Grant agreement no.: 27092

Curating and Preserving Collaborative Digital Experiments

Jose Enrique Ruiz
IAA-CSIC

May 19th 2011
2011 IVOA Spring Interop Meeting - Naples

SEVENTH FRAMEWORK PROGRAMME
European Commission
Information Society and Media
Information Society Technologies

Wf4Ever team

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

Astronomy research is entirely digital
Time has come to go "Beyond the PDF"

- Preserved experiments
- Methodology "in action"
- All data exposed
- Reproducible
- Repeatable
- Reusable
- Repurposeable
- Participatory
- Collaborative
- Formative

Wf4Ever goals

All components related to the research lifecycle should be available.

Preserved and easily retrievables

- Proposals
- Data
- Processes
- Workflows
- Publications

RESEARCH OBJECT

Research Objects in Astronomy

- Metadata (Author, Instrument, Research group, etc.)
- Description of the experiment (Strategy, Expected results, etc.)
- Observing proposal
- Auxiliary and raw data
- Reduced science-ready data
- Digital environment needed
- Scripting and software used
- Web services
- Scientific workflow
- Final data products
- Standard publication

Scientific Workflows

- Automation
- Repeatable
- Reproducible
- Encourage best practices
- Modular nature allows
  - Reuse
  - Repurpose
- Exposes the scientific method
- Formative
- Scientist friendly

Scientific Workflows

- Automation vs. The intrinsic exploratory nature of Science
- Documented vs. Hidden knowledge
- Web services vs. Local software
- On-line data vs. Local data
- Modular vs. Unstructured
- Open Science vs. Proprietary

- Preserved
- Classified and indexed
- Referenced and retrievable

Workflow preservation is complex

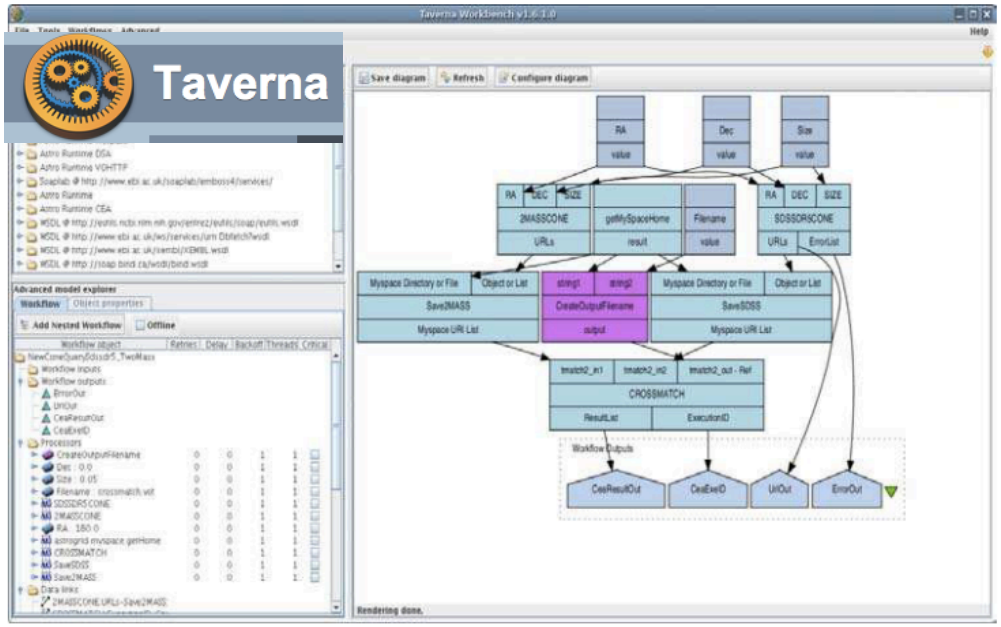
- Interpreted through their execution
- Complex models are required to describe them
- Provenance is a complex issue in a cloud of services
- Need of Web Semantics, Ontologies, Linked Data, etc..
- Resources are often beyond control of scientists

Scientific workflows: the cooking recipes

The oven

A workflow enactment and management system
University of Manchester

- AstroTaverna (AstroGrid)
  - SOAP
  - AstroRuntime
- Reflex (ESO)
- AladinJLOW Plugin (CDS)



Collaborative tools: "Le marché"

The recipes store

Oxford University

- Find workflows
- Share workflows and files
- Find people
- Build communities
- Publish packages
- Tag workflows
- Score and rate workflows
- Comment on workflows
- Write reviews

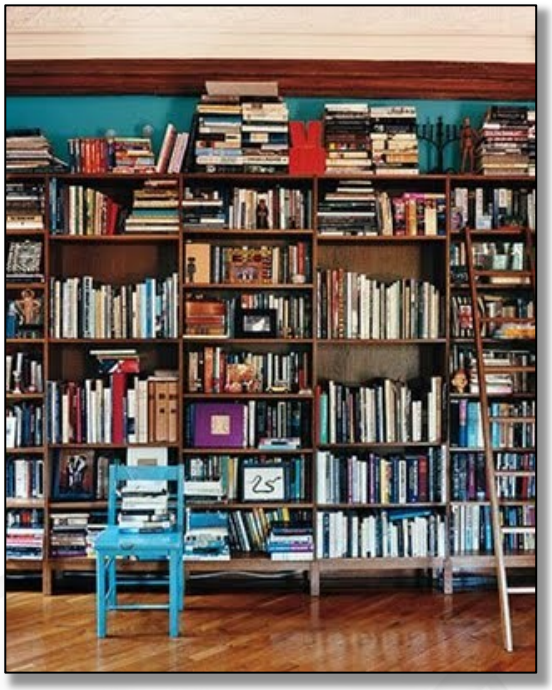
Living Working Research Objects

- Ubiquitous storing and computing
- Data archives and local data
- Web services and scripts
- Python based community
- VO standards

- Modular to reuse individual parts
- Access rights at different levels of granularity
- VOSpaces

Published Research Objects

- Archival
- Classification
- Indexing
- Retrieval
- Versioning



- Community reuse
- Rating, scoring and annotations
- Scalable in semantic repositories

- Permanent URIs, Linked Data, Semantics, etc.
- Interlink with catalogs/digital libraries

Users roles

Collaborator

Dealing with Living Working Research Objects in a research group.

Reader

Skims titles and abstracts of Published Research Objects.

Comparator

Looking for similar Research Objects to those she/he is working with.

Re-user

Extract modules from workflows and use them for his own purpose.

Publisher

Wants her/his work to be known.

Evaluator

He evaluates, rates, comments and recommend a specific Research Object.

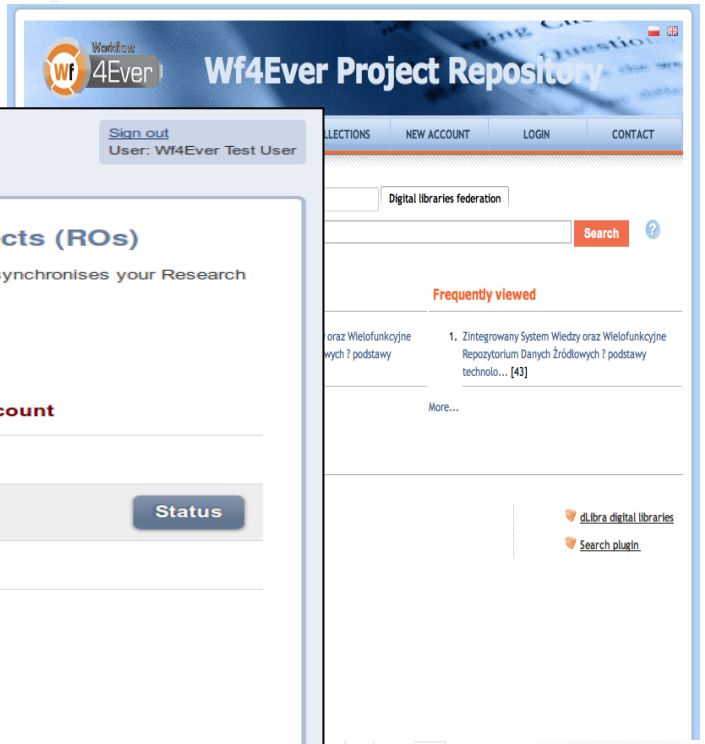
Most of them are **active roles** run the workflows with (different) data

ROBox: the basket

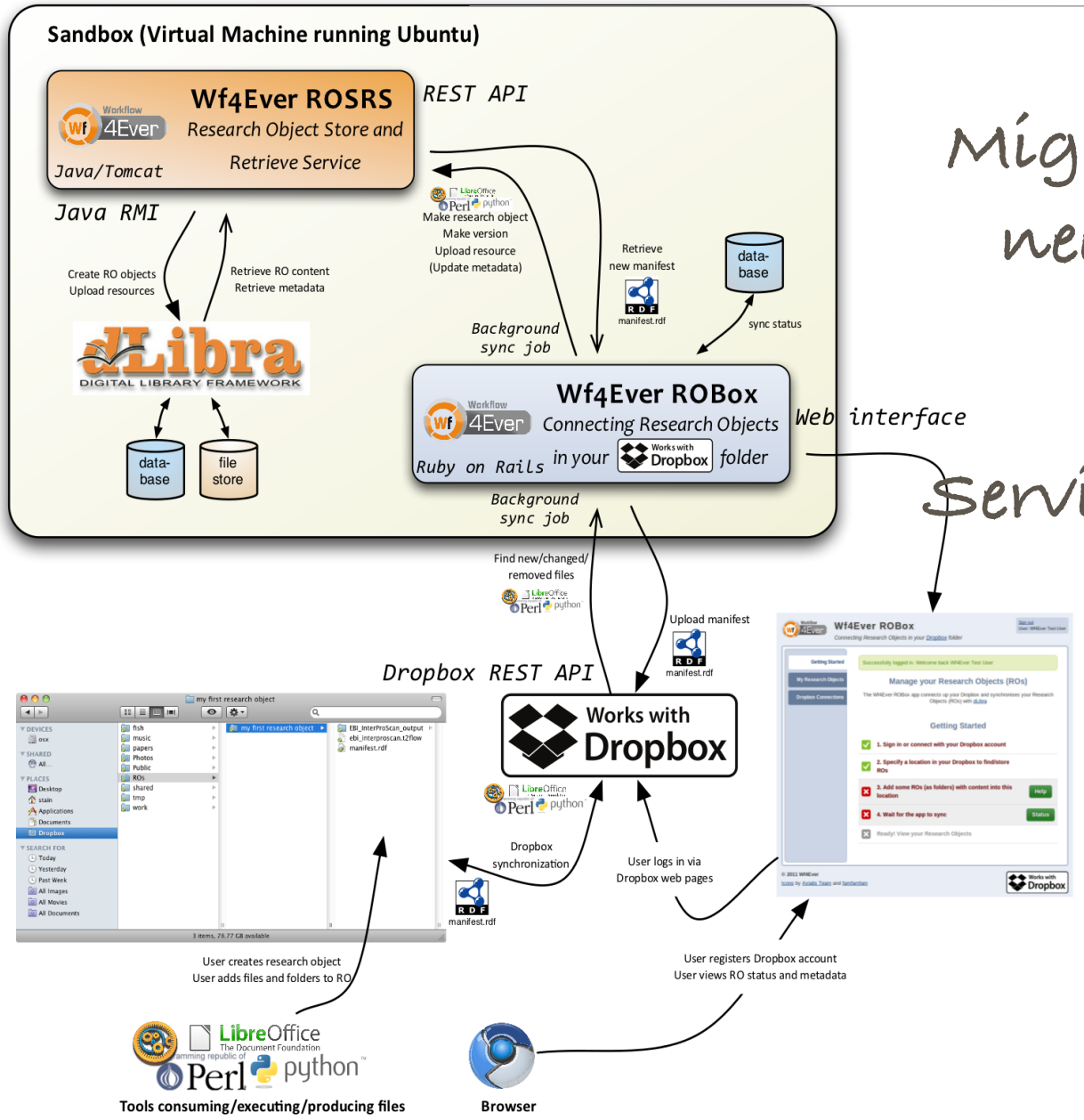
Seamless contribution to a collaborative platform

A shared folder in Dropbox becomes a Working Research Object

Automatic generation of metadata



Architecture of Prototype 1



Migration to VOSpace needed for Big Data Astronomy

Services should run where the data live

Open questions

We are moving into a world where
computing and storage are cheap and data movement is death.

In a Cloud of services and data, Web Services should benefit of the same privileges acquired by Data.

- Curation and preservation (identifiers)
- Discovery (semantics) of web services (linked "services"?)
- Characterization: input, outputs, functionality, etc.
- Copies (authenticity) or similar web services used as alternates
- Permissions, licenses, platform, costs, etc.
- Metrics for quality: popularity, use stats, logs uptime, etc.
- Versioning and authoring (referenced and acknowledged)