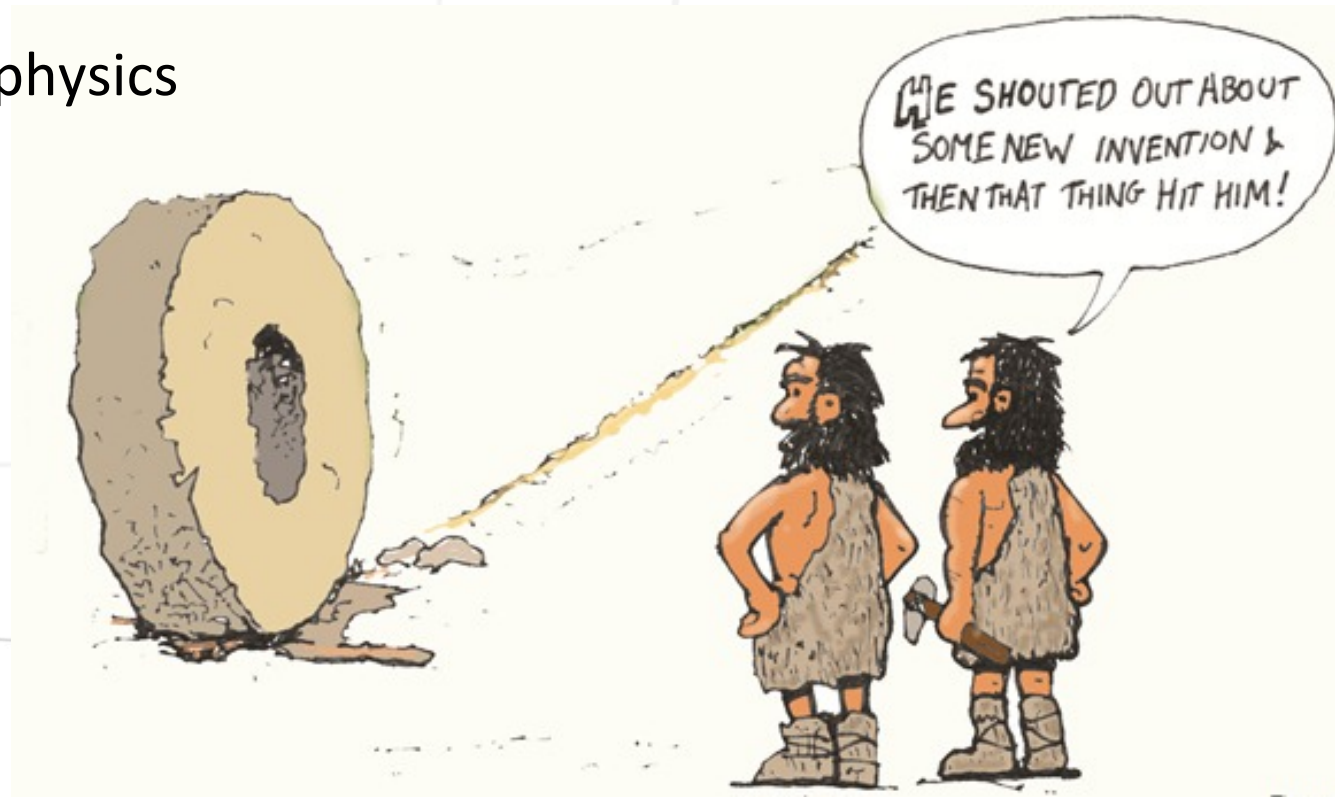
Instituto de Astrofísica de Andalucía – CSIC

*PI of VIA-SKA, Feasibility study of the Spanish technological participation in SKA

RadioNet Advanced Radio Astronomy, Commissioning Skills and Preparation for the SKA

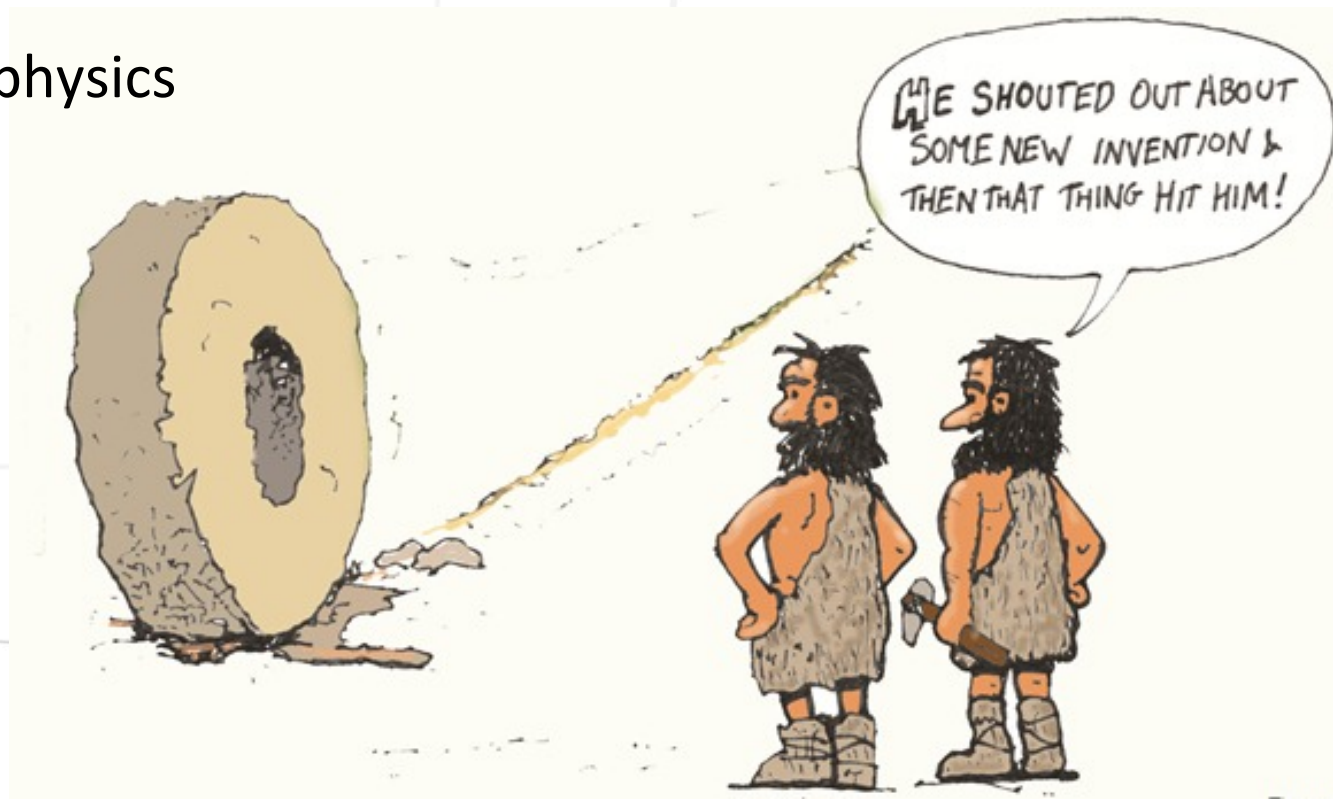
Manchester, November 15th 2012

- » The challenge: Reuse and Reproduce scientific experiments
 - Reusability, fundamental for incremental scientific development
 - Reproducibility, key for reliable science
- » How Wf4Ever is addressing these challenges
- » Impact in Astrophysics



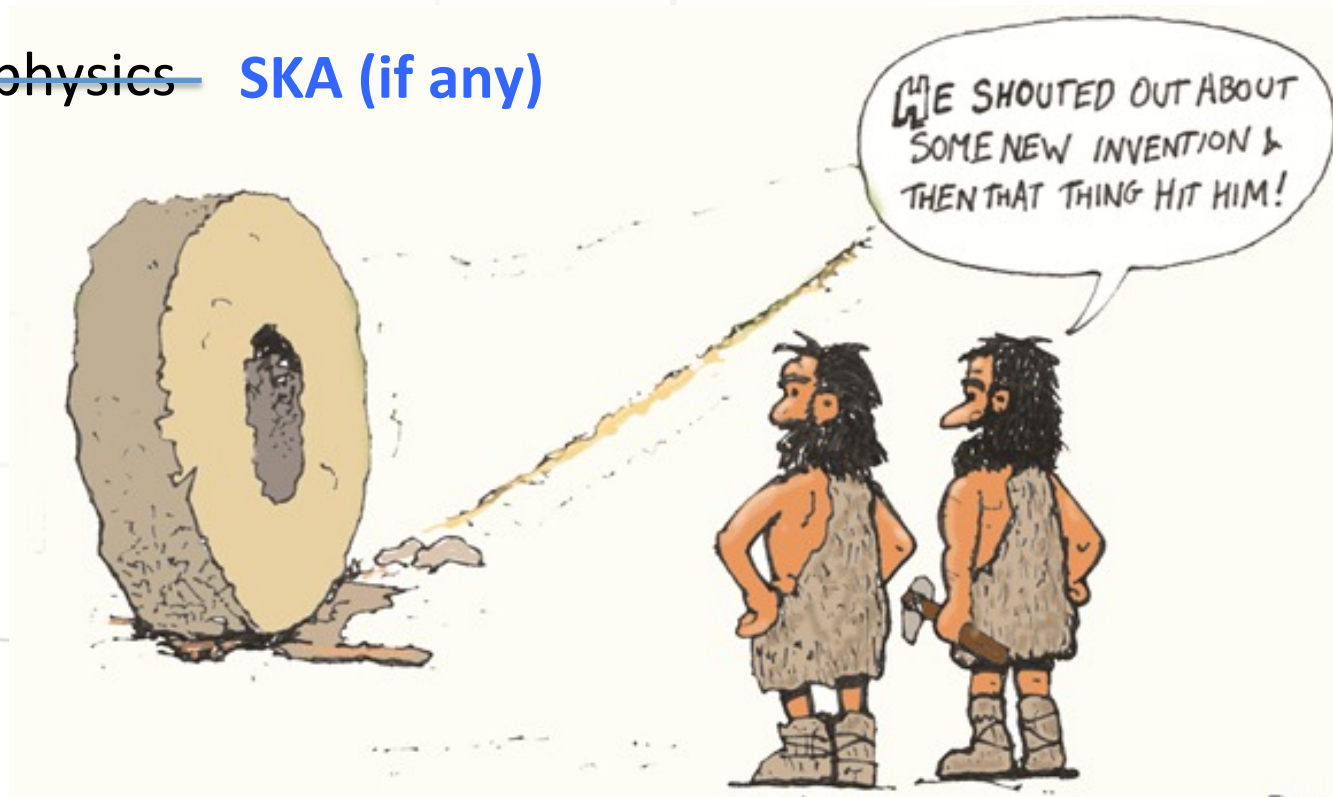
And commissioning looks soooo similar to a scientific experiment...

- » The challenge: Reuse and Reproduce scientific experiments
 - Reusability, fundamental for incremental scientific development
 - Reproducibility, key for reliable science
- » How Wf4Ever is addressing these challenges
- » Impact in Astrophysics



And commissioning looks soooo similar to a scientific experiment...

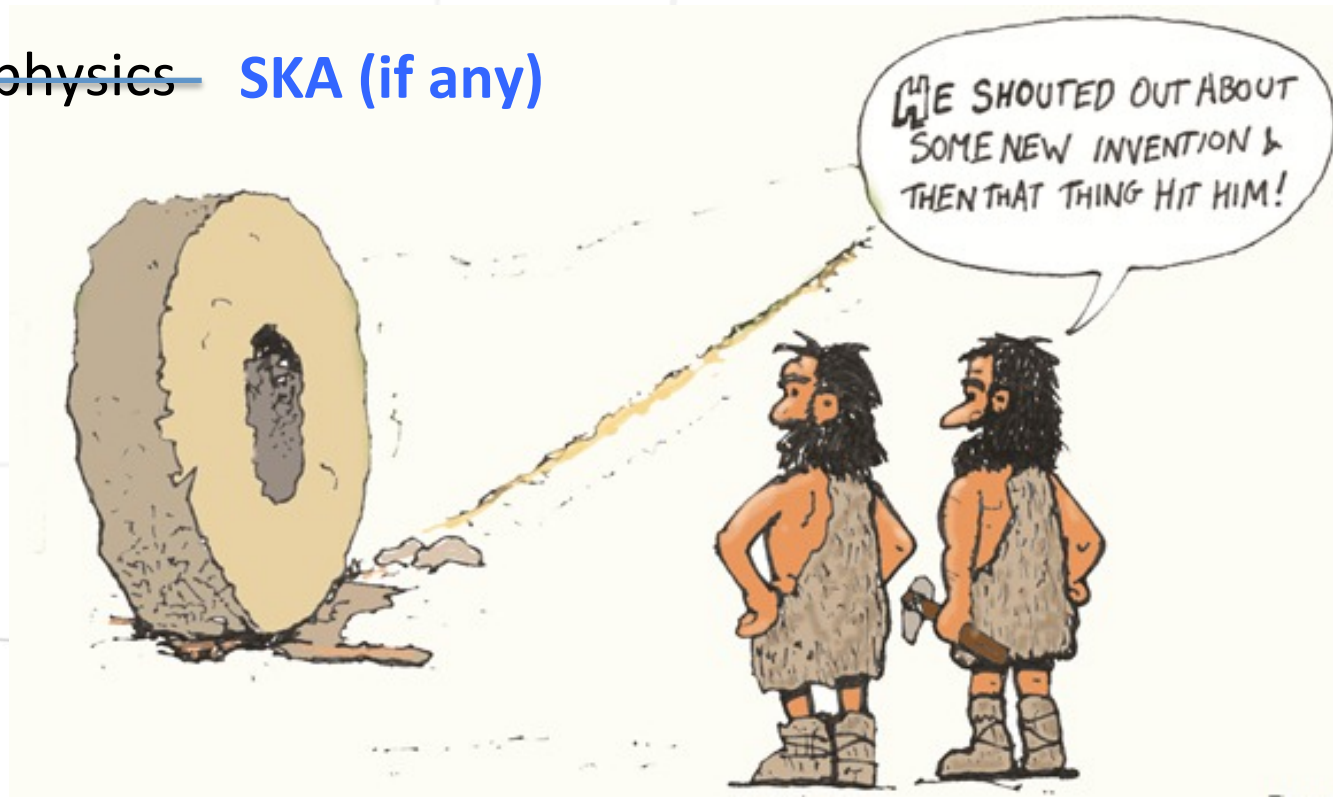
- » The challenge: Reuse and Reproduce scientific experiments
 - Reusability, fundamental for incremental scientific development
 - Reproducibility, key for reliable science
- » How Wf4Ever is addressing these challenges
- » Impact in ~~Astrophysics~~ **SKA (if any)**



And commissioning looks soooo similar to a scientific experiment...

An experiment is reproducible when someone else, working independently, can get the same results following the same methods and using the same inputs

- » How Wf4Ever is addressing these challenges
- » Impact in ~~Astrophysics~~ **SKA (if any)**

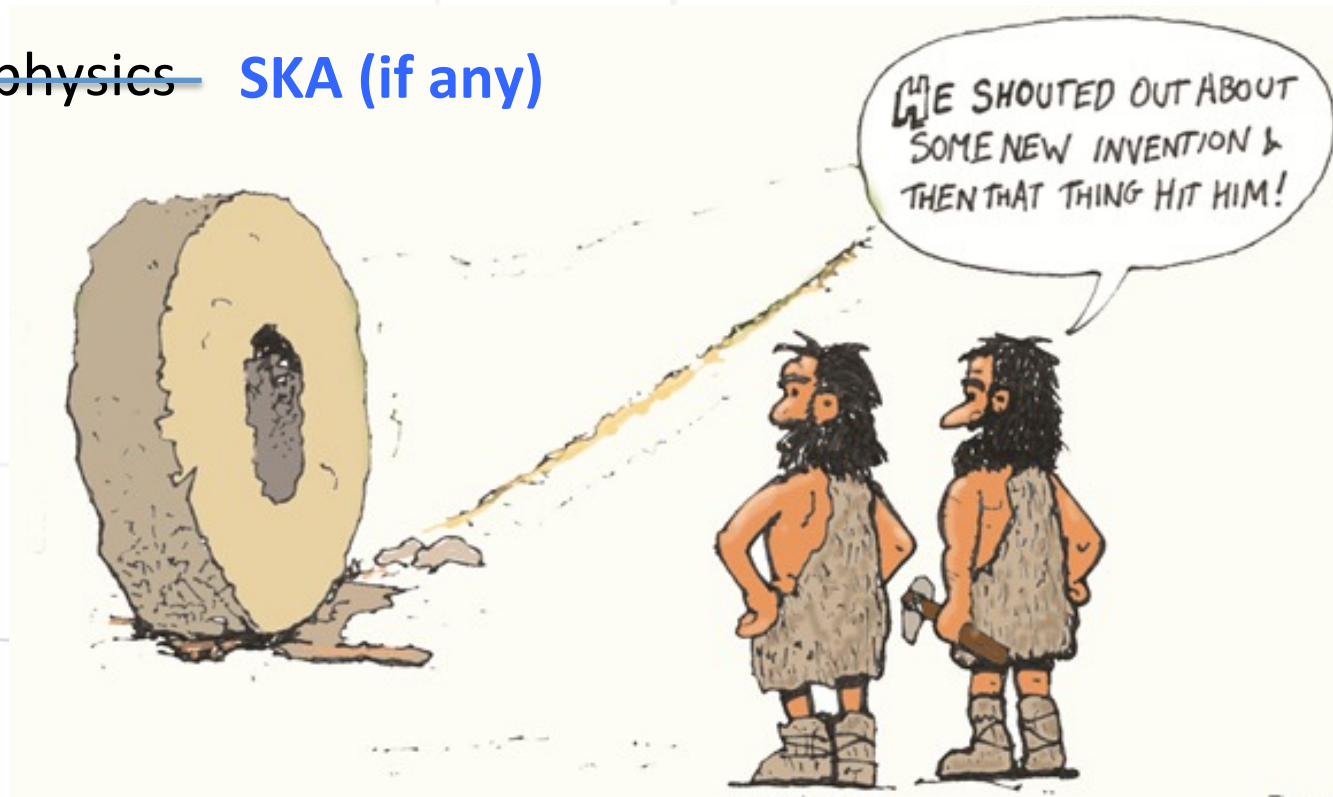


And commissioning looks soooo similar to a scientific experiment...

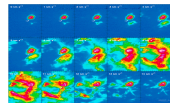
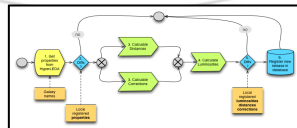
An experiment is reproducible when someone else, working independently, can get the same results following the same methods and using the same inputs

mmm... commissioning

- » How Wf4Ever is addressing these challenges
- » Impact in ~~Astrophysics~~ **SKA (if any)**



- » Compile all needed elements
 - » Share it with other ... **Commissioners**
 - » Describe it: methods, data, etc.
- » Make the experiment discoverable
- » Scientific Workflows: part of the solution
 - › Automation, encourage best practices, method for sharing
- » But more is needed:
 - › Share/provide the data, annotations, references, etc.
 - › Strategies for avoiding decay
 - › Tools for discovering the experiment



Workflow preservation

Wf4Ever - Preservation of scientific workflows in data-intensive science

EU funded FP7 STREP Project
December 2010 – December 2013



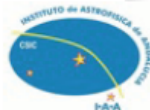
1. **Intelligent Software Components (ISOCO, Spain)**
2. University of Manchester (UNIMAN, UK)
3. Universidad Politécnica de Madrid (UPM, Spain)
4. Poznan Supercomputing and Networking Centre (PSNC, Poland)
5. University of Oxford (OXF, UK)
6. Instituto de Astrofísica de Andalucía (IAA, Spain)
7. Leiden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

isoco
enabling the networked economy



The University
of Manchester

MANCHESTER
1824



UPM



- Encapsulate the scientific methodology (the workflows and all the associated information/components of the digital experiment) in an artefact called **Research Object** (beyond the PDF).
- What to do with the Research Object?
 - Archival, classification and indexing in scalable semantic repositories
 - 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
- Use Cases:
 - Astronomy (IAA)
 - Genome-wide Analysis and Biobanking (LUMC)



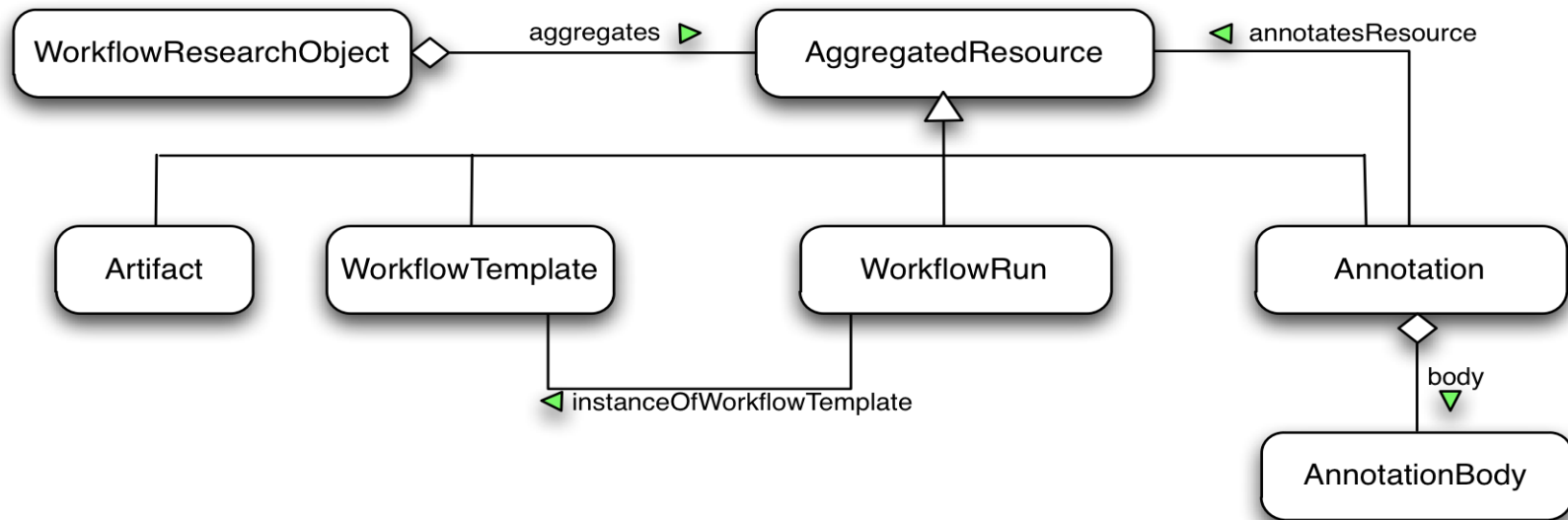
- Encapsulate the scientific methodology (the workflows and all the associated information/components of the digital experiment) in an artefact called **Research Object** (beyond the PDF).
- What to do with the Research Object?
 - Archival, classification and indexing in scalable semantic repositories
 - 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
- Use Cases:
 - Astronomy (IAA)
 - Genome-wide Analysis and Biobanking (LUMC)

Commissioning is multidisciplinary



Compile all the elements needed by the experiment

» Model for Workflow-Centric Research Objects [1]



» Semantic ontologies to implement the model

- › Object Exchange and Reuse (ORE) for specifying aggregation of resources
- › Annotation Ontology (OA) for annotating the resources

Metadata of the processes

» Provide digital libraries with RO preservation functionalities

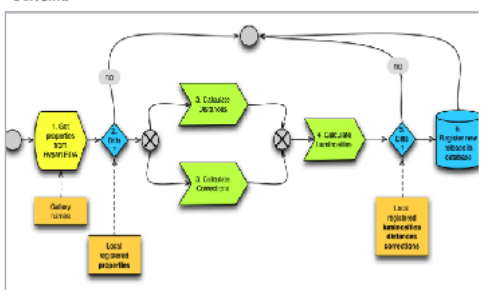

Curation of Astrophysical Quantities

Materials and Methods RO Versions [Prev](#) | [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) | [Next](#)

Conceptual **Relations**

Research Object

- Folder Gathering
 - Digital File
 - Hyperleda.tf2
- Folder Web Services
 - Hyperleda
 - NED CS
- Folder Input Data
 - ListCIGS
 - MorphoTypes
- Folder Results
 - Luminosities
- Folder Scripts
 - Calculation.py
- Folder Propagation
- Folder Comparison
- Folder Update
- Folder Bibliography
 - Help
 - Used
 - Produced

Content	Annotations	Internal Comments	Public Comments	Access Rights	Dependencies	Statistics
x Astronomy x Curation x Databases						
Created 12/01/2012 by John Doe						
Updated 24/06/2012 by Alice Darn						
Authors: John Doe; Simon Wright; Alice Darn						
Groups: AMIGA Group						
Size: 12 MB 26 files including 3 workflows						
Description and Purpose						
Schema						
						
Download Edit Info ★★★★★ Citations: 52 Downloads: 342 Comments: 24						
Check State Completeness <div style="width: 50%; height: 10px; background: linear-gradient(to right, blue, white);"></div> Health 						



» Provide digital libraries with RO preservation functionalities

Curation of Astrophysical Quantities



Materials and Methods

RO Versions [Prev](#) | [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) | [Next](#)

Conceptual Relations

Research Object

- Gathering
 - Digital File
 - Hyperleda.tf2
- Web Services
 - Hyperleda
 - NED CS
- Input Data
 - ListCIGS
 - MorphoTypes
- Results
 - Luminosities
- Scripts
 - Calculation.py
- Propagation
- Comparison
- Update
- Bibliography
- Help
- Used
- Produced

Content Annotations Internal Comments Public Comments Access Rights Dependencies Statistics

x Astronomy x Curation x Databases

Download Edit Info

Created 12/01/2012 by John Doe
Updated 24/06/2012 by Alice Darn
Authors: John Doe; Simon Wright; Alice Darn
Groups: AMIGA Group
Size: 12 MB 26 files including 3 workflows
Description and Purpose

Schema

Check State

Completeness

Health

Search for trustable reusable material for commissioning (e.g from other disciplines)

Useful for complex processes as commissioning?

How Wf4Ever is addressing these challenges

Describe it: methods, data and all the elements involved.

Mechanisms for

- » Defining relationships between the elements
- » Annotating each element and the whole RO

Curation of Astrophysical Quantities



Materials and Methods

Conceptual Relations

- Research Object
 - Gathering
 - Digital File
 - Hyperleda.tf2
 - Web Services
 - Hyperleda
 - NED CS
 - Input Data
 - ListCIGS
 - MorphoTypes

Conte... Annotations Provenance Internal

Role	Type	Annotation	Who
26/01/12	Purpose	Gather info from Hyperleda	John Doe
14/01/12	Description	Annot. coming from Taverna	Taverna
10/01/12	FileType	Workflow	System

Metrics and tools for QA Evolution with time

Compile all the elements needed by the experiment

75%

Runnable

Service up, software working, etc.

60%

Repeatable

The output can be reproduced

20%

Shareable

Enough annotation.

50%

Publishable

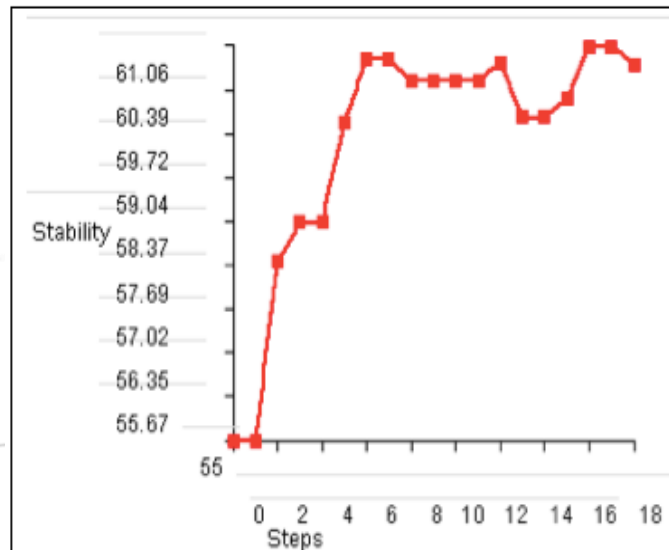
Service up, software working and all well commented

» Completeness

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

» 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

Metrics and tools for QA Evolution with time

How Wf4Ever is addressing these challenges

Be independent of the environment

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

» Decay

State of the services (up/down), of the applications (updated/deprecated), permissions to access the data

» Interoperability

- › RO level
- › Component level



Make the experiment discoverable

» Tracking

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

» Recommender Service

Which exploits semantic description, relations and other metadata to support advanced search mechanisms

» Collaboration Spheres

Visual mechanism to find similar elements (users, ROs, workflows) to others previously selected

Rating 

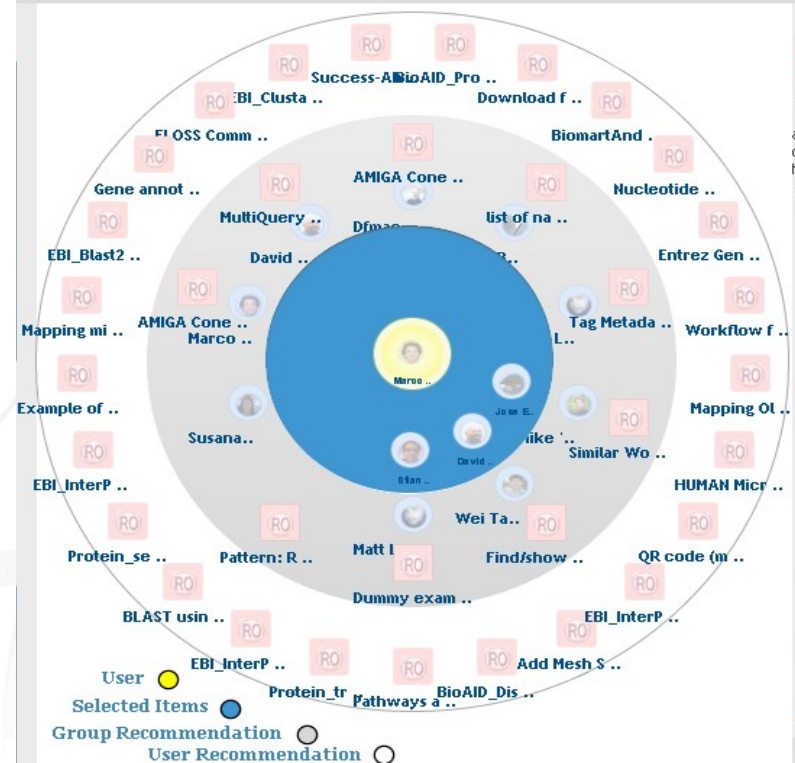
Downloads 36

Citations [2]

Re-used [1]

Comments [4]

Collaboration Spheres v.0.1



Large international teams to commission an experiment spread in two continents would be the most similar thing to a Social Network

How Wf4Ever is addressing these challenges

Make the experiment discoverable

Rating ★★☆☆☆

Downloads 36

Citations [2]

Re-used [1]

Comments [4]

» Tracking

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

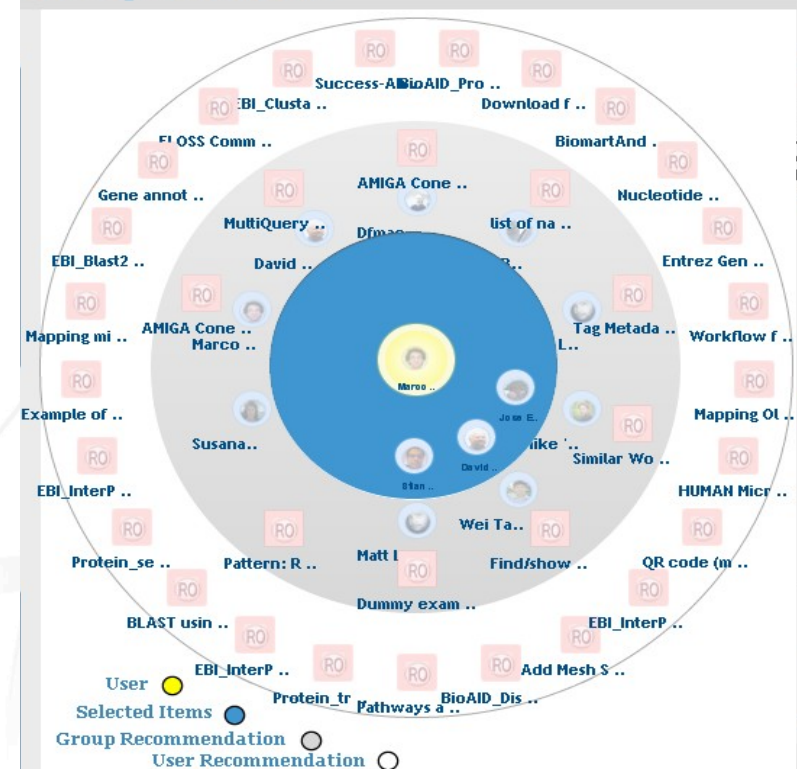
» Recommender Service

Which exploits semantic description, relations and other metadata to support advanced search mechanisms

» Collaboration Spheres

Visual mechanism to find similar elements (users, ROs, workflows) to others previously selected

Collaboration Spheres v.0.1



Building complex commissioning tools from existing ones

Impact in Astrophysics
Changing the way astrophysics work

- » AstroTaverna plugin <http://wf4ever.github.com/astrotaverna/>
- » Connection with Virtual Observatory

Taverna Workbench 2.4.0

Registry: <http://registry.euro-vo.org/services/RegistrySearch>

Keywords:

Cone Search SIA Search SSA Search

14 results for ConeSearch: amiga

Short na...	Title	Subjects	Identifier	Publisher
AMIGACS	AMIGA C...	[The AMI...	ivo://svo...	The AMI...
J/A+A/4...	The AMI...	[Position...	ivo://CD...	CDS
J/A+A/4...	AMIGA V...	[Galaxies]	ivo://CD...	CDS
J/A+A/4...	AMIGA III...	[Galaxies]	ivo://CD...	CDS
J/A+A/4...	AMIGA. I...	[Galaxie...	ivo://CD...	CDS
J/A+A/4...	AMIGA. I...	[Galaxies]	ivo://CD...	CDS

J/A+A/411/391: The AMIGA project. Revised positions for CIG galaxies (Leon+ 2003)

CATALOG

We present revised positions for the 1051 galaxies belonging to the Karachentseva Catalog of Isolated Galaxies (CIG, Cat. 82>). New positions were calculated by applying SExtractor to the Digitized Sky Survey CIG fields with a spatial resolution



- » Services for managing and visualizing VOTable
- » Astronomical utilities: coordinate transforms.

Service panel

Filter:

- ▼ Astro tools
 - Add Column - Add column using a expression
 - Add sky coordinates - Add sky coordinates
 - Cat n-tables - Cat a list of tables
 - Cat tables - Cat two tables
 - Check template filler - Check Template filler
 - Coordinates transformation - Coordinates transformation in a

This talk was an example of R-use!

<http://amiga.iaa.es/p/212-workflows.htm>

<http://www.wf4ever-project.org>



The R dimensions

Best practices for everything... and gRRRowing

Reusable. The key tenet of Research Objects is to support the sharing and reuse of data, methods and processes.

Repurposeable. Reuse may also involve the reuse of constituent parts of the Research Object.

Repeatable. There should be sufficient information in a Research Object to be able to repeat the study, perhaps years later.

Reproducible. A third party can start with the same inputs and methods and see if a prior result can be confirmed.

Replayable. Studies might involve single investigations that happen in milliseconds or protracted processes that take years.

Referenceable. If research objects are to augment or replace traditional publication methods, then they must be referenceable or citeable.

Revealable. Third parties must be able to audit the steps performed in the research in order to be convinced of the validity of results.

Respectful. Explicit representations of the provenance, lineage and flow of intellectual property.

Replacing the Paper: The Twelve Rs of the e-Research Record” on <http://blogs.nature.com/eresearch/>

Dave de Roure (Oxford e-Research Center)