Research Objects in Wf4Ever

Jose Enrique Ruiz
jer@iaa.es
On behalf of the Wf4Ever Team

October 25th 2012
2012 IVOA Fall Interop Meeting - Sao Paolo

2011 - 2013

Wf4Ever
Advanced Workflow Preservation Technologies for Enhanced Science

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
5. University of Oxford and OeRC (OXF, UK)

6. Instituto Astrofísica Andalucía (IAA-C
7. Leiden University Medical Centre

Reproducible Science

The University of Manchester

MANCHESTER 1824

INSTITUTO de ASTROFISICA de ANDALUCIA CSIC IAA

POZNAN SUPERCOMPUTING NETWORKING

iSOCO enabling the networked economy

LUMC

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Astronomy Research Lifecycle

Astronomy research lifecycle is **entirely digital**

- Observation proposals
- Data reduction pipelines
- Analysis of science ready data
- Catalogs of objects and data
- Publish process

  - Final data results
  - Experiment in DL ADS/arXiv

Reproducible research is still not possible in a digital world

Efficient use of rich data infrastructure (VO) may be improved

A normalized preservation of methodology is needed

Tools

Efficiency and Reuse

Optimize return on investments made on big facilities

- Avoid duplication of efforts and reinvention
- How to discover and not duplicate ?
- How to re-use and not duplicate ?
- How to make use of best practices ?
- How to use the rich infrastructure of data ?
- **Intellectual contributions are encoded in softw**

More data in archives does not imply more knowledge

- Time has come to go beyond the PDF
- Expose complete scientific record, not the story
- Allow easy discovery of methods and tools

Barriers to Data and Code Sharing in Computational Science

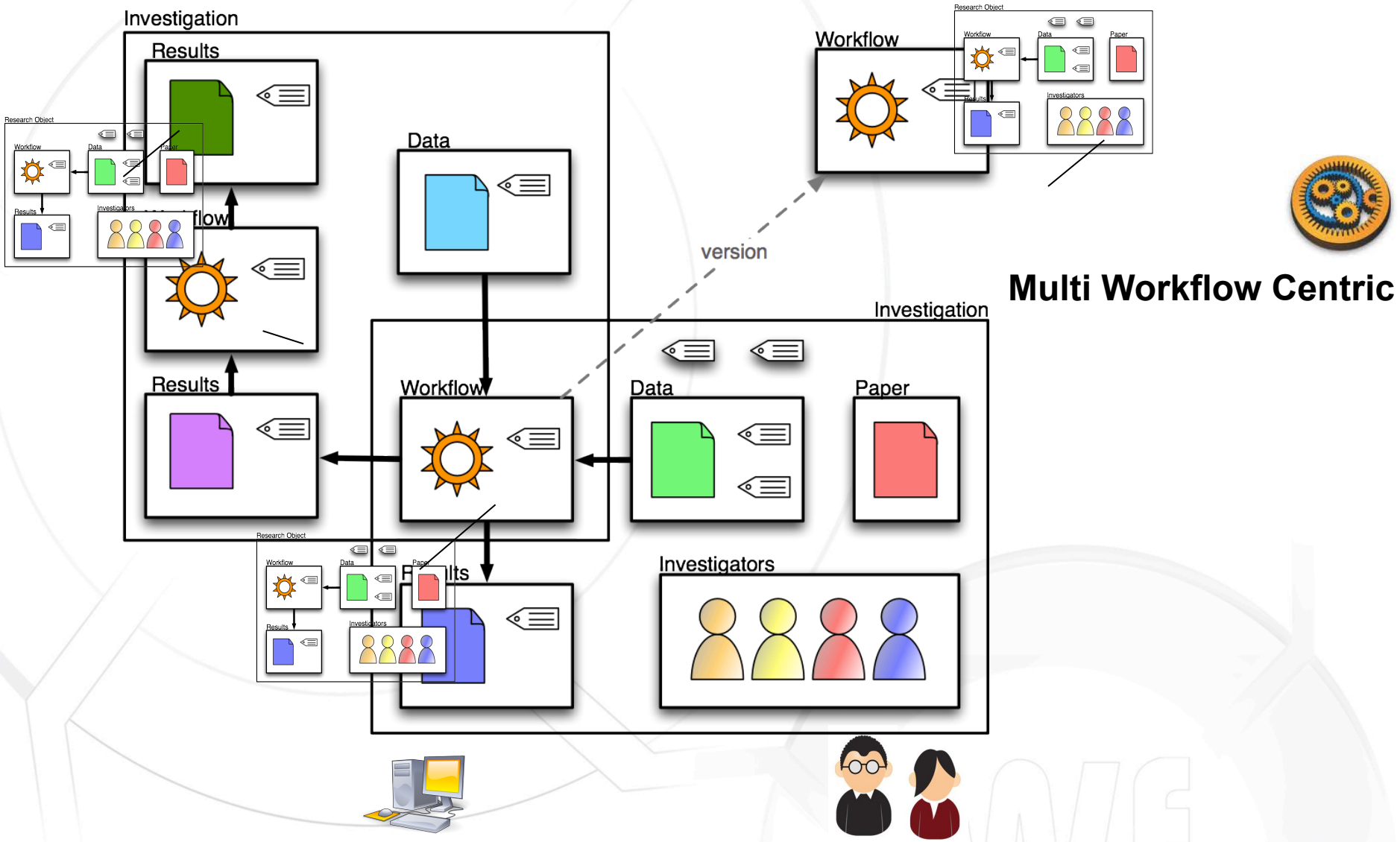
Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

I don't know how

Tools

Research Objects in Wf4Ever



RO Content

- Process (workflows), data, external resources and bibliography
- Execution environment set-up and local software dependencies
- Experimental protocol followed
- Roles, types and relationships among all digital components
- Provenance of intermediate and final results
- Decomposable attribution and authoring
- Fine-grained access control and permissions
- Example datasets for demonstration, reproducibility, monitoring, etc

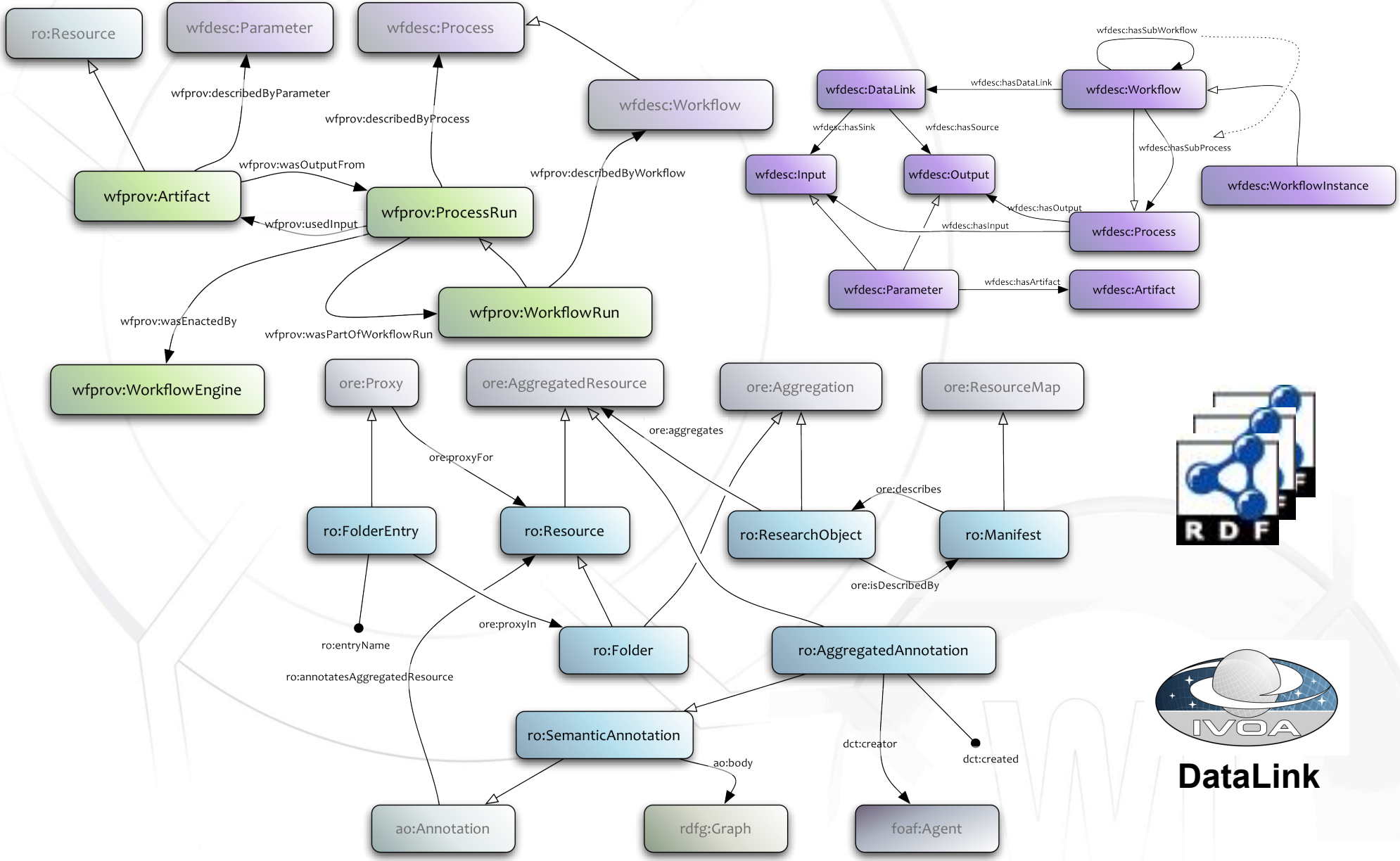
RO Template

- Placeholders to ease the aggregation process
- Completeness checking/quality assessment

Semantic Annotations

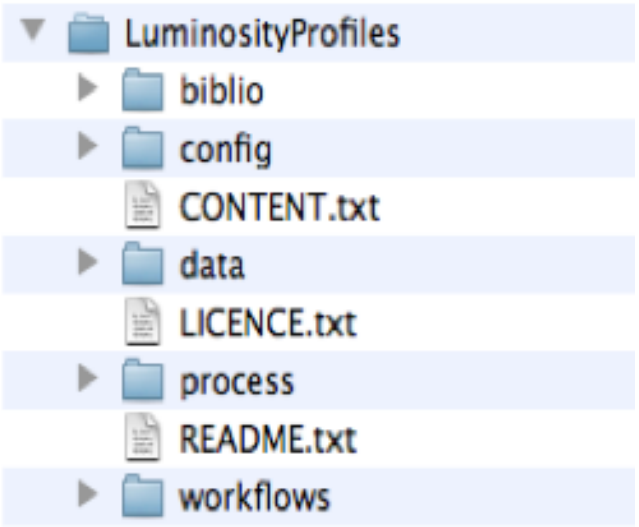
- Author of an **annotation**
- **Author and co-authors** of a **workflow**; reference link to a re-used workflow and its author
- Who has performed the **execution** of a workflow leading to the results provided in the RO
- Computing execution environment of the RO and local software **dependencies**
- Special access requirements to web services
- Datasets **provider**: person, webpage, survey, data release, etc.
- How much **time** does it take to run a workflow using the full data and the provided subsample
- The number of **elements** of the sample dataset where one workflow and/or RO iterates
- Previous and subsequent workflows to be executed, as in the experimental **protocol**
- Research institution, country, and scientific domain of the RO
- The actual **size** of the RO and/or a folder
- The version of a workflow

Research Object Wf4Ever Semantic Model



DataLink

Luminosity Profiles RO



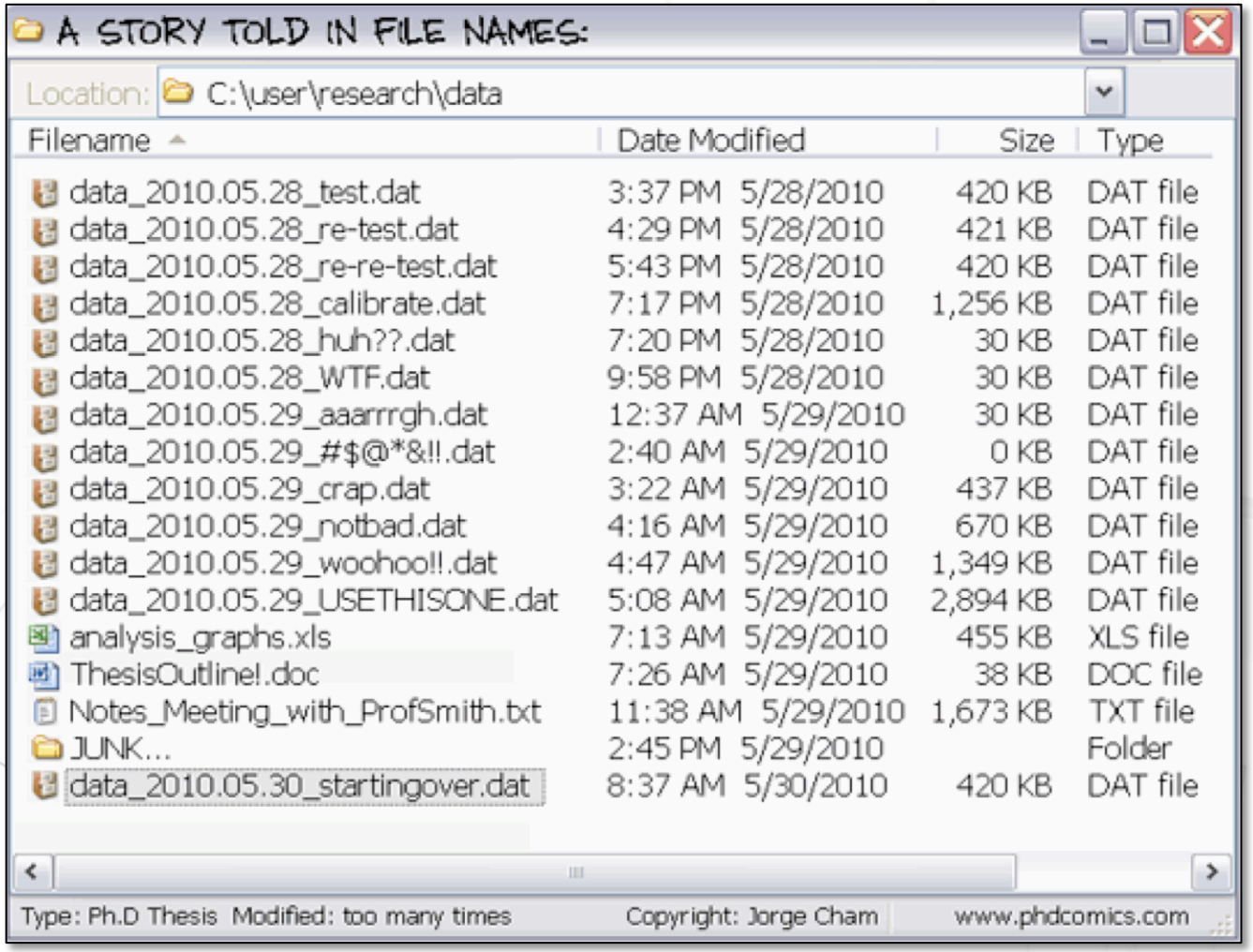
1010 Files, 200 MB
External Sources ~ 8 GB

5 Main Workflows, 14 Nested Workflows, 25 Scripts, 11 Configuration files
10 Software dependencies, 1 Web Service

Dataset: 90 galaxies observed in 3 bands

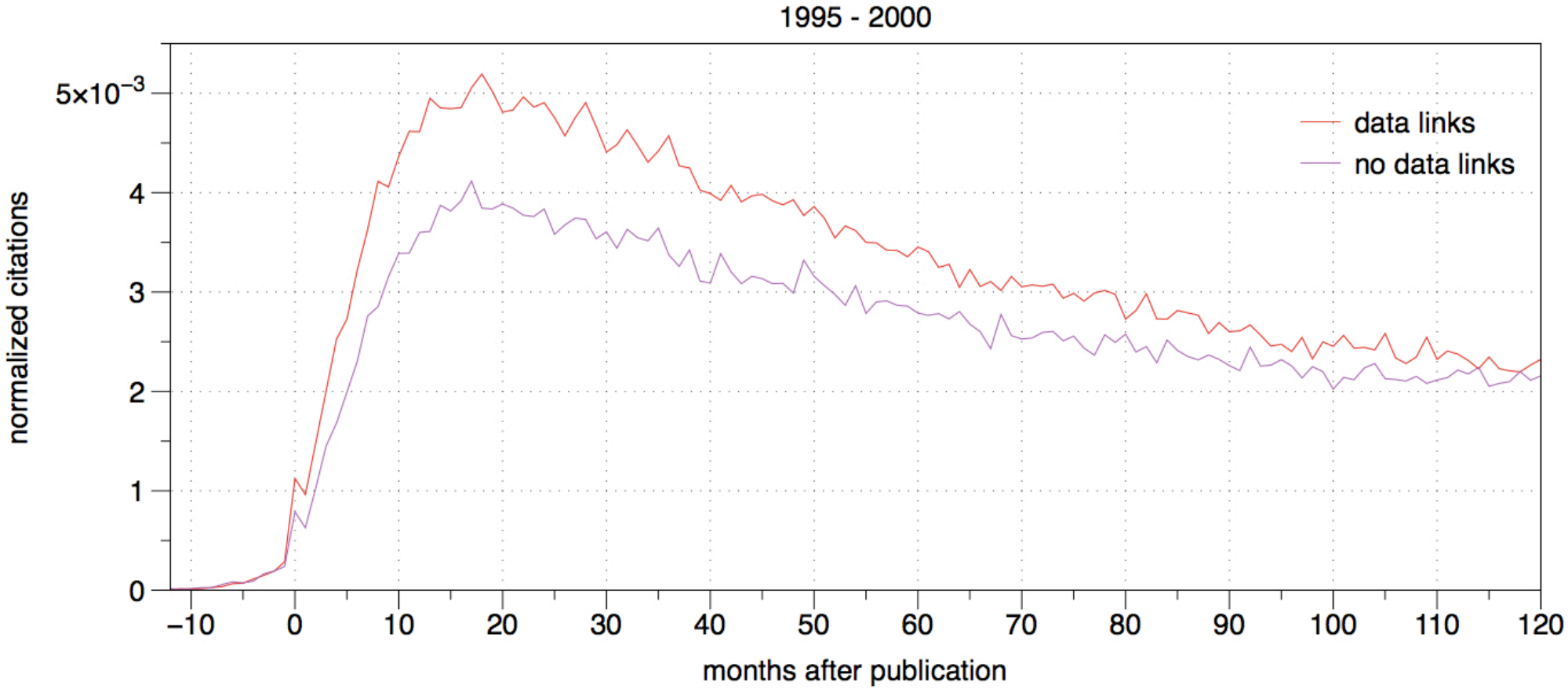
Reproducibility

When organization is better than automation



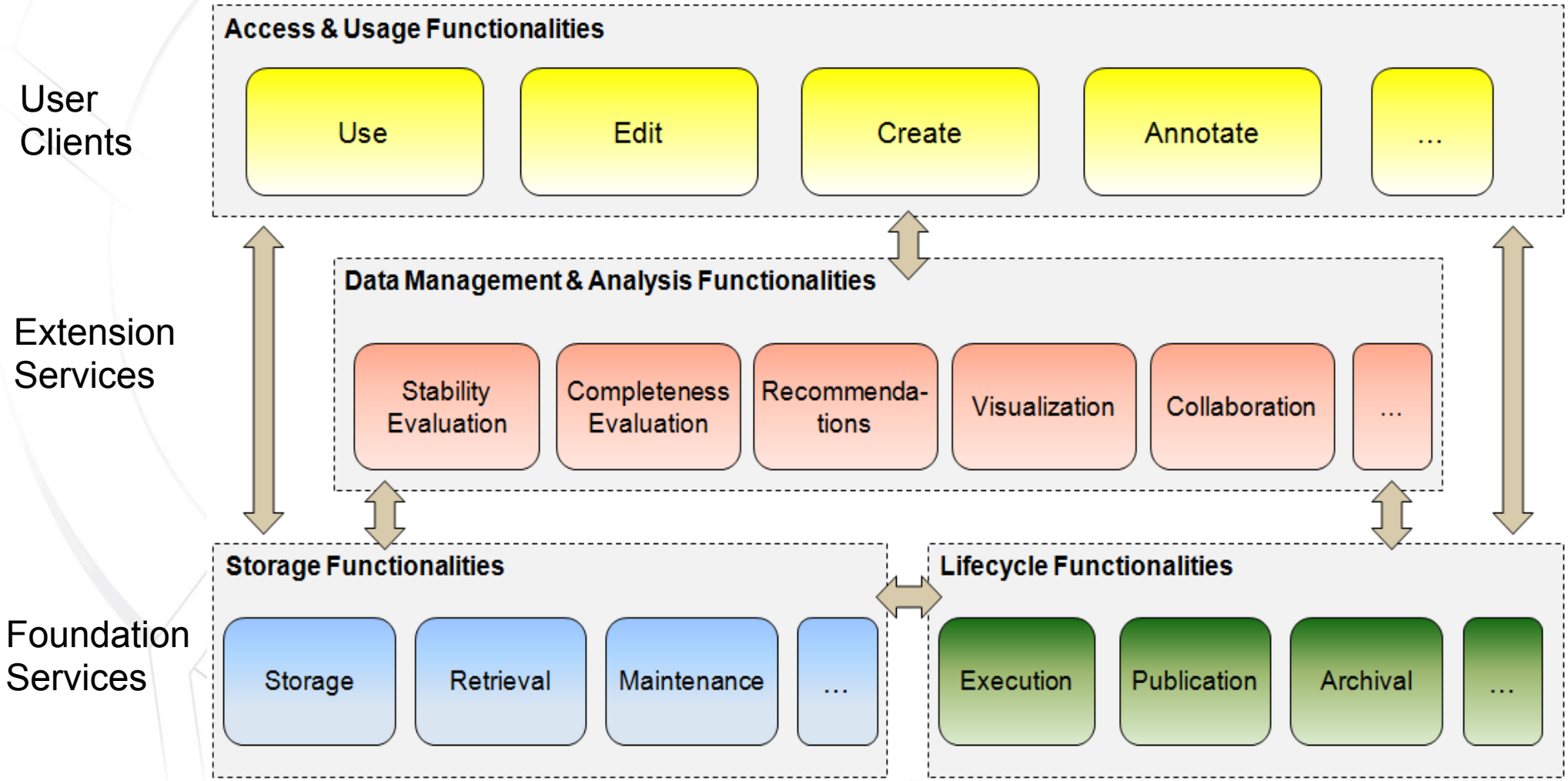
Credit and attribution

Papers with data links are cited more than those without

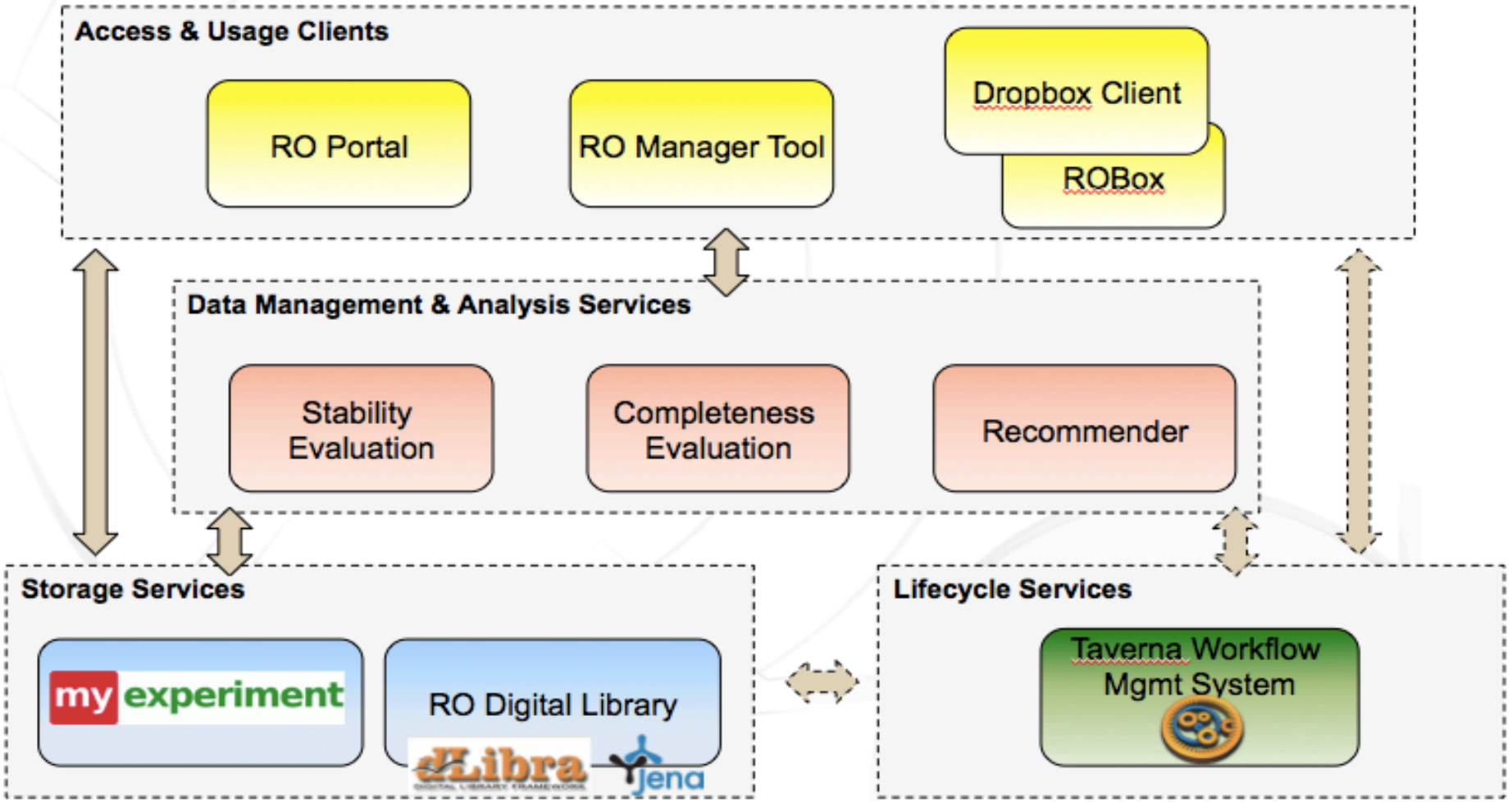


Effect of E-printing on Citation Rates in Astronomy and Physics 2006. Edwin A. Henneken et al.

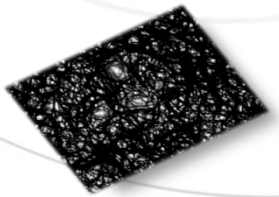
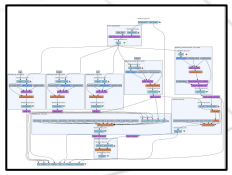
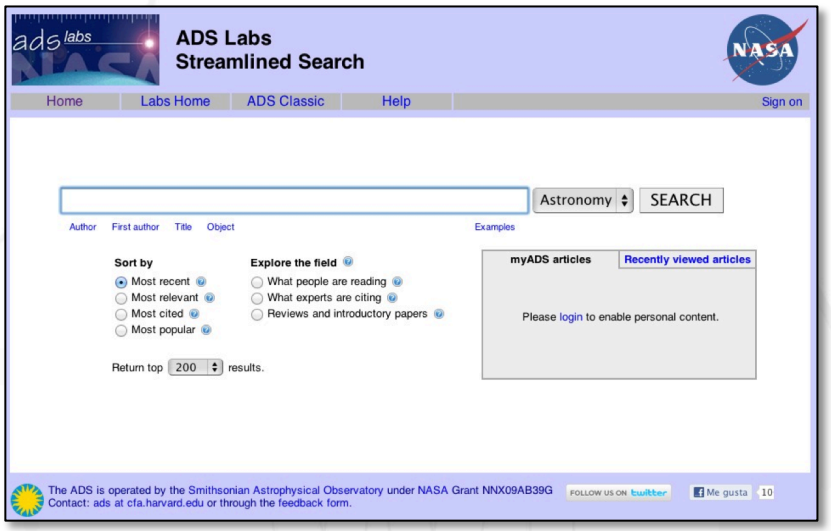
Research Object Digital Library Architecture



Research Object Digital Library Architecture



ADSLabs Research Objects

ADO Linked Components

- Authors
- Publications
- Journals
- Objects SIMBAD
- Tabular data behind the plots CDS
- ASCL reference of used software
- Observing time Proposals
- Used facilities, surveys or missions