Digital Science

Reproducibility and Visibility in Astronomy

José Enrique Ruiz on behalf of the Wf4Ever Team

SCIOPS 2013
ESAC, FRIDAY 13th SEPTEMBER 2013

Wf4Ever
Advanced Workflow Preservation Technologies for Enhanced Science
2011 - 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 (Poland)
5. University of Oxford and OeRC (OXF, UK)

6. Instituto Astrofísica Andalucía (IAA-CSIC, Sp
7. Leiden University Medical Centre (LUMC

Reproducible Science

The University of Manchester

MANCHESTER 1824

POZNAN SUPERCOMPUTING NETWORKING

INSTITUTO de ASTROFISICA de ANDALUCIA CSIC I·A·A

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 archives
- Publish process
  - Final data results 
  - Experiment in DL ADS/arXiv

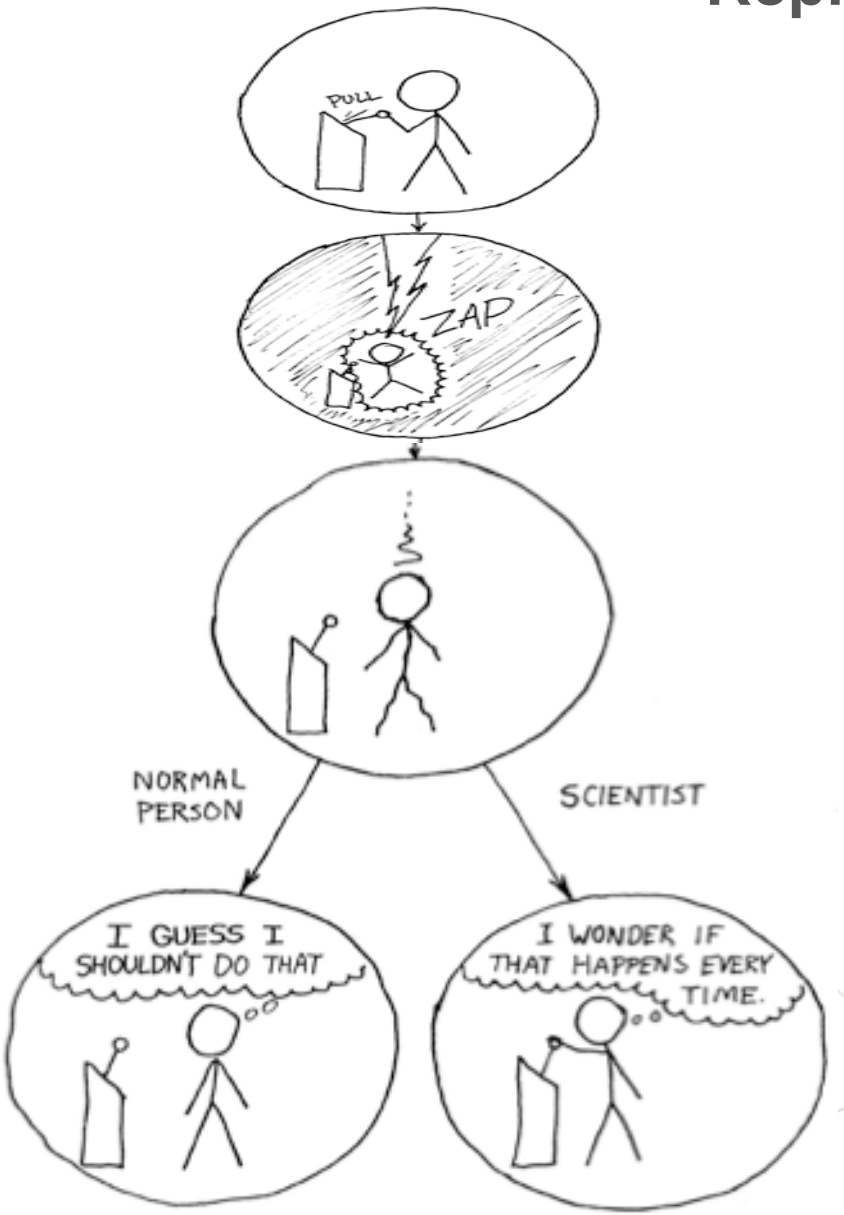
Reproducible research is still not possible in a digital world

A rich infrastructure of data is not efficiently used

A normalized preservation of methodology is needed

Tools

Reproducibility and The Scientific Method



http://xkcd.com/242/

Benefits

- Publishing knowledge, not advertising
- The author, the referee, the re-user
- Reputation, prestige and respect
- Higher quality of publications
  - Authors will be more careful
  - Many eyes to check results

Challenges

- Hard and time consuming
- Need incentives – not rewarded now

Barriers to Data and Code Sharing in Computational Science

Survey of Machi … len, 2010):

I don't know how

Code		Data
77%	Time to document and clean up	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%

Visibility, 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 contribs are encoded in software**

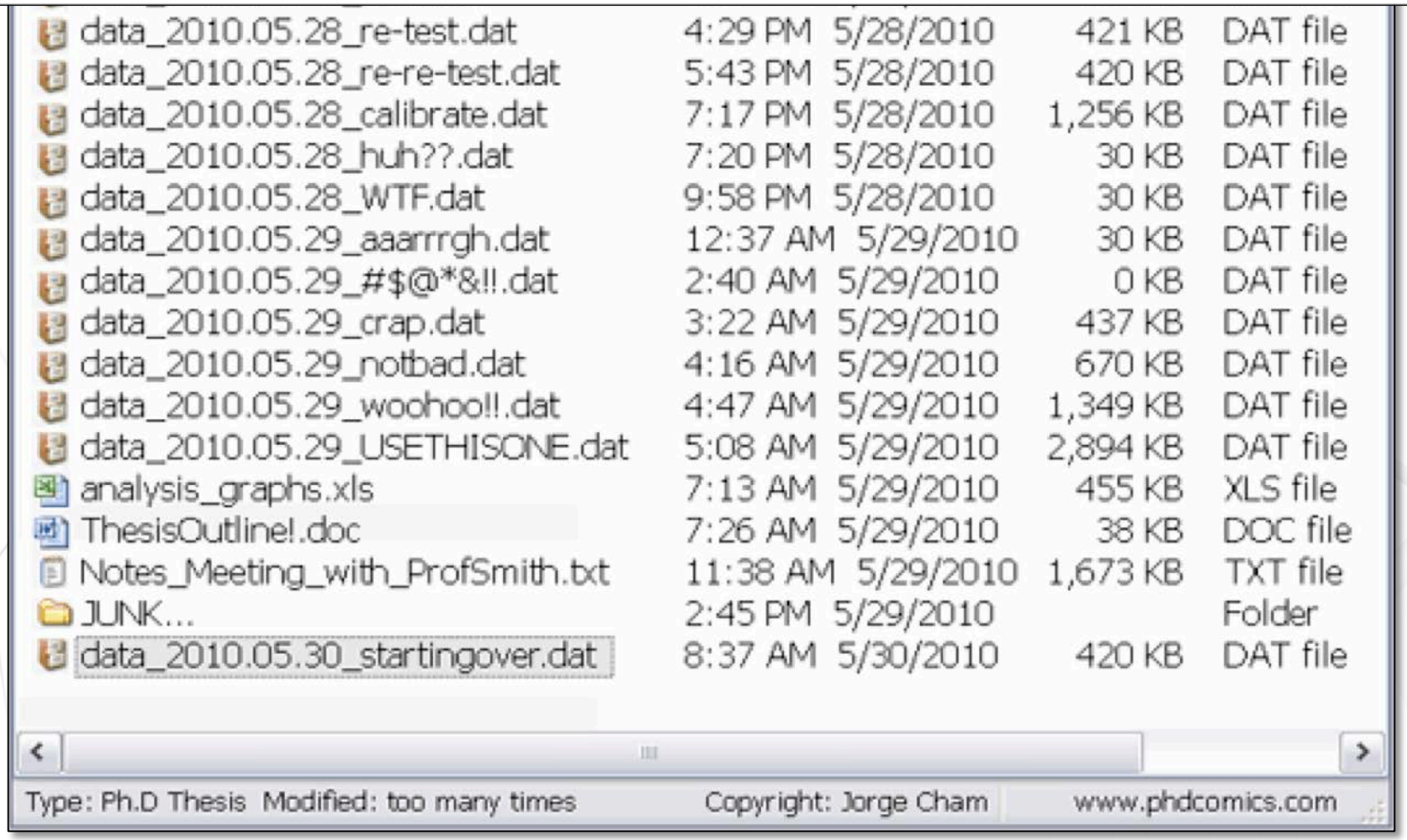
More data in archives does not imply more knowledge

- Expose **complete scientific record**, not the story
- Allow easy **discovery** of methods and tools

Paper discovery: the social dimension

Time has come to go beyond the PDF

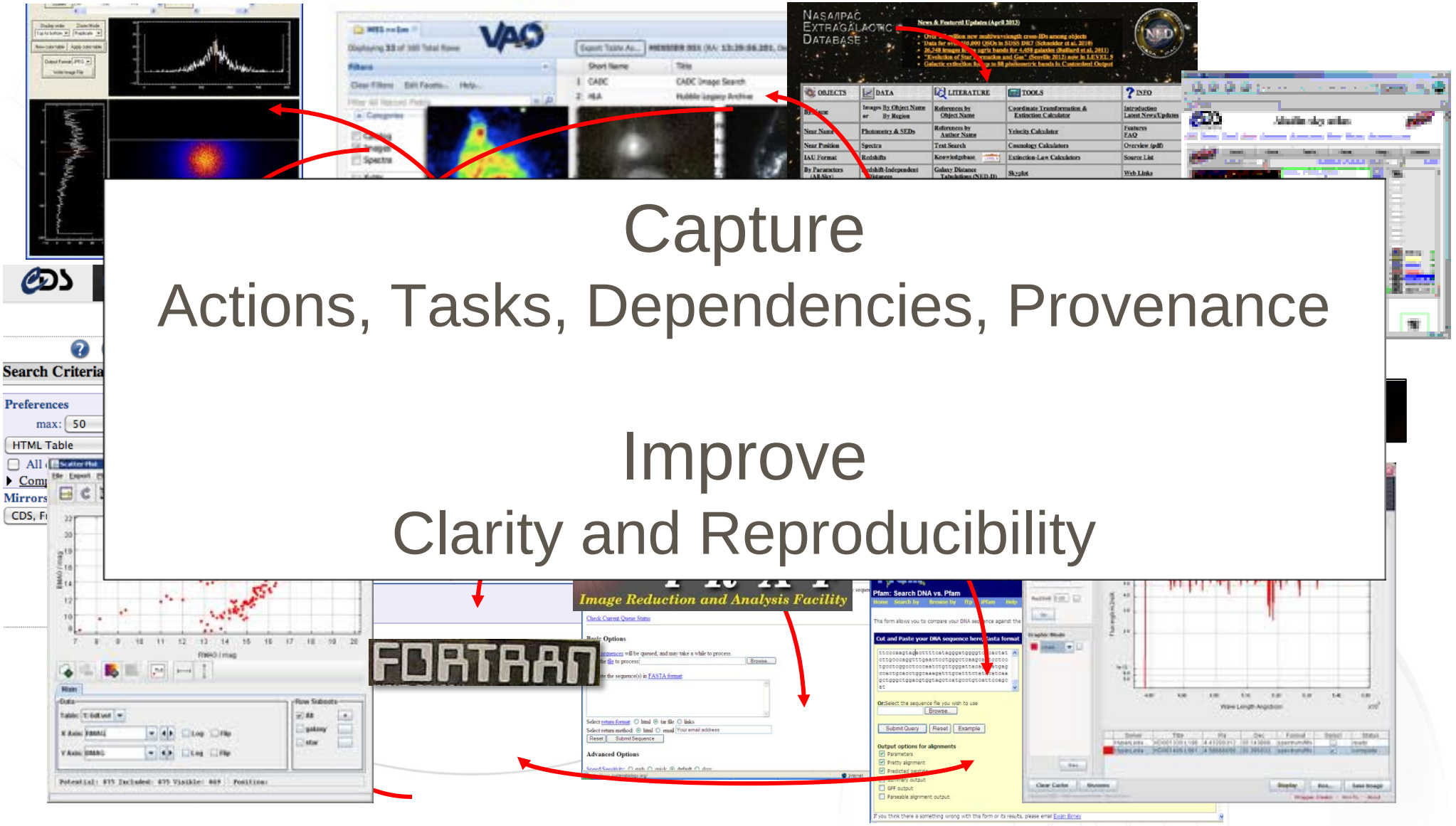
Going beyond automation
Organization



data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

Digital Astronomy in the Local Desktop



Scientific Workflows

Scientific Workflows

Related Initiatives

- ER-Flow
- VAMDC
- Helio-VO
- Cyber-SKA
- IceCore
- Montage
- **Astro-WISE**
- AstroGrid

Software

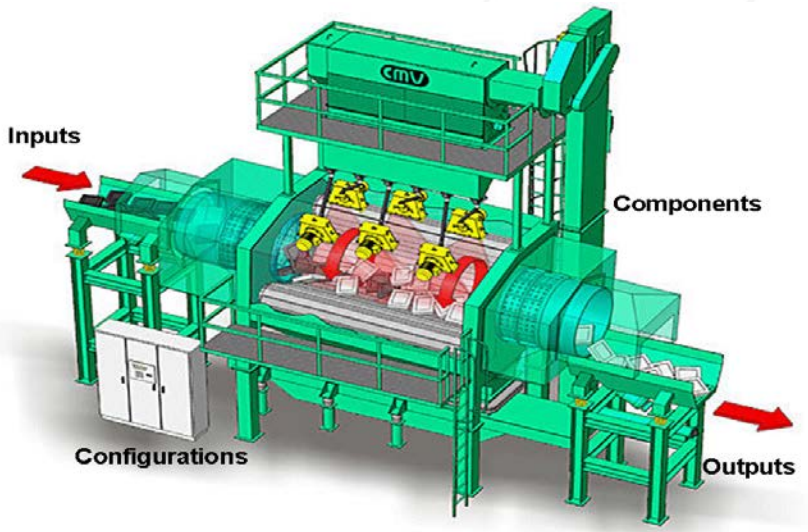
- Taverna
- Kepler
- Pegasus
- Triana
- **ESO Reflex**

IVOA

- AstroGrid
- Grid&WS WG
- VO France Wf WG

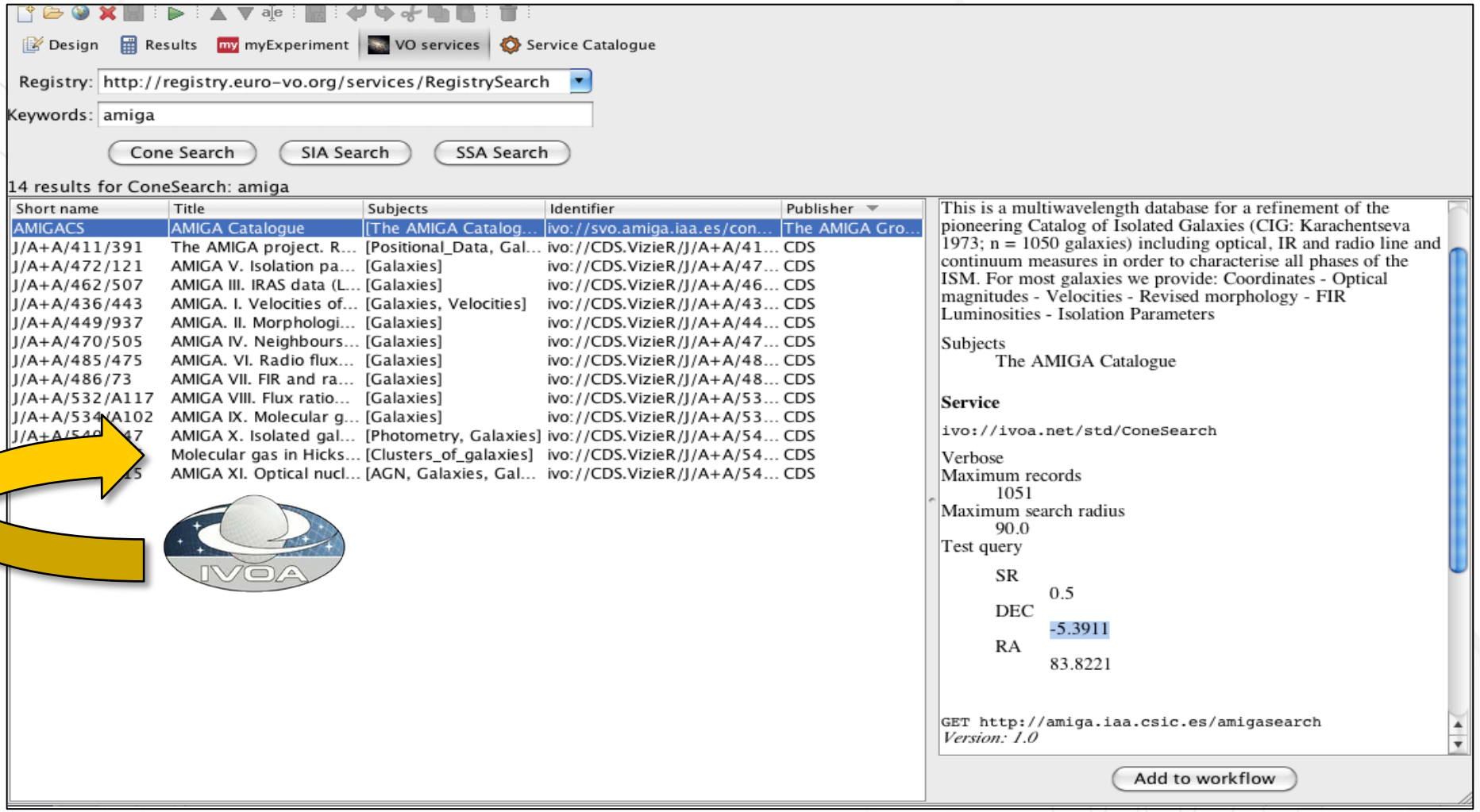
Self descriptive WS

- PDL
- SimDAL, S3



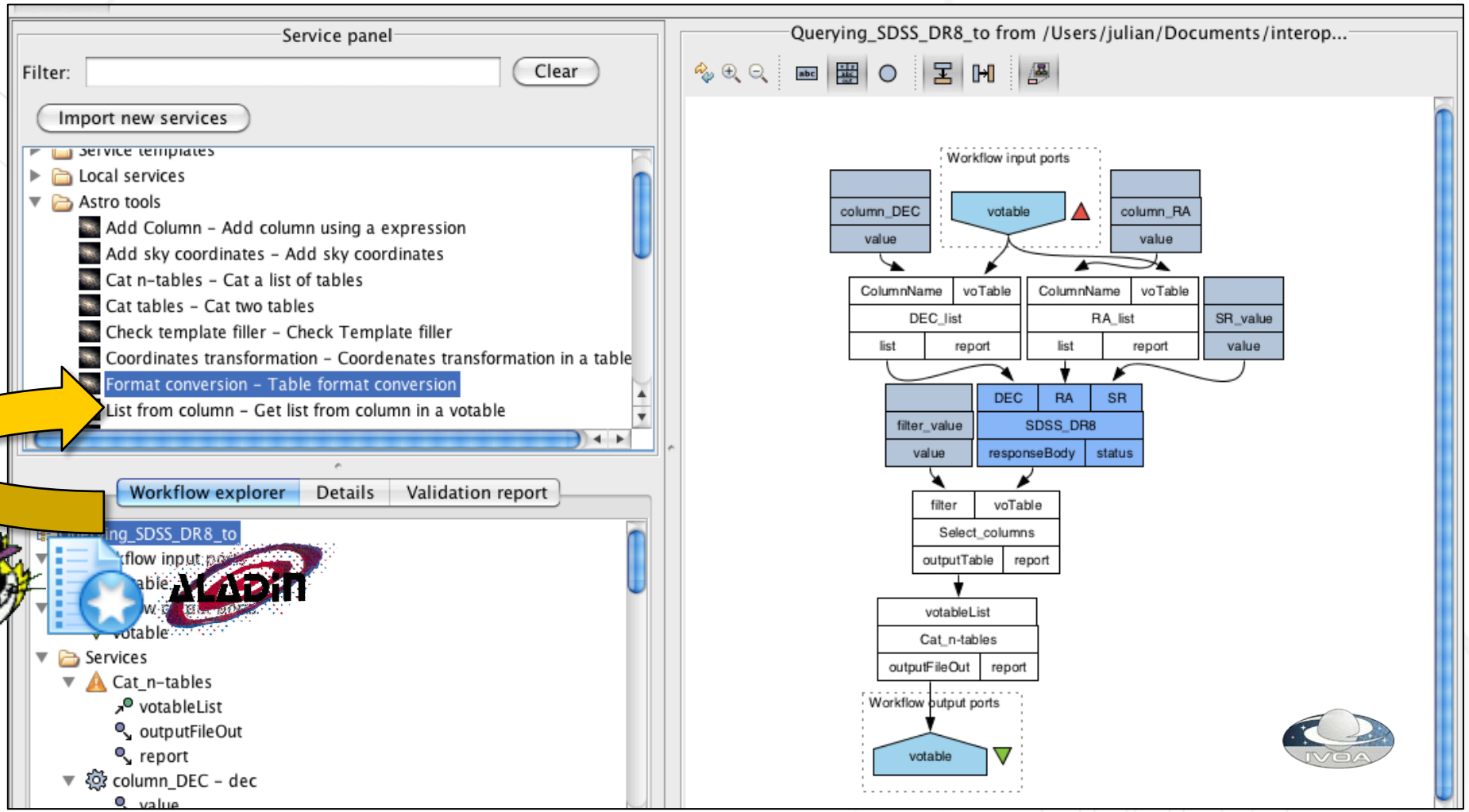
Interoperability y Standards

AstroTaverna: Create, annotate and run a workflow

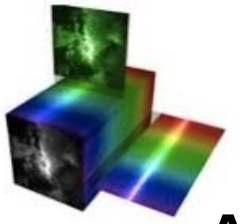


http://amiga.iaa.es/p/290-astrotaverna.htm

AstroTaverna: Create, annotate and run a workflow



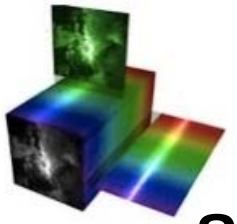
http://amiga.iaa.es/p/290-astrotaverna.htm



ASKAP Datacubes

	Low Res		High Res		Extreme Res	
Number	4 Bytes	4B	4 Bytes	4B	4 Bytes	4B
Resolution	2,048 x 2,048	16MB	8,192 x 8,192	268MB	12,288 x 12,288	603MB
Channels	16,384	0.27TB	16,384	4.39TB	16,384	9.8TB
Stokes & Weighting	1	0.27TB	1	4.39TB	4 + 1	49.5TB

Prof. Kevin Vinsen



SKA Datacubes

Spectral Line Datacube

- Dish
  - Assume 30,000 channels
  - 27,000 x 27,000 x 30,000 x 4
  - ≈80TB
- AA
  - Assume 40,000 channels
  - 28,000 x 28,000 x 40,000 x 4
  - ≈125TB
- Stokes parameters and Weighting Map
  - Multiple by 5
  - Dish ≈ 400TB
  - AA ≈ 625TB

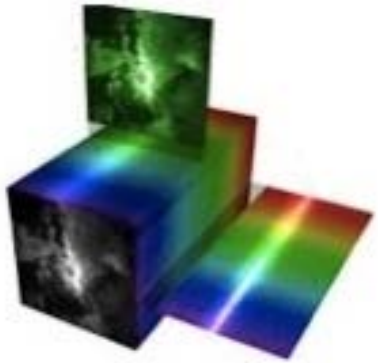
The next generation of archives

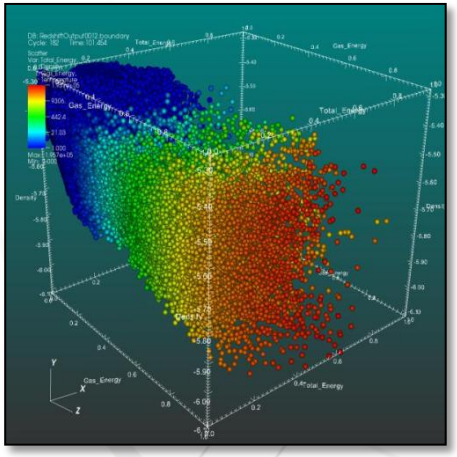
Much wider FoV and spectral coverage

» Large volumes for a single observed dataset

Automated surveys

» Huge amounts of tabular data





Extraction of scientifically relevant info from a multiD param. space

» Exploration services
» Anomaly detection
» Cross-matching data
» Dimensionality reduction

Detailed inspection and subset

» Filtering
» Extraction
» Re-Projection
» Analysis services

We are moving into a world where

» computing and storage are cheap
» **data movement is death**

The *move computing to data* parad

» A cloud of Web Services

Archives should evolve from … into
» Virtual Data provider…
» Software Tasks p…

» Archives speaking W…

Astronomy …s/facilities/wavelength
Interco… …roperable archives
» …bservatory
» …sks

Web Services based Scientific Workflows

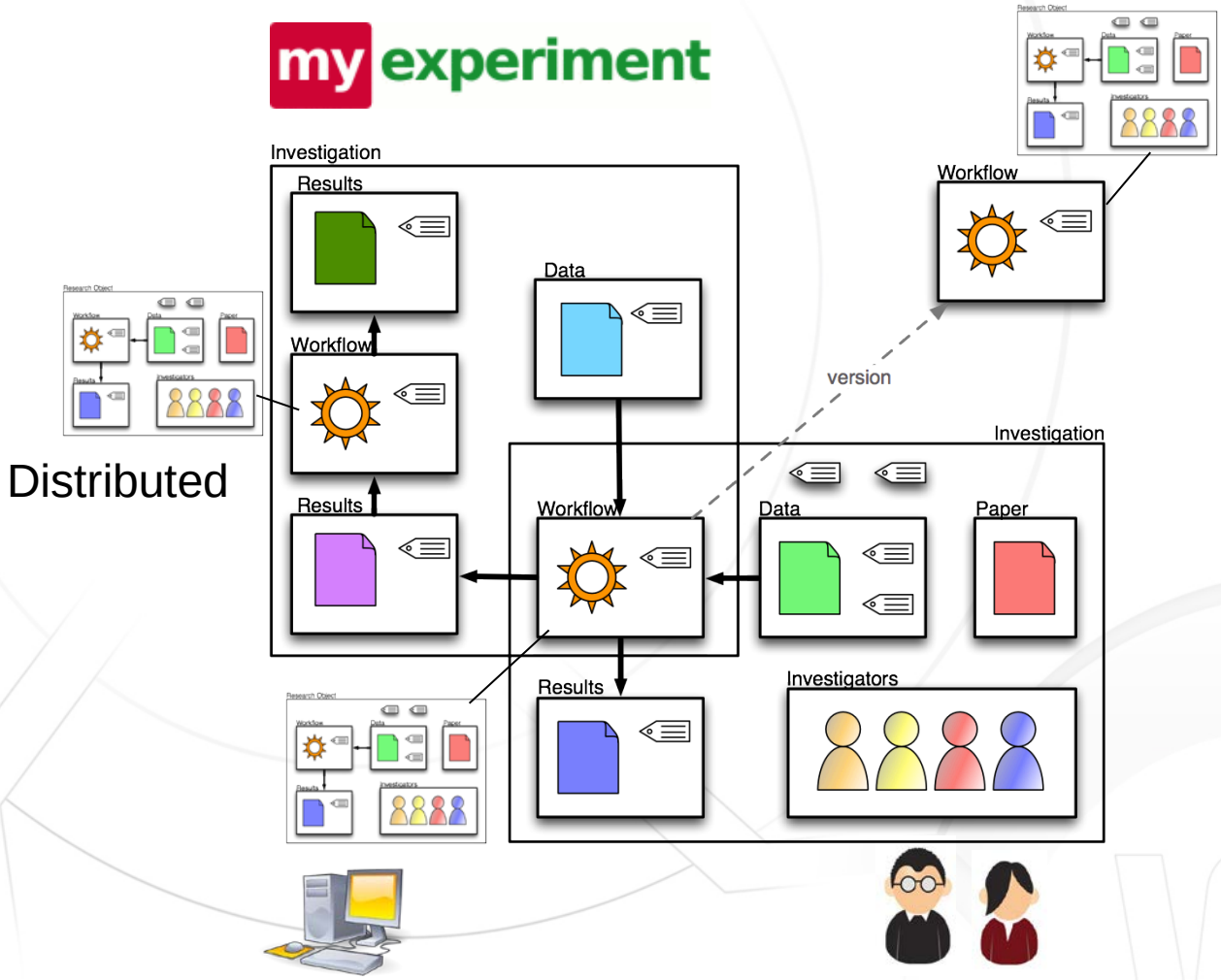
Preservation

Process should benefit of the same privileges acquired by data

Preserving the method ensures replication of final results at any moment

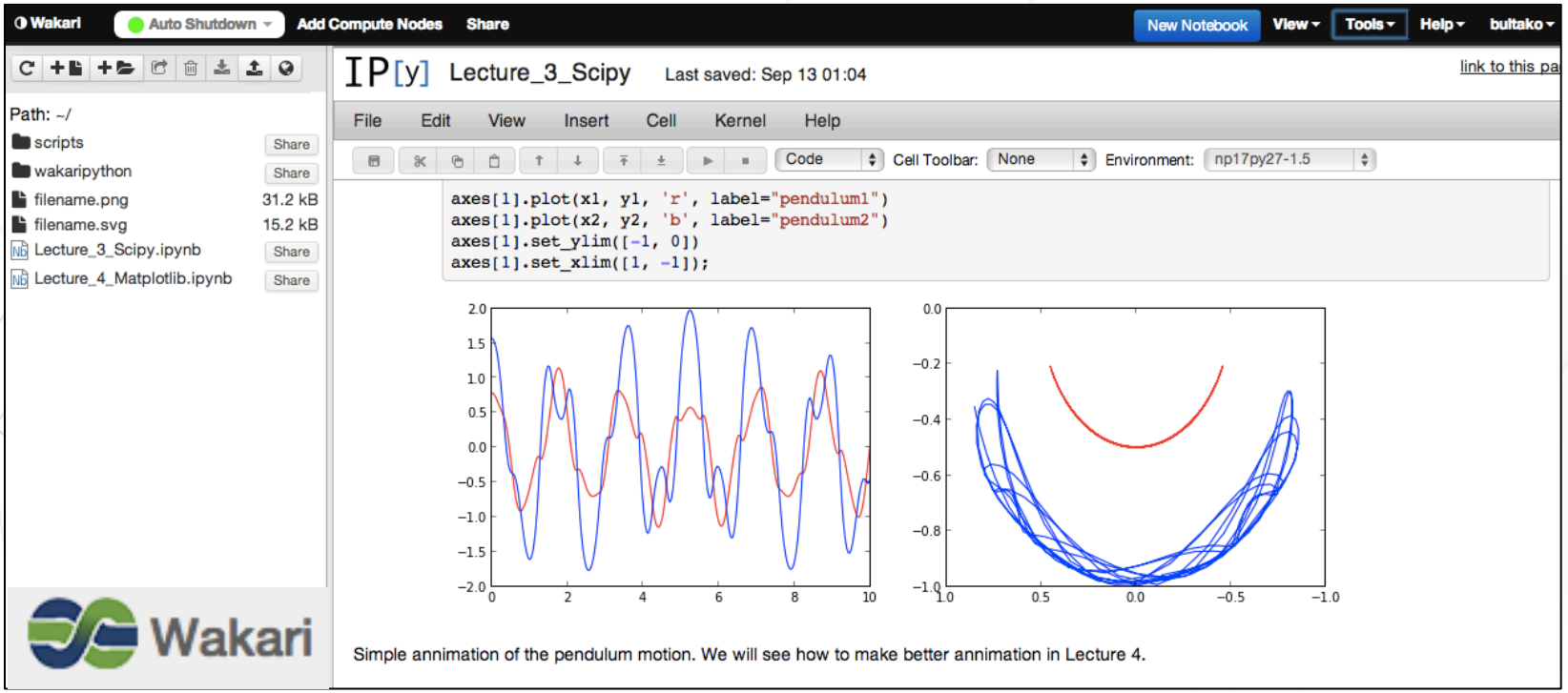
Research Objects

Expose experimental context in a structured way in order to be understood



IPython Notebook solutions

- Web-browser as the working desktop
- Python code, plots and data, living with rich-text documentation
- Cloud-based adaptive scalable computing environment
- Fully shareable, re-usable and executable wikis
- Social platform and Git versioning



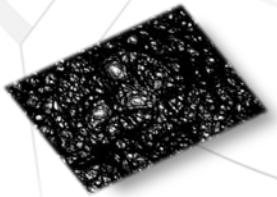
Research Objects

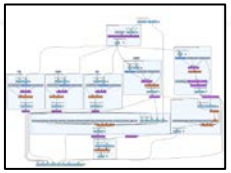
Similar Initiative to ESO Telbib

ADSLabs

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



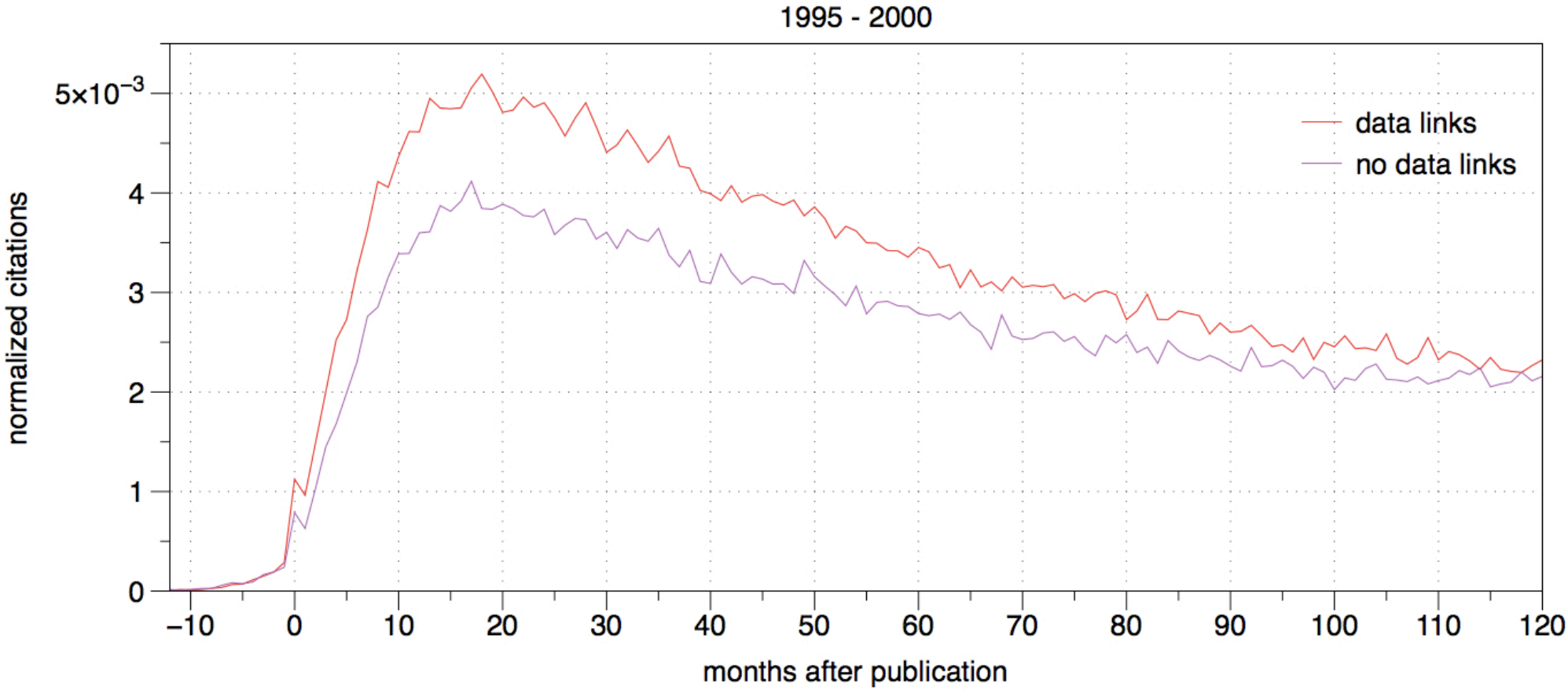


Incentives

http://labs.adsabs.harvard.edu/

The Incentive

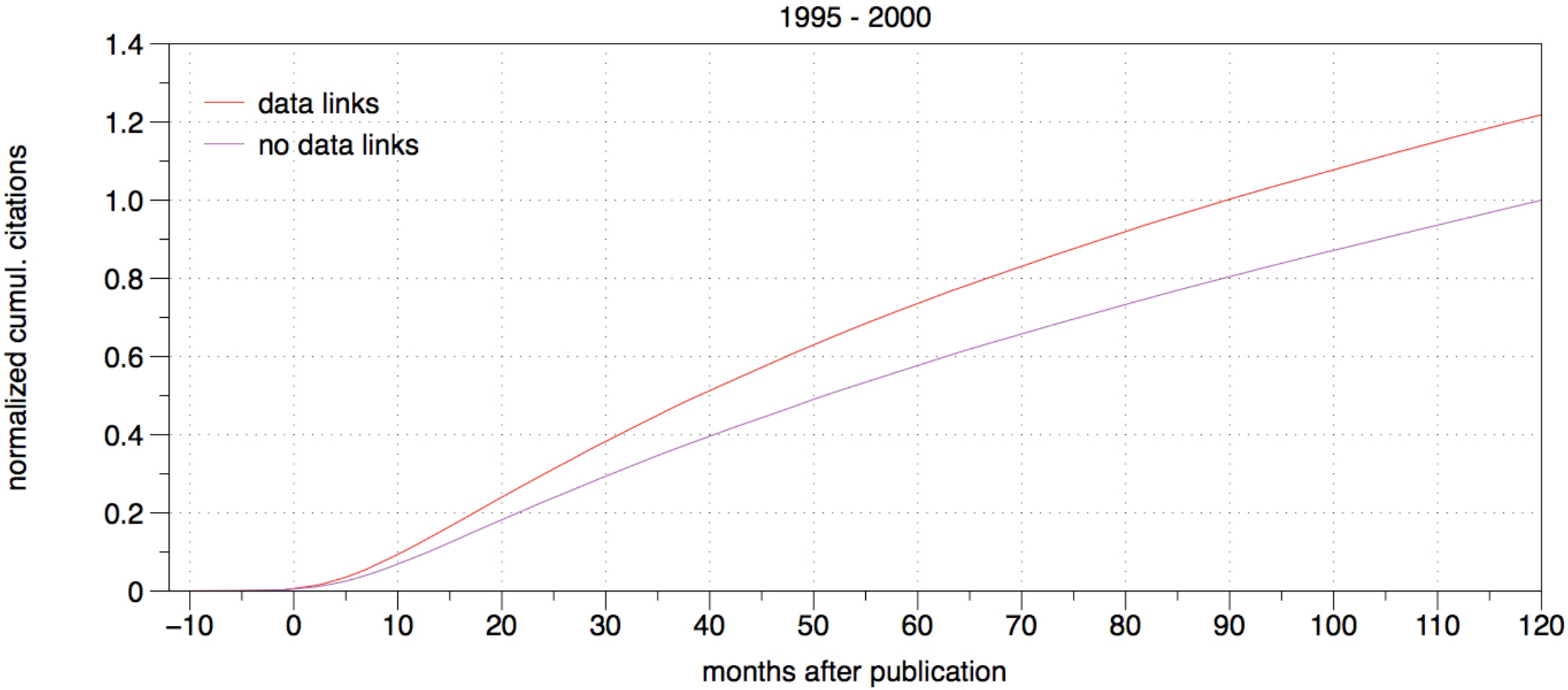
Papers with data links are cited more than those without



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

The Incentive

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.

Conclusions

- Reproducibility is at the very heart of the scientific method
- Improving visibility is key in order to avoid reinvention
- Social dimension of science stressed in the discovery process
- Highly specialized science needs re-use to achieve efficiency
- In a digital world, publish decomposable executable papers
- Capture provenance and structure in the local desktop
- Scientific workflows go beyond automation: provide clarity and structure
- Transfer rate is more than an issue for next generation of archives
- The move computing to data paradigm -> back to old terminals
- Process should benefit of the same benefits acquired by data
- Digital libraries of web-services-based workflows
- The distributed digital workflow-centric Research Object
- Preserving knowledge - not only data or advertising

jer@iaa.es