Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

Grant agreement no.: 27092

Data Curation & Preservation Session

Jose Enrique Ruiz
IAA-CSIC

8th December 2010
IVOA 2010 Fall Interoperability Meeting - Nara

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

The 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)

The Consortium Classified

Partners

- One SME
- Six public organizations

Technological Core Competencies

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

Major Sectors

- Education
- IT
- Astronomy
- Bioinfiormatics

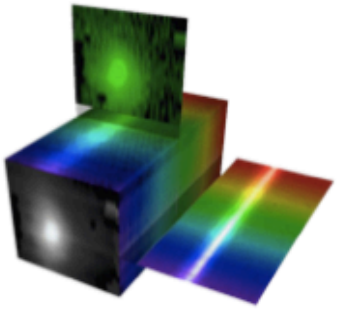
Case Studies

- Workflow Preservation in Astronomy (IAA)
- Workflow Preservation for Genome-wide Analysis and Biobanking (LUMC)

Workflow 4Ever

VO Scientific Workflow

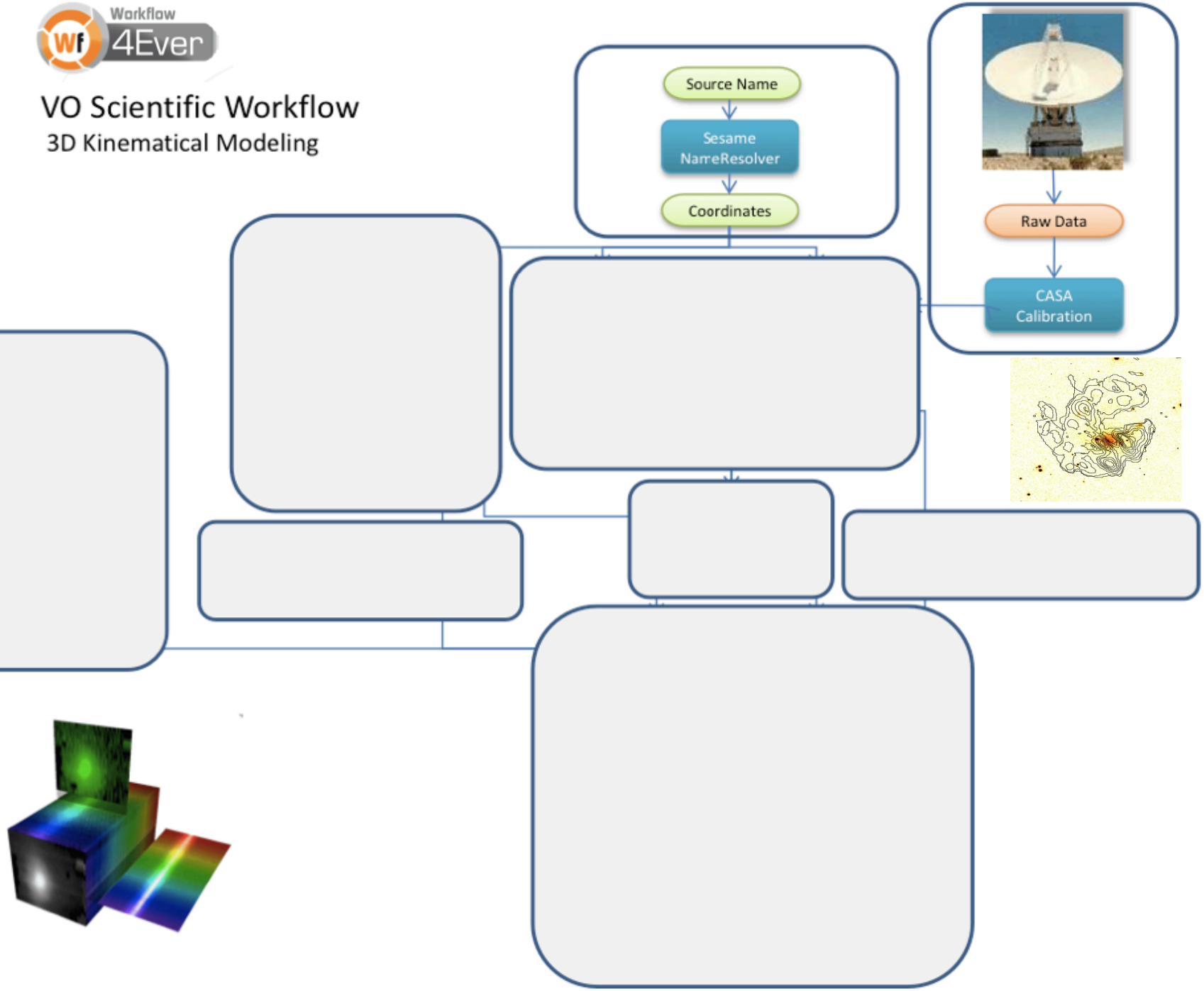
3D Kinematical Modeling



Workflow 4Ever

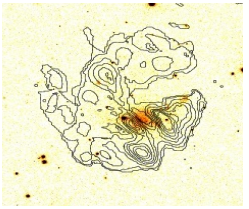
VO Scientific Workflow

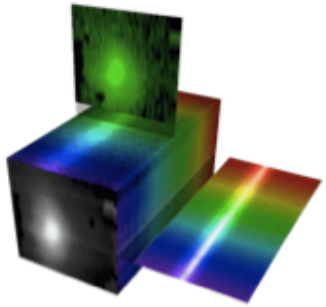
3D Kinematical Modeling



VO Scientific Workflow

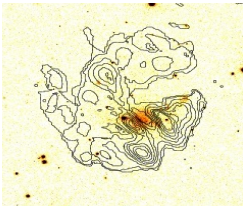
3D Kinematical Modeling

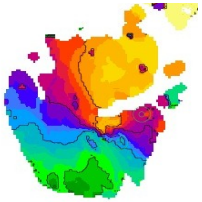


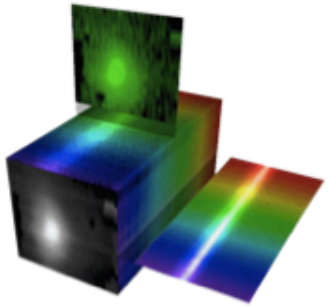


VO Scientific Workflow

3D Kinematical Modeling

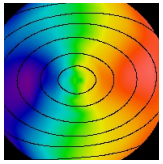


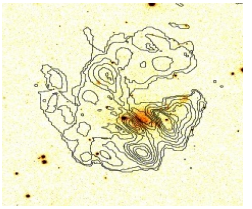


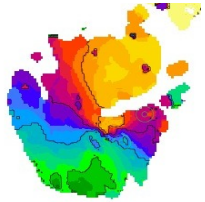


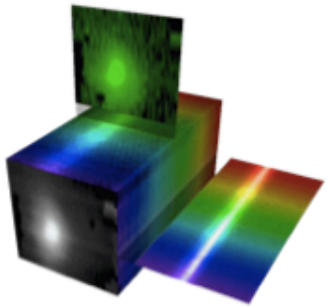
VO Scientific Workflow

3D Kinematical Modeling









Source Name

Sesame NameResolver

Coordinates

Raw Data

CASA Calibration

VO Catalog

CO ALMA 3D VO Archive Calibrated Data

Center, Inclinations, etc.

CO Datacube

3D VO Archive Modeled Data

VO Publishing

Gipsy 3D Modeling

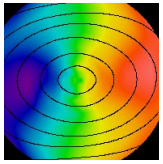
Subtraction

3D CO Residuals

Modeled CO Datacube

VO Scientific Workflow

3D Kinematical Modeling



Source Name

Sesame NameResolver

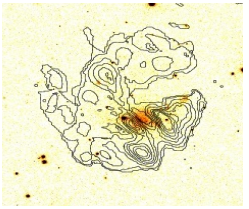
Coordinates

Raw Data

VO Catalog

CO ALMA 3D VO Archive Calibrated Data

CASA Calibration



3D VO Archive Modeled Data

Center, Inclinations, etc.

CO Datacube

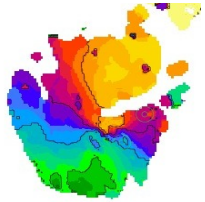
Gipsy 3D Modeling

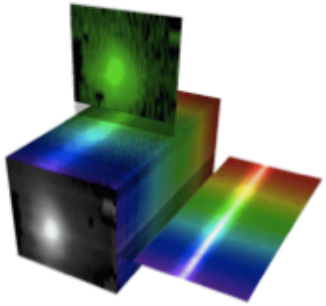
Subtraction

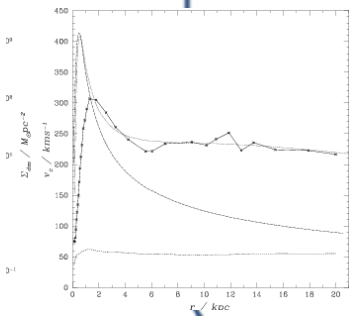
3D CO Residuals

VO Publishing

Modeled CO Datacube







Rotation Curve Extraction

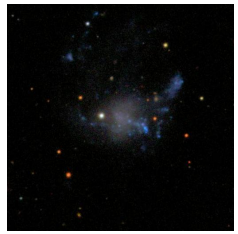
CO Rotation Curve

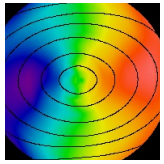
CO Dark Matter Model

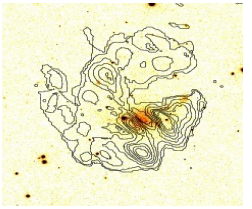
Workflow 4Ever

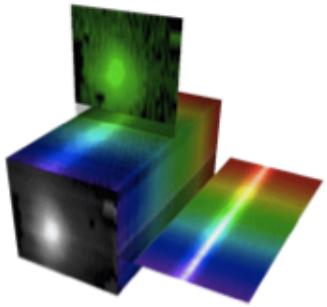
VO Scientific Workflow

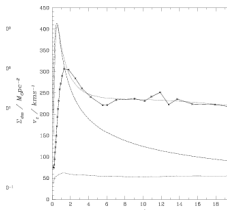
3D Kinematical Modeling

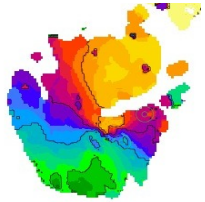


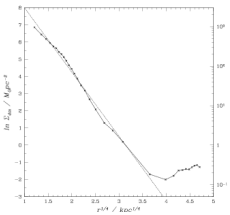












Source Name

Sesame NarreResolver

Coordinates

Raw Data

CASA Calibration

HIALMA 3D VO Archive Calibrated Data

VO Catalog

CO ALMA 3D VO Archive Calibrated Data

3D VO Archive Modeled Data

HI Datacube

Center, Inclinations, etc.

CO Datacube

Gipsy 3D Modeling

VO Publishing

3D HI Residuals

Subtraction

Subtraction

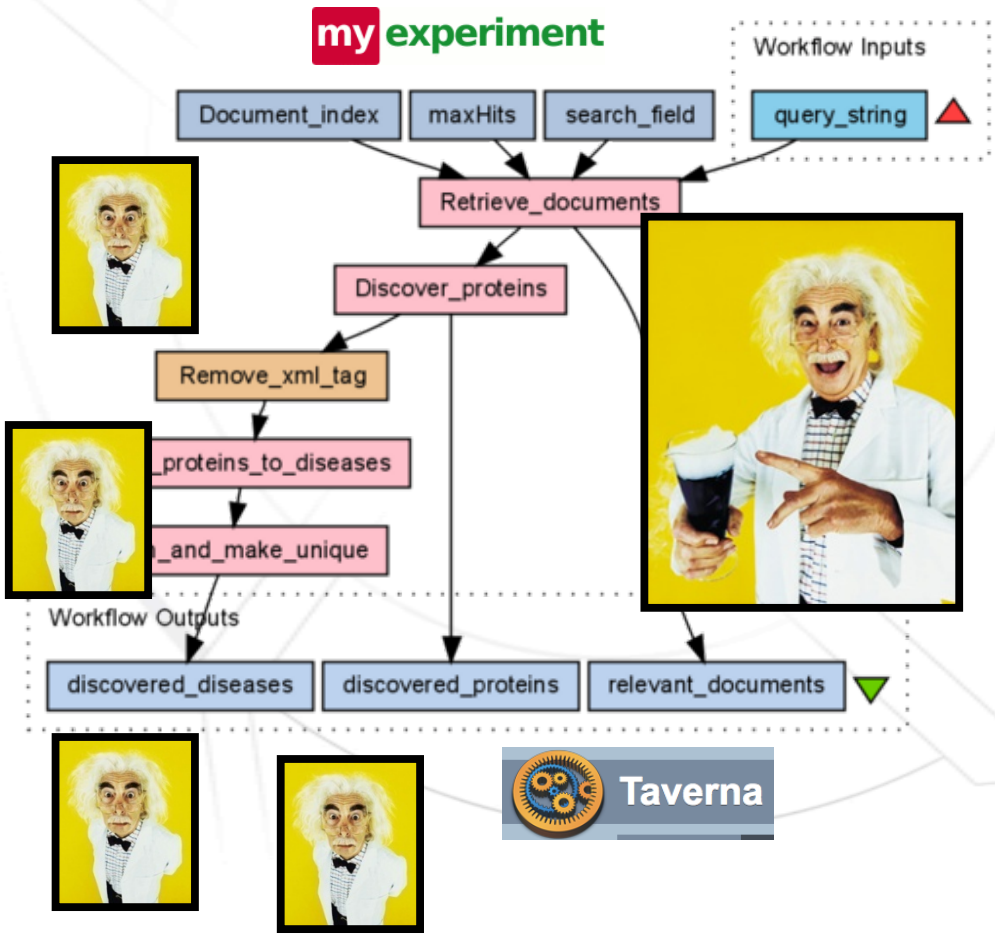
3D CO Residuals

Modeled HI Datacube

Modeled CO Datacube

YOUR FAVOURITE MASS MODEL TOOL

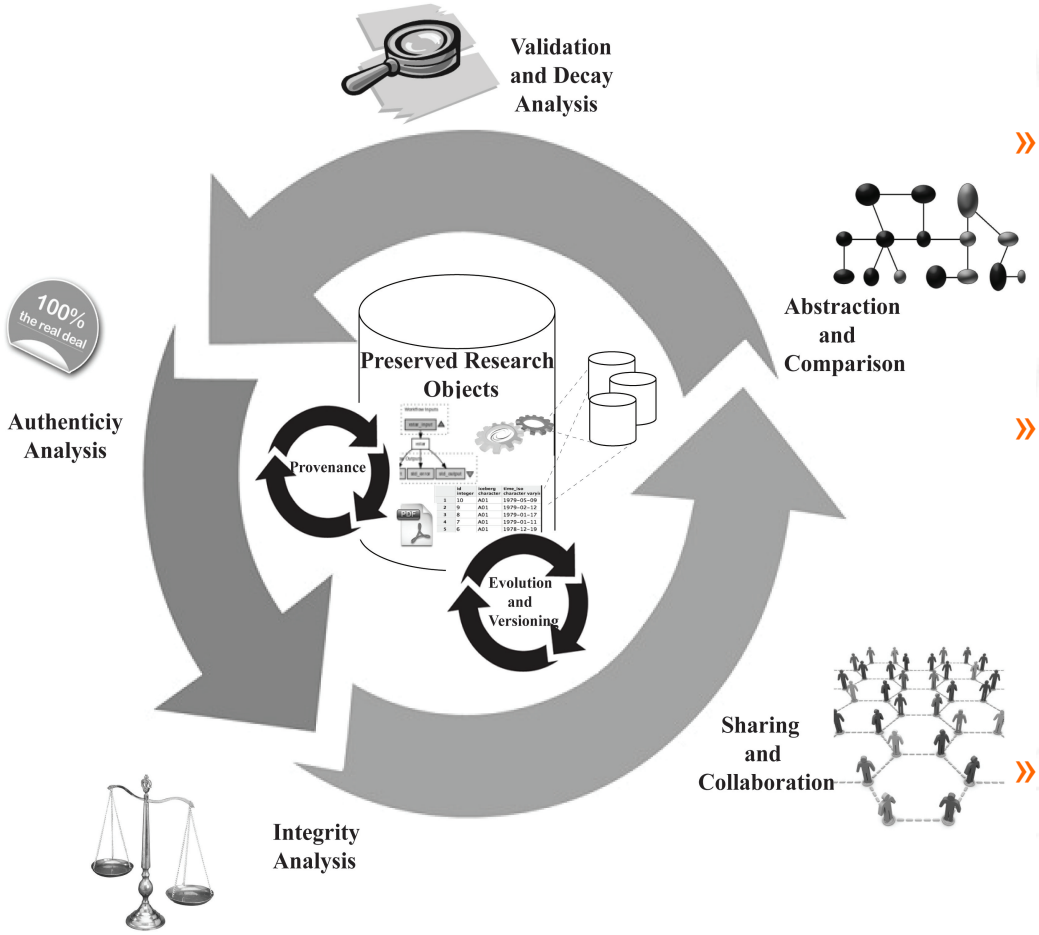
A **Scientific Workflow** can be seen as the combination of data and processes into a configurable, structured set of steps that implement semi-automated computational solutions in scientific problem-solving



- Central in experimental science
  - Enable automation
  - Make science repeatable (and sometimes reproducible)
  - Encourage best practices
- Scientist-friendly
  - Aimed at (some types of) scientists, possibly even without strong computational skills
- Communities: need for scientific data preservation
  - Enhance scientific development by building on, sharing, and extending previous results within scientific communities
- However, workflow preservation is especially complex
  - Workflows not only specified statically at design time but also interpreted through their execution
  - Complex models are required to describe workflows and related resources, including documents, data and services
  - Resources often beyond control of scientists

Goals

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



» Creation and management of complex **Research Objects** that take into account the dual nature (static and dynamic) of scientific workflows

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

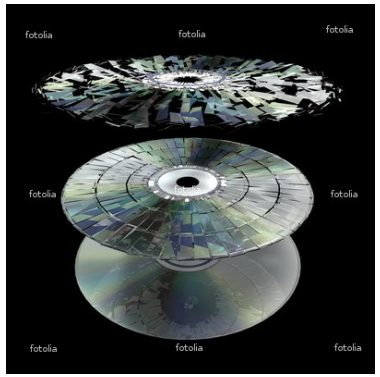
» Creation of scientific communities to collaboratively share, reuse and evolve workflows and their parts, stimulating the development of new scientific knowledge

Definitions

Integrity

» The quality or condition of being whole, complete and unaltered

» Crucial for ensuring the quality of preserved data in Research Objects



Authenticity

» Authenticity has a twofold dimension: data origin and entity authenticity

» Data origin: Proof of the origin of data, their genuineness, trustworthiness and realness

» Entity: ensuring that an entity, e.g. a person or other kind of actor, is the one it claims to be

Objectives

Evaluate and preserve the integrity and authenticity of archived Research Objects

- To ensure the data can be accessed and interpreted unchanged, complete, and correct today and in the future
- To preserve the integrity of archived Research Objects by tracking and verifying changes in archived objects as well as related resources
- To assist scientists in anticipating potential inconsistencies caused by uncontrolled changes in such resources
- To verify and proof the authenticity of authors and contributors to Research Objects as well as of internal and related resources

Provenance-based means to calculate measures of integrity and authenticity

Common needs from the Community

What do you want to know when accessing a workflow ?

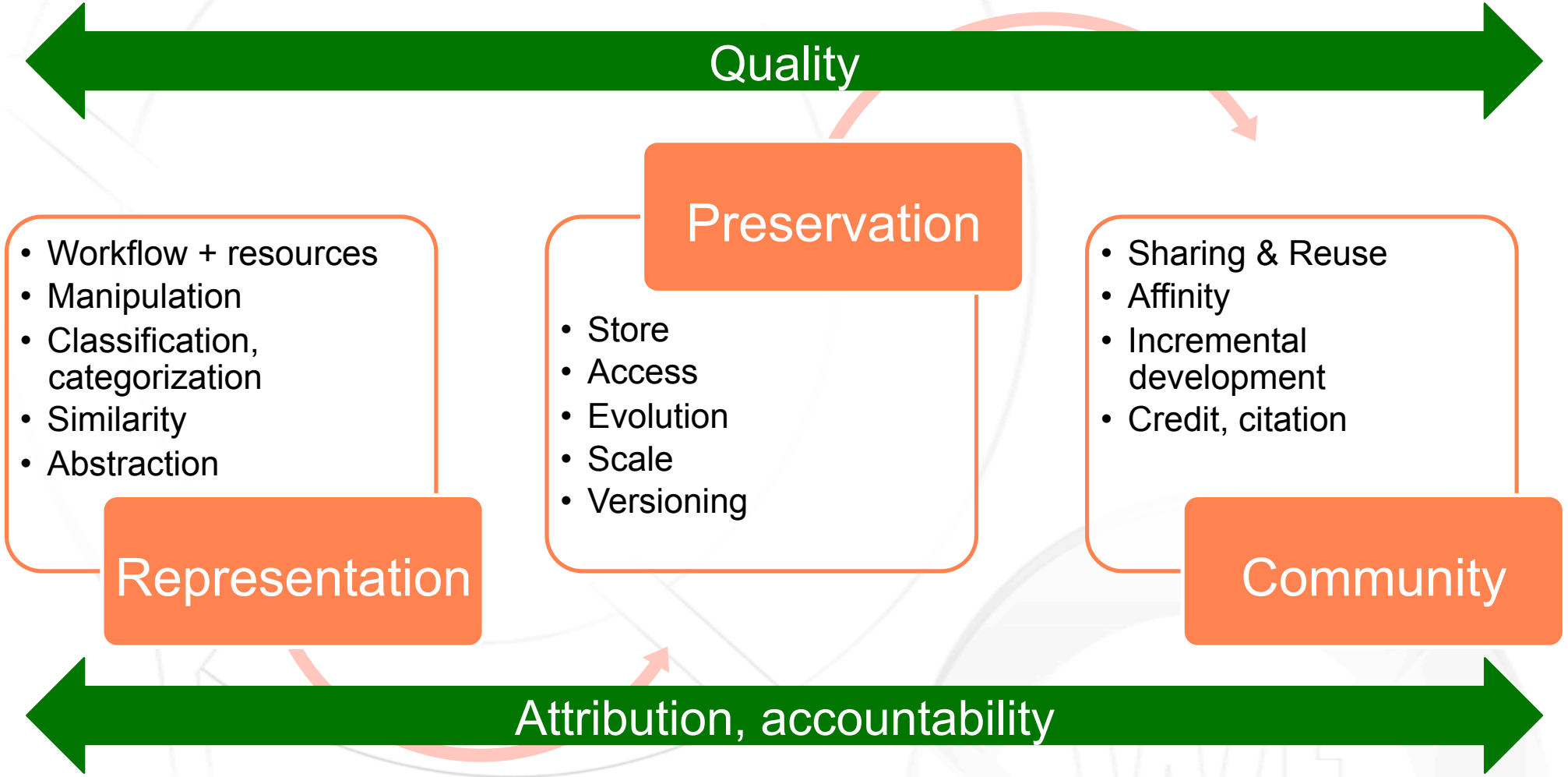
- If I can use it for my purposes (in my words)
- If I can expect it to run, when it was last run, by whom
- What it does quickly, by one of
  - example input / output (and trying it)
  - a description
  - ‘reading’ its key parts
  - what is was used for
  - related workflows
  - its creator
  - contacting the creator or last user
- How I need to cite the author and workflow

Common needs from the Community

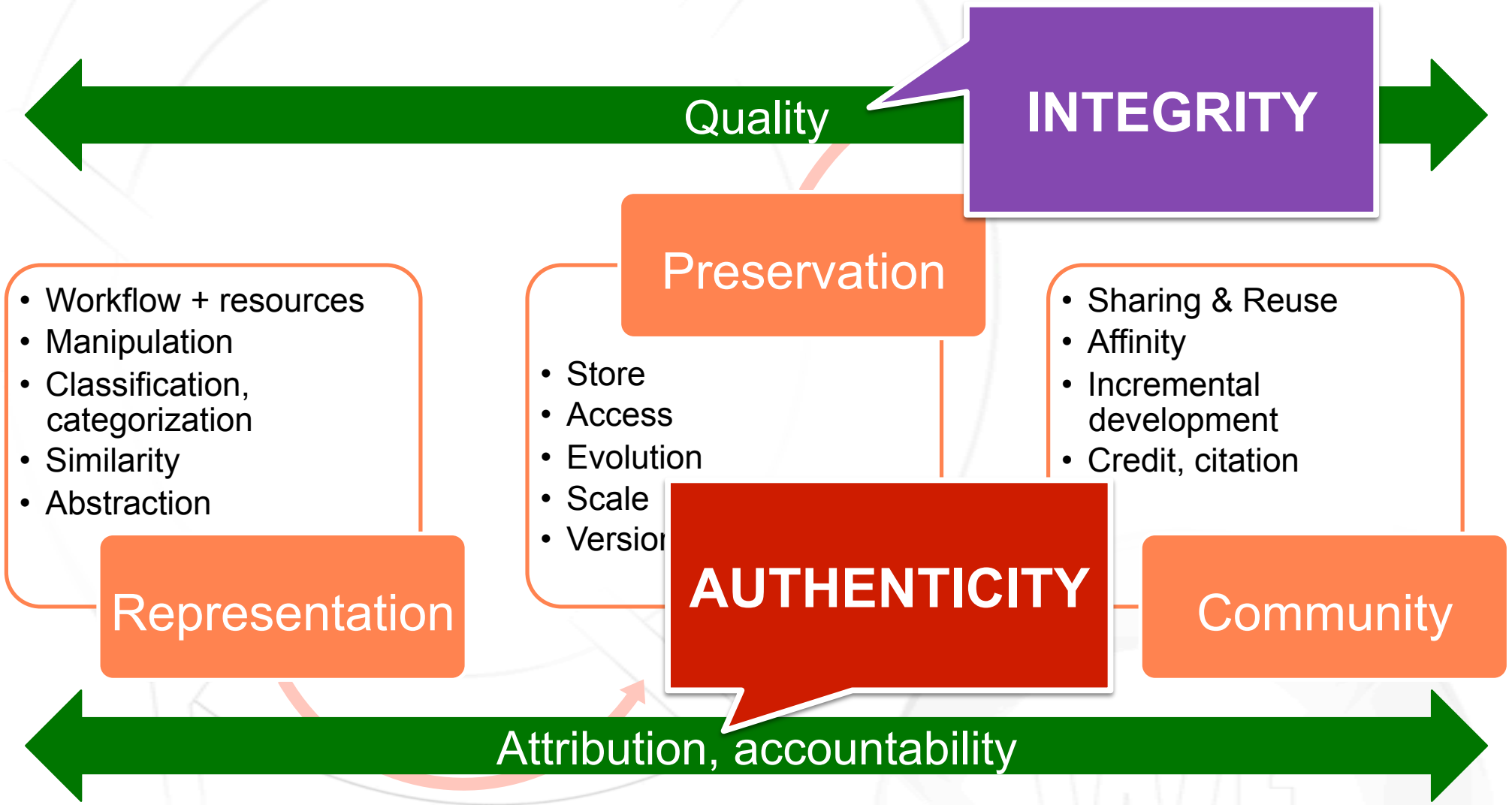
What do you want to know when sharing a workflow ?

- What rights others have
- What a good workflow is to get a good score
  - Make my workflow findable, reusable, and ready for review
  - Instructions to authors
  - Two types of contributions: serious science, preliminary/playing around
- If my workflow may have issues
- What the system or other users think it does
  - How it relates to other things
- Share freely or anonymously upon request

Main Challenges



Main Challenges



Wf4Ever Case Studies

Two workflow-intensive scientific case studies in the domains of Astronomy and Genomics

Astronomy

- Application area: Virtual Observatory (VO) data processing
  - Astrophysical quantities propagation
  - Source extraction on CCD images
  - Modeling of galaxy 3D data
- Focus on bringing workflow-based methodologies into Astronomy
- Creation of Golden Exemplars
- Beachhead

Genomics

- Application areas: Biobanking and Genome-Wide analisys
  - Interpretation of GWAS data
  - Gene expression studies
- Focus on authenticity and experimental reproducibility
- Community! Lots of available workflows
  - myExperiment
  - SysMO-DB
- Long tradition of workflow application

Overall goals

- To collect and preserve existing workflows and their related objects in each area
- To create scientific communities around the use and preservation of scientific workflows
- To apply, evaluate, and provide feedback on the results obtained from system and component-level research

Main Goals

WP5: Workflow Preservation in Astronomy

Scientific contribution

»Development of an online community of scientists working on Astronomy
»Introduction of workflow and workflow preservation needs in Astronomy and the Virtual Observatory
»Provide a set of workflows for frequently used complex task-combinations and demands in the Astronomy domain

Technological contribution

»Online repository of preserved Astronomy workflows identifying preservation needs
»Creation of three Golden Exemplars workflows:

- using Wf4Ever results
- involving additional implementations for wrapping VO Web services

Workflow-based methodology deployed in the VO community through exemplars and preservation methodologies

Questions

http://www.wf4ever-project.org