Workflow 4Ever

Grant agreement no.: 27092

Workflows Preservation

Data Curation and Preservation Session

Jose Enrique Ruiz
IAA-CSIC

October 17th 2011
2011 IVOA Fall Interop Meeting - Pune

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

Information Society
Technologies

Who am I ?

Instituto Astrofisica de Andalucia - CSIC

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

iSOCO enabling the networked economy

MANCHESTER 1824 The University of Manchester

UPM

SUPERCOMPUTING POZNAN NETWORKING

INSTITUTO de ASTROFISICA de ANDALUCIA CSIC I·A·A

LUMC

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 organizations

Technological Core Competencies

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

Case Studies

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

Goals

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

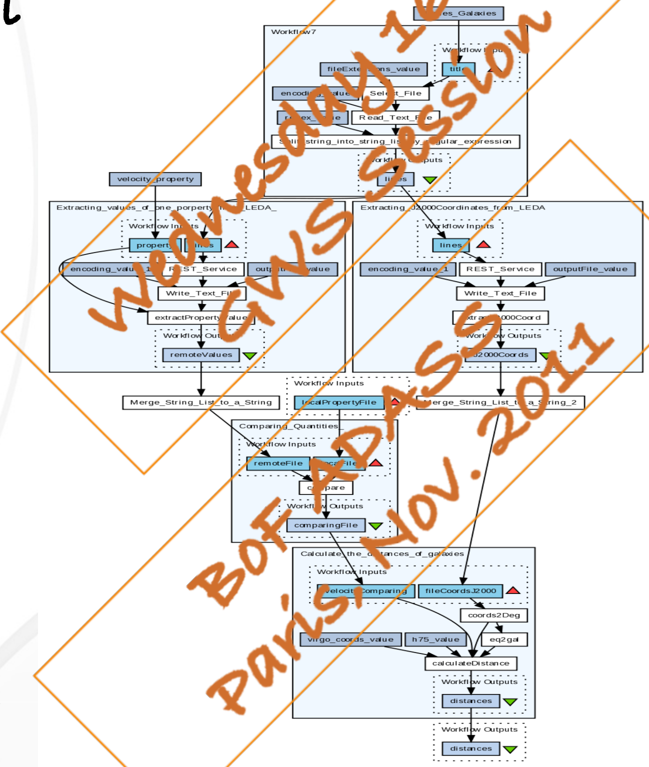
Creation of scientific communities to collaboratively share, reuse and evolve workflows and their parts, stimulating the development of new scientific knowledge

What are workflows ?

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

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)



Wednesday 16:15 GWS Session

BoF ADASS Paris Nov. 2011

(*) Scientifically exploitable results vs. scientific insight

Easily accessible and reproducible

Why workflow preservation is important ?

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

Preserved experiments

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

~~(Data)~~ Workflow preservation

- 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

~~(Data)~~ Workflow preservation

- Versioning of the whole or its components
- Restricted access on data and processes
- Permissions, licenses, platform, costs, etc.
- Semantic discovery of Wfs, processes, web services
- Metrics for quality: use stats, logs uptime, etc.

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

Preserve, Retrieve, Reconstruct, Replay

- Retrieve
  - Functionality of the Wf or its modules
  - What are the inputs and outputs
  - Authority…
- Reconstruct
  - Understand dependencies and components
  - Technical specificities
- Replay
  - Check the success of the preservation method

- Referenced and acknowledged

Characterization

Modeling

Tools

RO . The Research Object

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

Preserved and easily retrievable

- Proposals
- Data
- Processes
- Workflows
- Publications

LINKED

Wf4Ever Update

User Requirements

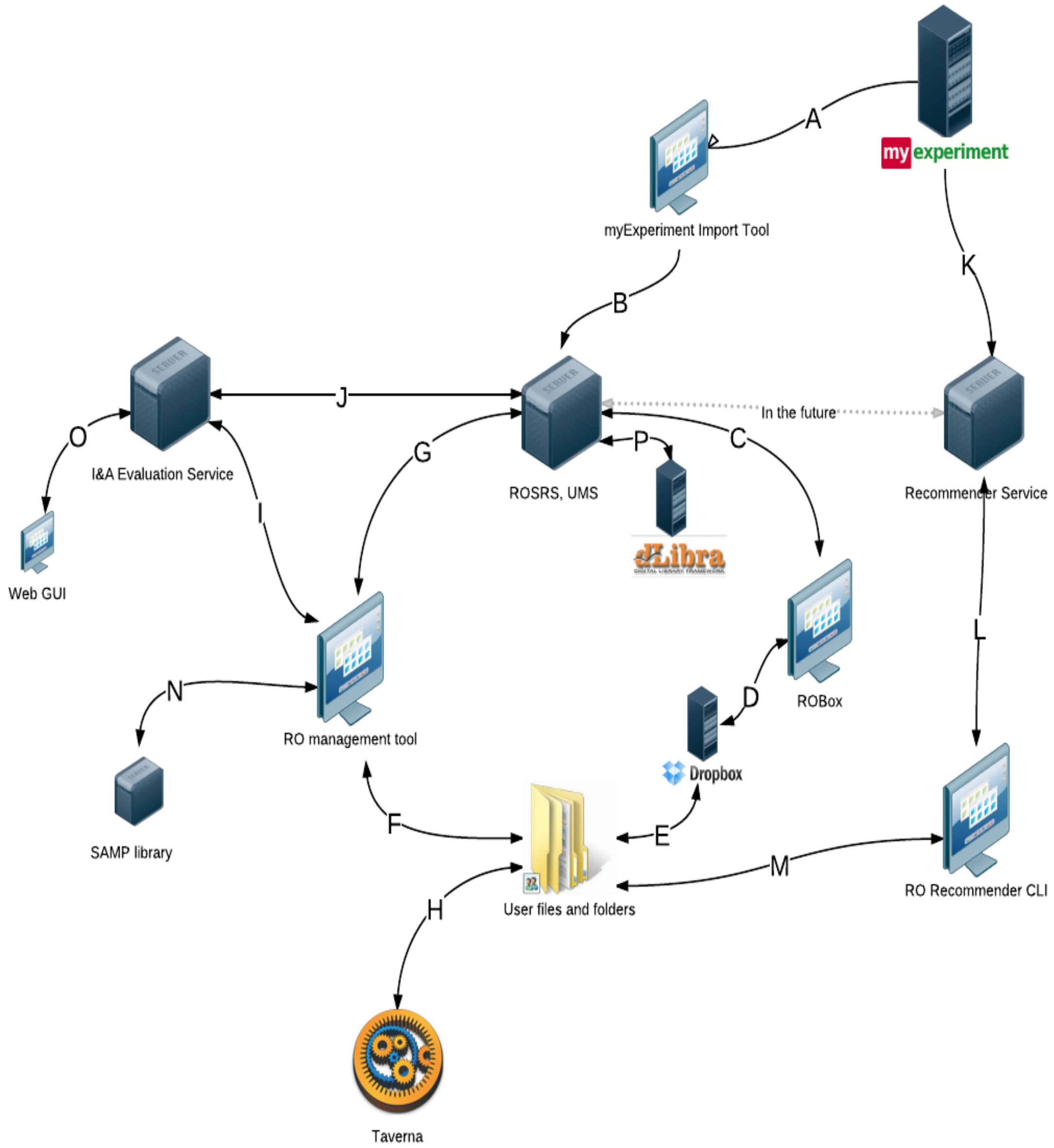
- Functional requirements for Wf4Ever "working" platform
- Focused on improving collaboration and reuse
- Interoperability in exchanging scientific methodology
- Expose experiment in a structured way to be understood by others

We need to build what we would like 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



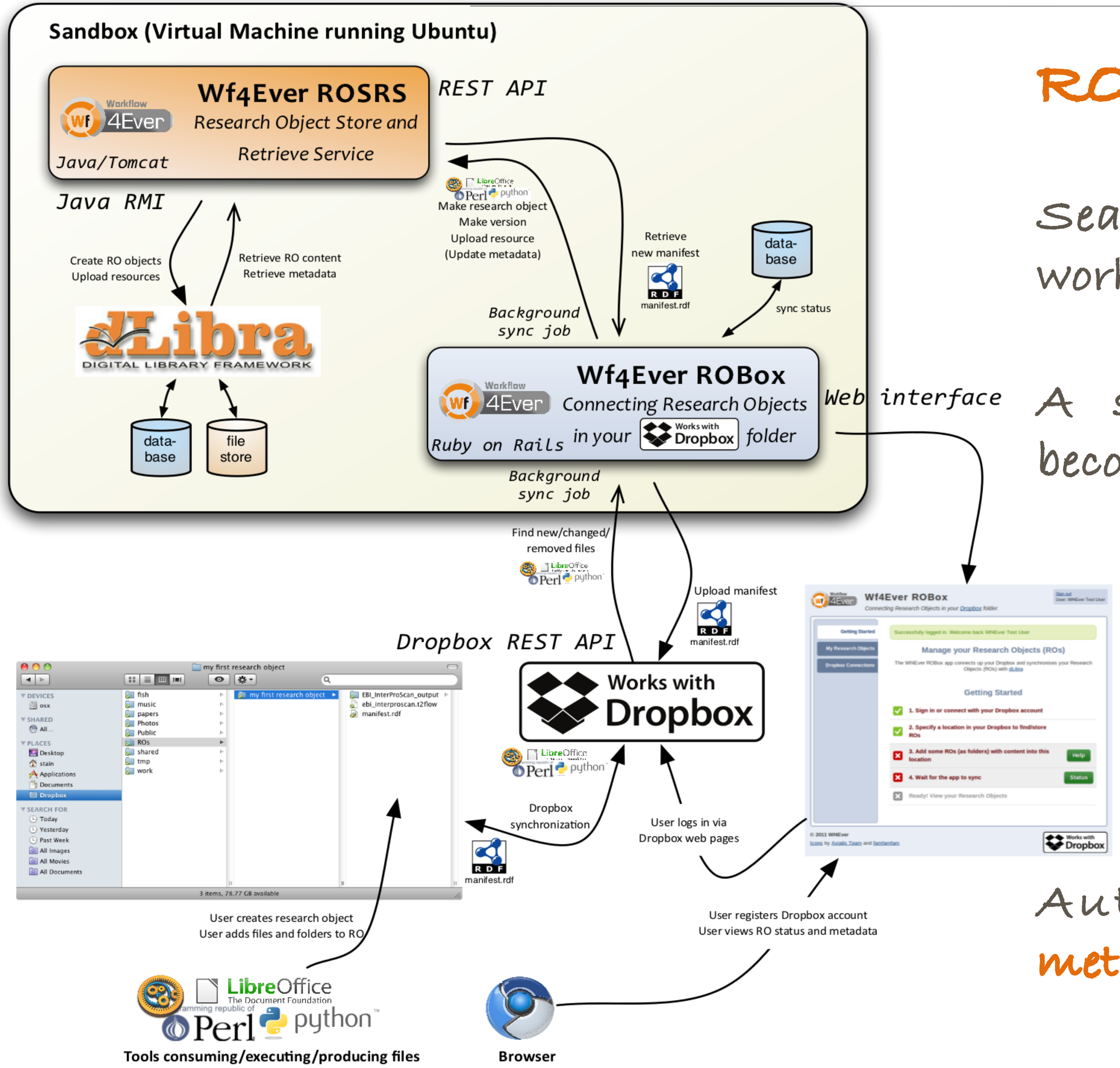
Architecture

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

User-Tools Prototypes

- RO Command Line Tool
- RO Annotator
- RO Box

Architecture of Prototype 1



Successfully logged in. Welcome back Wf4Ever Test User

Manage your Research Objects (ROs)

The Wf4Ever ROBox app connects up your Dropbox and synchronises your Research Objects (ROs) with dLibra

Getting Started

☑ 1. Sign in or connect with your Dropbox account

☑ 2. Specify a location in your Dropbox to find/store ROs

☒ 3. Add some ROs (as folders) with content into this location

☒ 4. Wait for the app to sync

☒ Ready! View your Research Objects

© 2011 Wf4Ever
Icons by Axialis Team and famfamfam

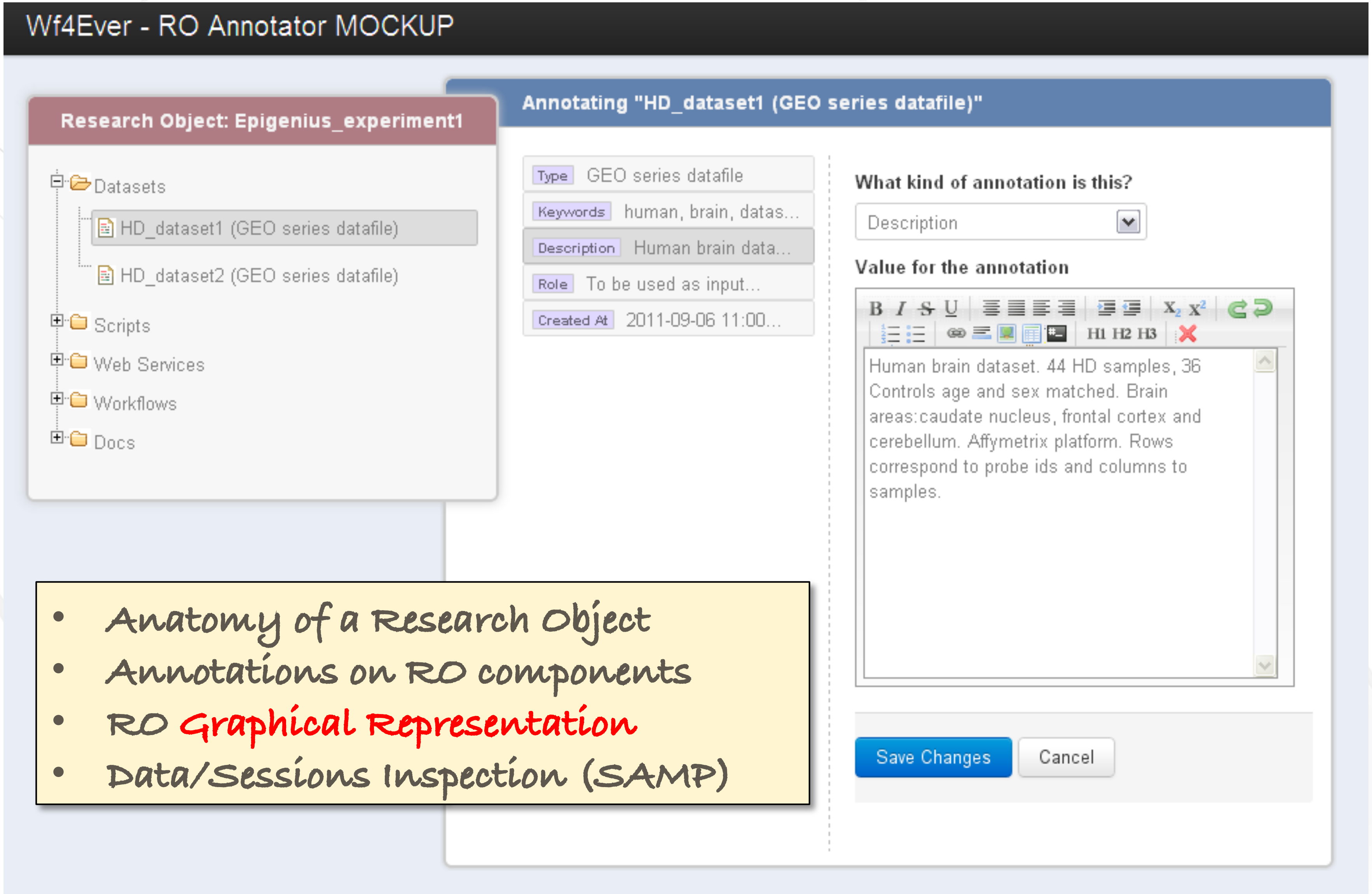
ROBox

Seamless contribution to a working collaborative platform

A shared folder in Dropbox becomes a Working RO

Automatic generation of metadata

Wf4Ever update



Wf4Ever - RO Annotator MOCKUP

Research Object: Epigenius_experiment1

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

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...

What kind of annotation is this?

Description

Value for the annotation

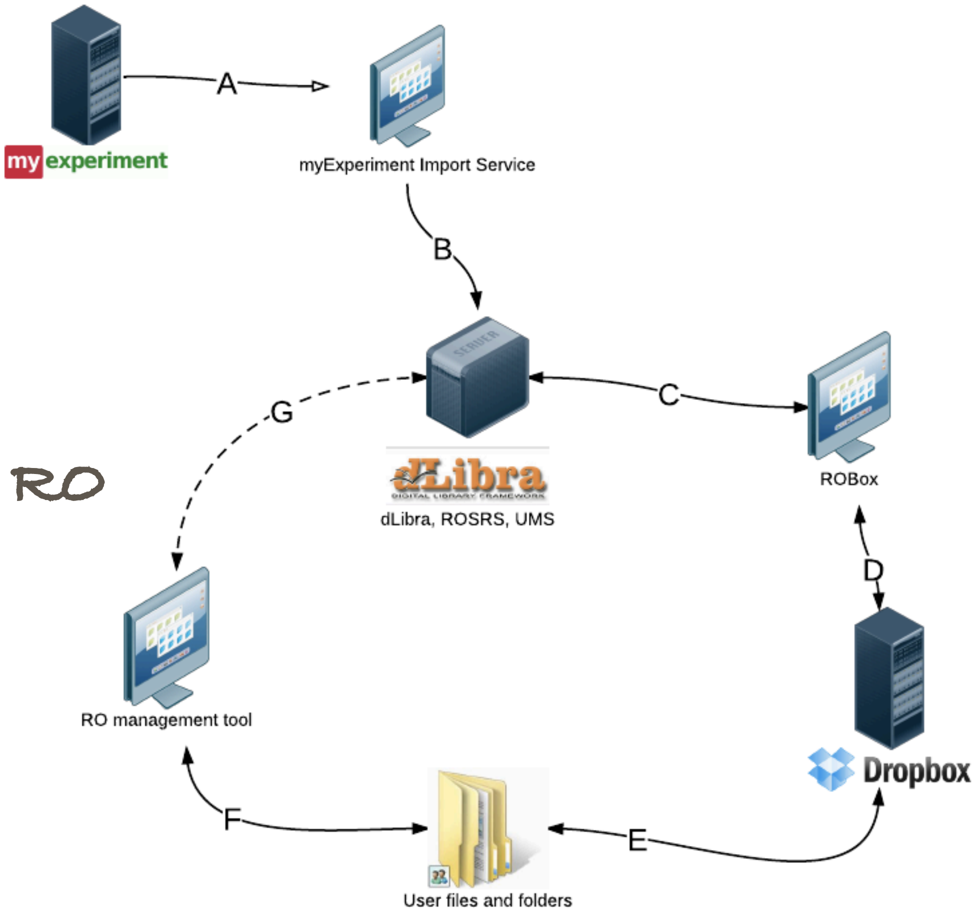
Human brain dataset. 44 HD samples, 36 Controls age and sex matched. Brain areas:caudate nucleus, frontal cortex and cerebellum. Affymetrix platform. Rows correspond to probe ids and columns to samples.

- Anatomy of a Research Object
- Annotations on RO components
- RO Graphical Representation
- Data/Sessions Inspection (SAMP)

Notification Service for Authors

What should be notified ?

- Fails
- Downloads
- Annotations
- Linked/Similarity
- Modifications on Working RO
- Acknowledgements



Notification Management Tool

Avoid spam

Related activities in the VO

US VAO

Work on semantic linking of proposals, publications, data

IVOA Working Groups

- Data Modeling
  Characterization, Provenance..
- Semantics
  Ontologies, Vocabularies, Annotations..
- Data Access Layer
  Self-descriptive Protocols..
- Grid and Web Services
  UWS, VOSpace, SSO..
- IG . Data Curation and Preservation
  Persistent Identifiers, Curation of VO Resources..

workflow@ivoa.net

IVOA Note

André Schaaff & Jose Enrique Ruiz