

# 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, Spain)
7. Leiden University Medical Centre (LUMC, Netherlands)



**Reproducible  
Science**



# Digital Science - Reproducibility and Visibility in Astronomy

## 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**

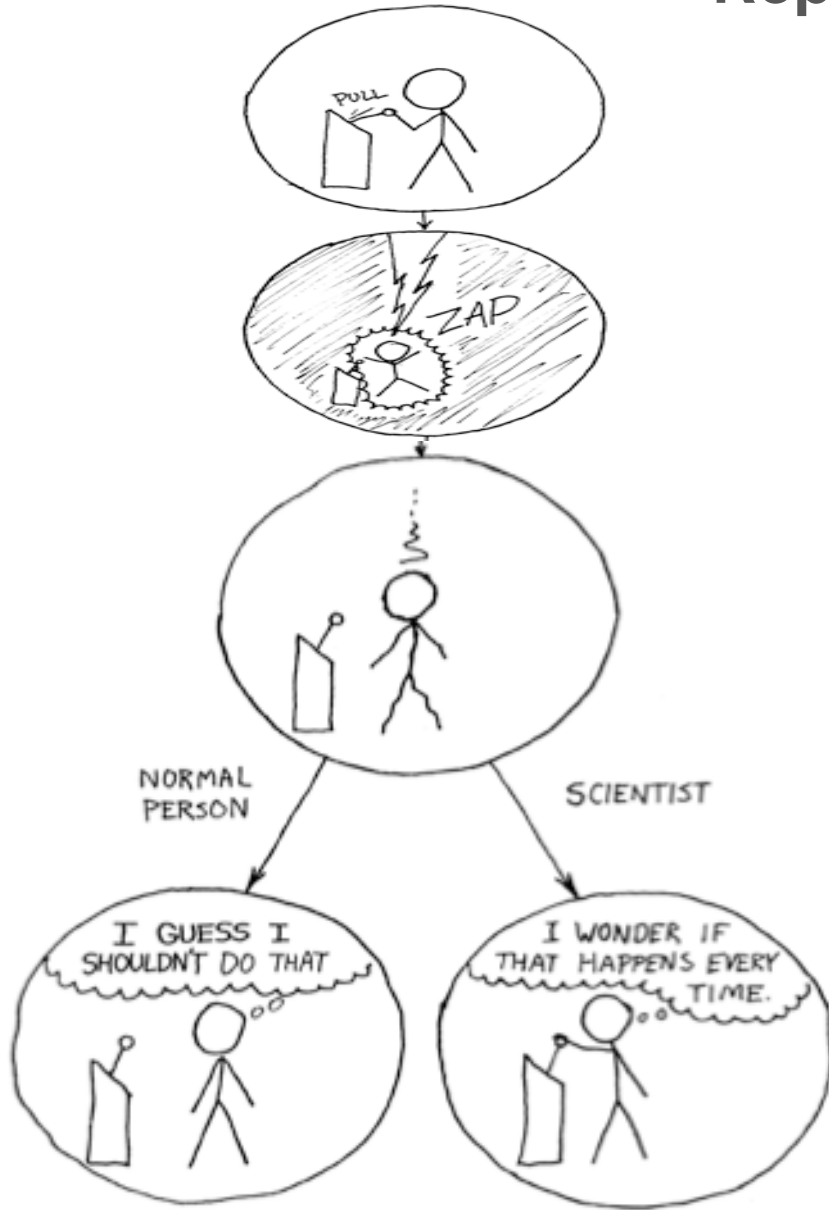


**A normalized preservation of methodology is needed**

**Tools**

# Digital Science - Reproducibility and Visibility in Astronomy

## Reproducibility and The Scientific Method



### 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

**I don't know how**

(en, 2010):

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%

# Digital Science - Reproducibility and Visibility in Astronomy

## 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



# Digital Science - Reproducibility and Visibility in Astronomy

## Visibility and Social Discovery

### Paper discovery: the social dimension

peer evaluation  
empowering scholars

MENDELEY

YouTube

citeulike

Search citeulike

[Browse](#) | [FAQ](#) | [Log in](#)

Install the Web Importer  
The Web Importer lets you import references and documents from over 30 academic databases with a single click. You can add it to your browser here.

Edit My Profile  
Fill out your research profile to increase your impact in the Mendeley network and to enable

citeulike is a free service for managing

BibSonomy

ResearchGate

Search

twitter

klænk  
Spread your research results

Collabgraph!

delicious  
social bookmarking

AstroBetter

Tips and Tricks for Professional Astronomers

Blog About Archives Support Wiki

Learning Python - The Interactive Way  
by Jessica on June 18, 2012

One of the nice things about programming in python is that it is free, relatively easy to use, and there is lots of support and development online. One of the downsides is that there is not a standard python package to install... you have to know about all the interesting add-on bits to get maximum usage out of python. Luckily, some enterprising astronomers and other programming-scientists have put together pre-packaged python distributions containing many of the most useful add-ons like numpy, scipy, matplotlib, atpy, etc.

In this post, I wanted to introduce you to some other interesting features and packages that may be brand-new or not yet widely advertised. I'll also include some new book/sites for learning python.

Learning Python

Search  
To search, type and hit enter

Follow AstroBetter

- Subscribe
- Subscribe by Email
- Facebook
- Follow on Twitter

Contributors

collaborating in your field of research. Just **upload** or upload a bibtex file, containing your **graph** will create a fancy graph showing

slideshare  
BETA

zotero

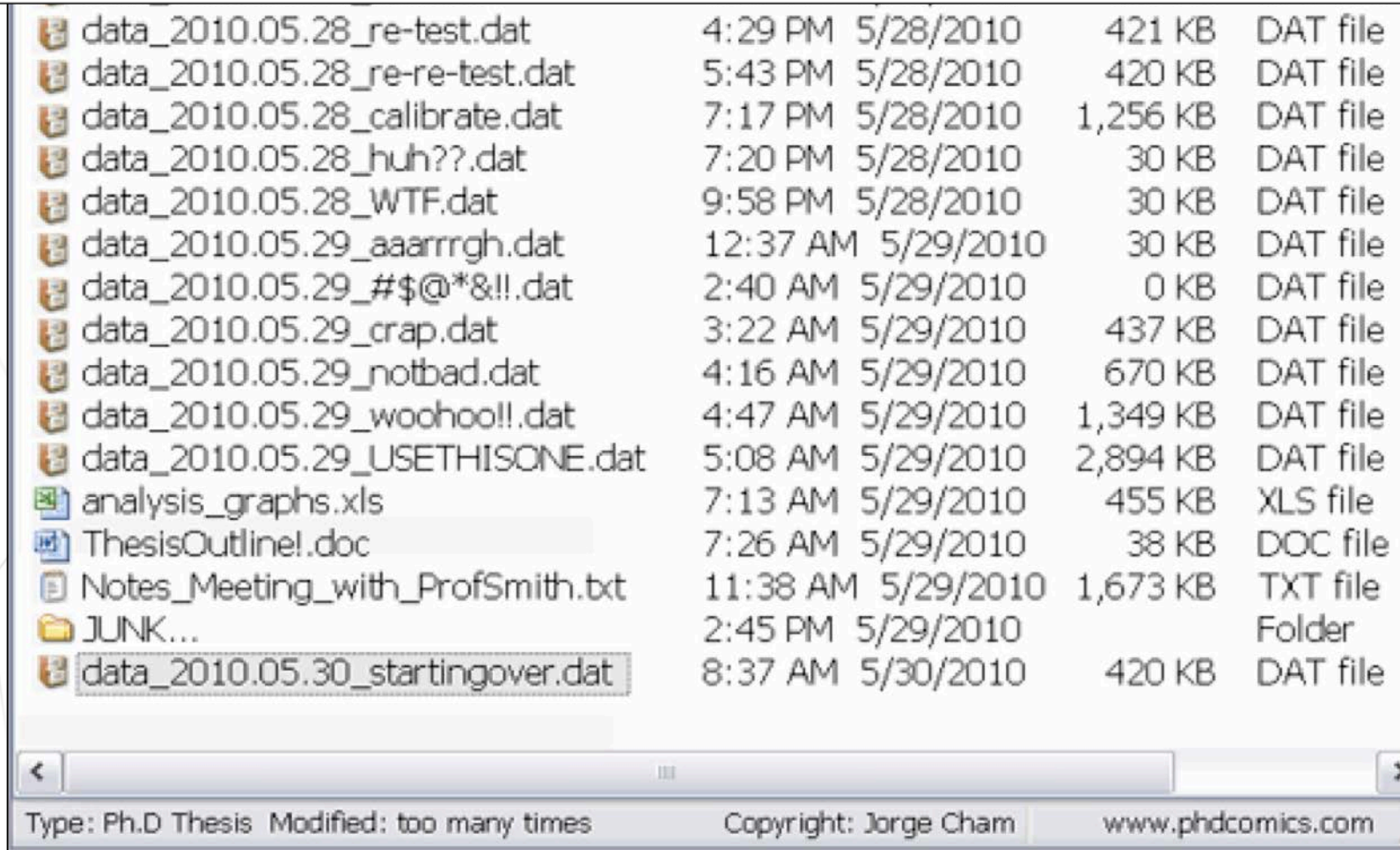


Time has come to go **beyond the PDF**





# 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_aaarrgh.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

# Workflows to Access and Massage VO Data Digital Astronomy in the Local Desktop

Capture  
Actions, Tasks, Dependencies, Provenance

Improve  
Clarity and Reproducibility

The collage features several key astronomical software interfaces:

- VAO (Virtual Astronomy Observing):** A central interface for data discovery and management, showing search filters and object lists.
- NASAMPAC EXTRAGALACTIC DATABASE:** A specialized database for extragalactic objects, displaying a table of objects and their associated data.
- CDS (Centre for Data Services):** A platform for data access and visualization, showing search criteria and preferences.
- Image Reduction and Analysis Facility (IRAF):** A suite of tools for image reduction and analysis, including a 'FORTRAN' logo and various processing options.
- Sequence Analysis Tools:** A tool for comparing DNA sequences, showing a 'Cut and Paste your DNA sequence here' field and a 'Submit Query' button.

Red arrows trace a workflow path: from the VAO search results, through the NASAMPAC database, to the CDS search criteria, and finally to the IRAF image reduction and analysis tools.

# Digital Science - Reproducibility and Visibility in Astronomy

## Scientific Workflows

---



## Related Initiatives

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

## Software

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

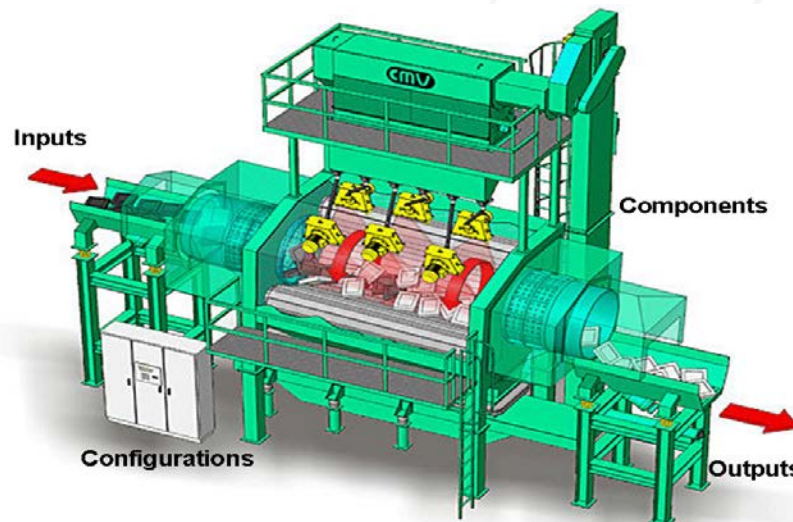
## IVOA



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

## Self descriptive WS

- › **PDL**
- › SimDAL, S3



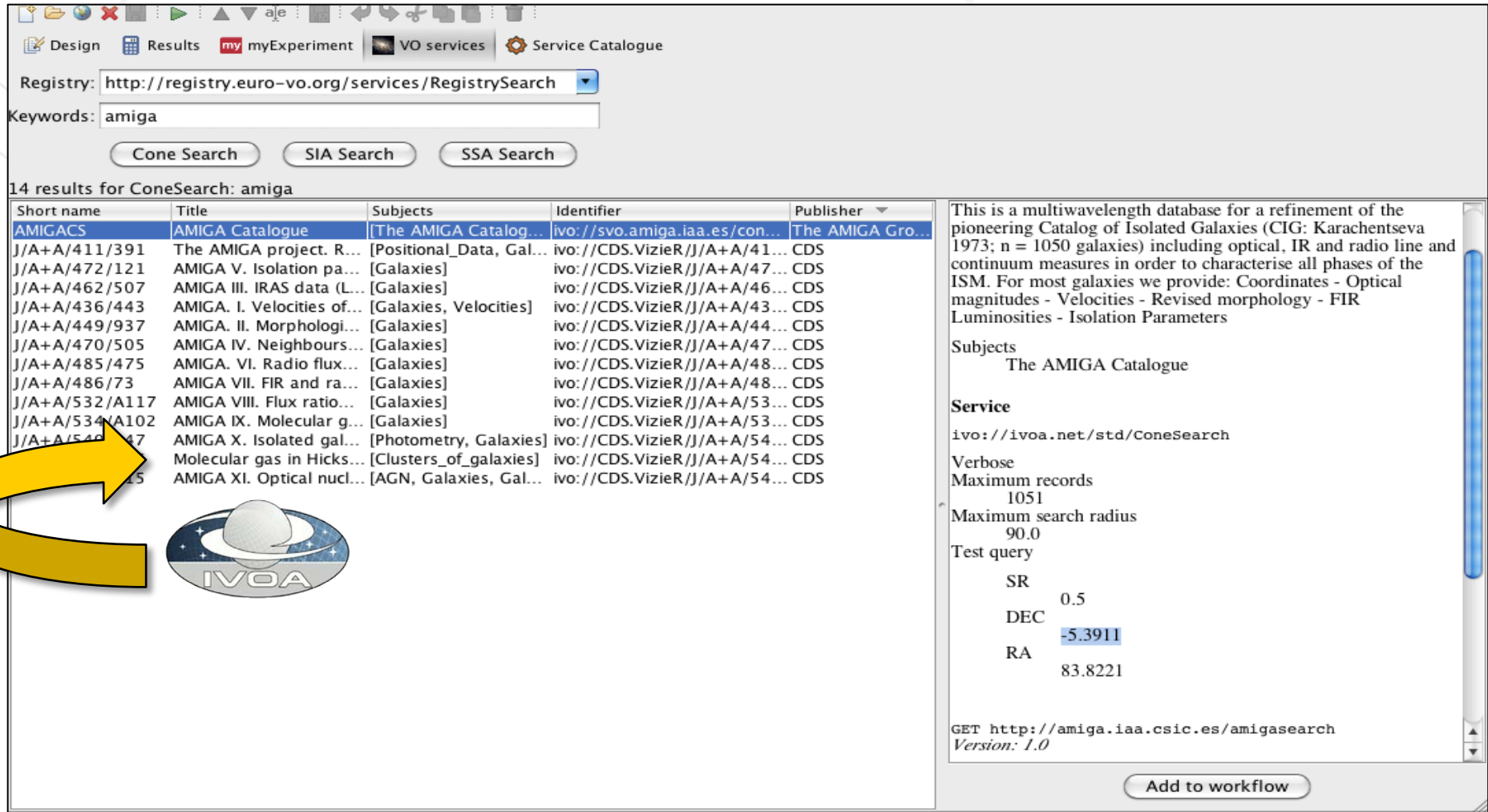
Interoperability  
Standards



# Digital Science - Reproducibility and Visibility in Astronomy

## Astronomical Research Objects in Action

### AstroTaverna: Create, annotate and run a workflow



The screenshot displays the AstroTaverna web interface. At the top, there are tabs for 'Design', 'Results', 'my myExperiment', 'VO services', and 'Service Catalogue'. Below these, the 'Registry' is set to 'http://registry.euro-vo.org/services/RegistrySearch' and the search 'Keywords' are 'amiga'. There are buttons for 'Cone Search', 'SIA Search', and 'SSA Search'. The main area shows '14 results for ConeSearch: amiga' in a table with columns for Short name, Title, Subjects, Identifier, and Publisher. A yellow arrow points from the table to the IVOA logo. On the right, a detailed view of the 'AMIGACS' service is shown, including its description, subjects, and service parameters.

Short name	Title	Subjects	Identifier	Publisher
AMIGACS	AMIGA Catalogue	[The AMIGA Catalog...	ivo://svo.amiga.iaa.es/con...	The AMIGA Gro...
J/A+A/411/391	The AMIGA project. R...	[Positional_Data, Gal...	ivo://CDS.VizieR/J/A+A/41...	CDS
J/A+A/472/121	AMIGA V. Isolation pa...	[Galaxies]	ivo://CDS.VizieR/J/A+A/47...	CDS
J/A+A/462/507	AMIGA III. IRAS data (L...	[Galaxies]	ivo://CDS.VizieR/J/A+A/46...	CDS
J/A+A/436/443	AMIGA. I. Velocities of...	[Galaxies, Velocities]	ivo://CDS.VizieR/J/A+A/43...	CDS
J/A+A/449/937	AMIGA. II. Morphologi...	[Galaxies]	ivo://CDS.VizieR/J/A+A/44...	CDS
J/A+A/470/505	AMIGA IV. Neighbours...	[Galaxies]	ivo://CDS.VizieR/J/A+A/47...	CDS
J/A+A/485/475	AMIGA. VI. Radio flux...	[Galaxies]	ivo://CDS.VizieR/J/A+A/48...	CDS
J/A+A/486/73	AMIGA VII. FIR and ra...	[Galaxies]	ivo://CDS.VizieR/J/A+A/48...	CDS
J/A+A/532/A117	AMIGA VIII. Flux ratio...	[Galaxies]	ivo://CDS.VizieR/J/A+A/53...	CDS
J/A+A/534/A102	AMIGA IX. Molecular g...	[Galaxies]	ivo://CDS.VizieR/J/A+A/53...	CDS
J/A+A/540/47	AMIGA X. Isolated gal...	[Photometry, Galaxies]	ivo://CDS.VizieR/J/A+A/54...	CDS
J/A+A/540/45	Molecular gas in Hicks...	[Clusters_of_galaxies]	ivo://CDS.VizieR/J/A+A/54...	CDS
J/A+A/540/45	AMIGA XI. Optical nucl...	[AGN, Galaxies, Gal...	ivo://CDS.VizieR/J/A+A/54...	CDS

**AMIGACS**  
This is a multiwavelength database for a refinement of the pioneering Catalog of Isolated Galaxies (CIG: Karachentseva 1973; n = 1050 galaxies) including optical, IR and radio line and continuum measures in order to characterise all phases of the ISM. For most galaxies we provide: Coordinates - Optical magnitudes - Velocities - Revised morphology - FIR Luminosities - Isolation Parameters

**Subjects**  
The AMIGA Catalogue

**Service**  
ivo://ivoa.net/std/ConeSearch

**Verbose**  
Maximum records  
1051  
Maximum search radius  
90.0  
Test query  
SR 0.5  
DEC -5.3911  
RA 83.8221

GET http://amiga.iaa.csic.es/amigasearch  
Version: 1.0

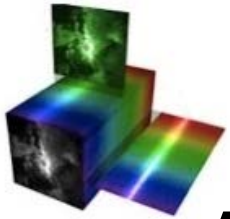
Add to workflow

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

### AstroTaverna: Create, annotate and run a workflow

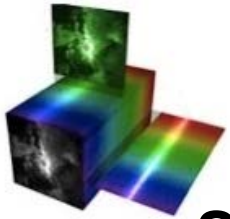
The screenshot displays the AstroTaverna interface. On the left, the 'Service panel' lists various services under 'Astro tools', with 'Format conversion - Table format conversion' highlighted. A yellow arrow points from this service to the workflow diagram on the right. The workflow diagram, titled 'Querying\_SDSS\_DR8\_to from /Users/julian/Documents/interop...', shows a sequence of steps: 'Workflow input ports' (column\_DEC, value, column\_RA, value) feed into 'Select\_columns' (filter\_value, value, filter, voTable), which outputs 'outputTable' and 'report'. This leads to 'Cat\_n-tables' (votableList, outputFileOut, report), which finally outputs to 'Workflow output ports' (votable). The IVOA logo is visible in the bottom right corner of the interface.

<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



## SKA Datacubes



### Spectral Line Datacube

- Dish
  - Assume 30,000 channels
  - $27,000 \times 27,000 \times 30,000 \times 4$
  - $\approx 80\text{TB}$
- AA
  - Assume 40,000 channels
  - $28,000 \times 28,000 \times 40,000 \times 4$
  - $\approx 125\text{TB}$
- Stokes parameters and Weighting Map
  - Multiple by 5
  - Dish  $\approx 400\text{TB}$
  - AA  $\approx 625\text{TB}$

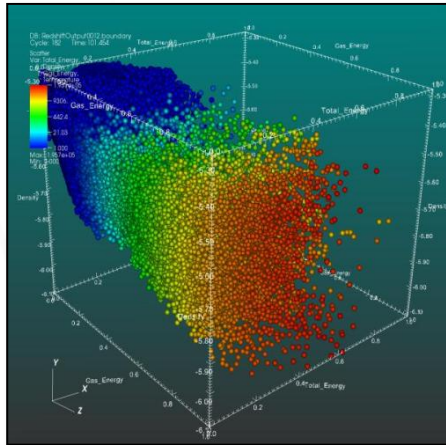
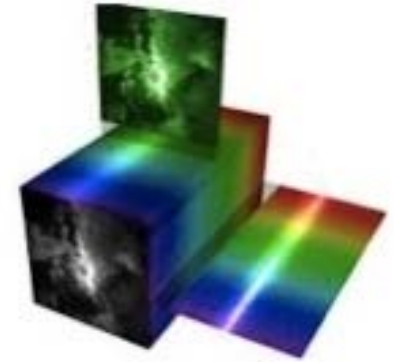


### Much wider FoV and spectral coverage

- » 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* paradigm

- » A cloud of Web Services

Archives should evolve from *static* to *dynamic*

- » Virtual Data providers
- » Software Tasks providers

- » Archives speaking Web Services

Astronomy *data* / *facilities* / *wavelength*

Interconnected *interoperable* archives

- » *observatory*
- » *tasks*

Web Services based Scientific Workflows



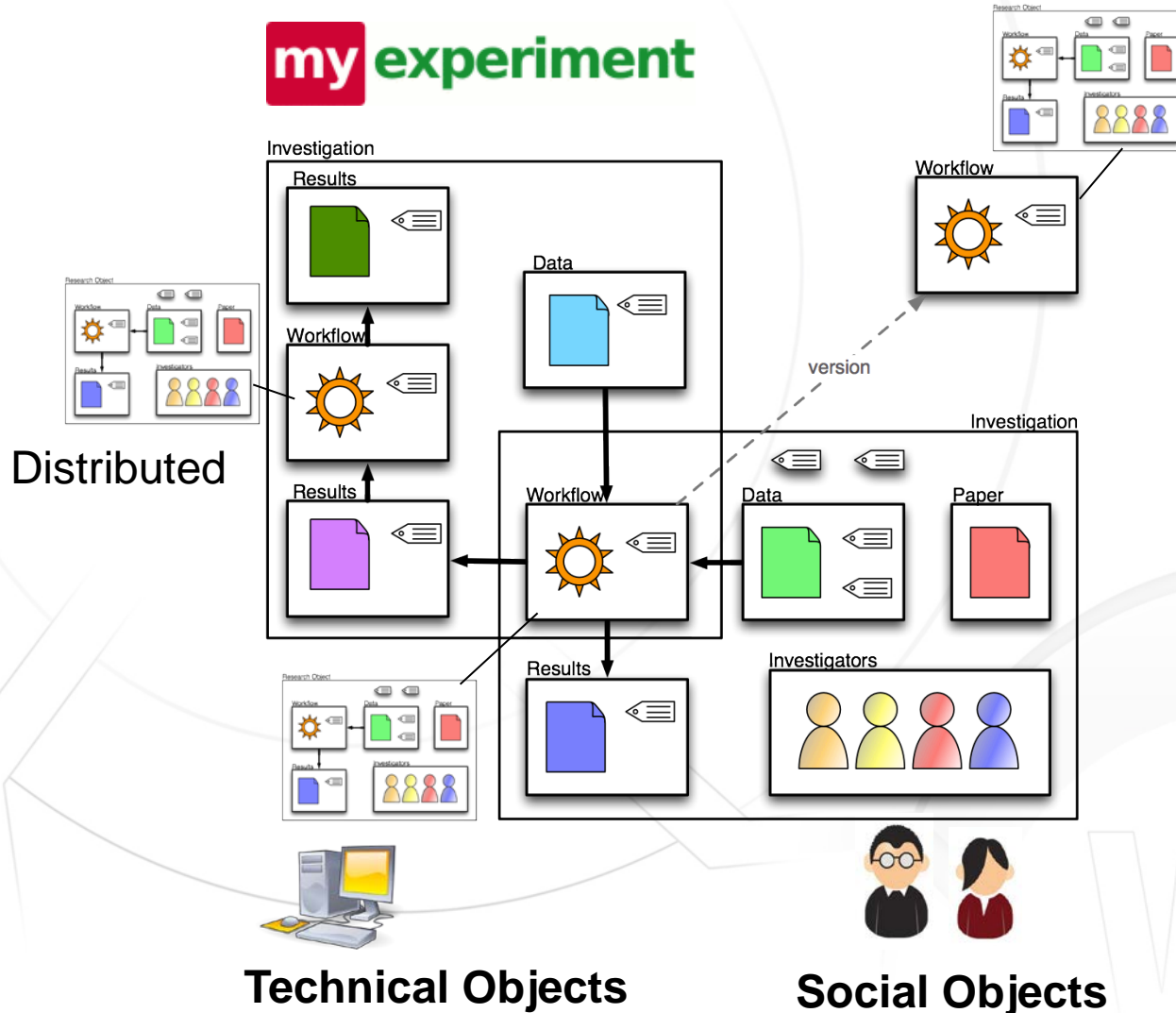
Process should benefit of the same privileges acquired by data

Preserving the method ensures replication of final results at any moment

# Digital Science - Reproducibility and Visibility in Astronomy

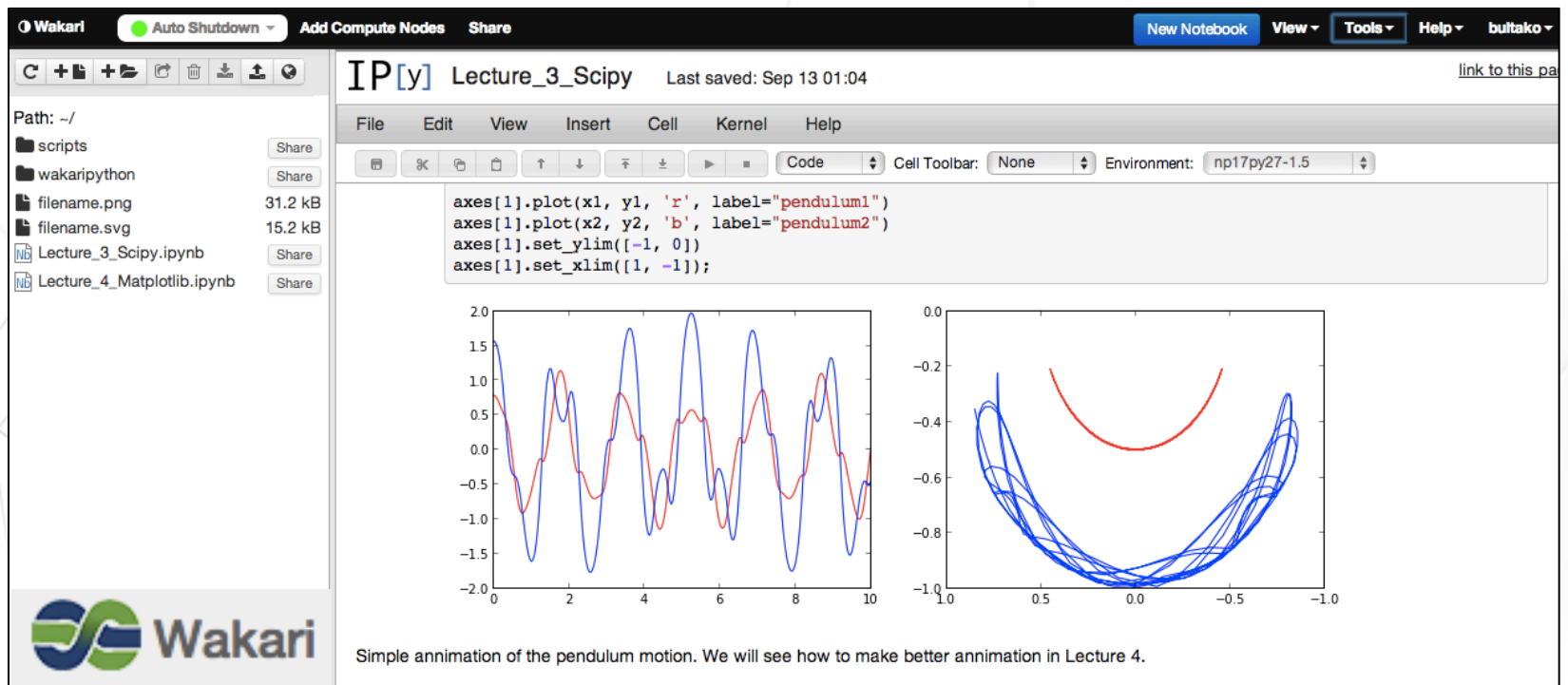
## Research Objects

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**



The screenshot displays the Wakari IPython Notebook interface. The top navigation bar includes 'Wakari', 'Auto Shutdown', 'Add Compute Nodes', 'Share', 'New Notebook', 'View', 'Tools', 'Help', and 'builtako'. The main window title is 'IP[y] Lecture\_3\_Scipy' with a timestamp 'Last saved: Sep 13 01:04'. The left sidebar shows a file explorer with a path of '~/' containing folders 'scripts' and 'wakaripython', and files 'filename.png' (31.2 kB), 'filename.svg' (15.2 kB), 'Lecture\_3\_Scipy.ipynb', and 'Lecture\_4\_Matplotlib.ipynb'. The central code cell contains the following Python code:

```
axes[1].plot(x1, y1, 'r', label="pendulum1")
axes[1].plot(x2, y2, 'b', label="pendulum2")
axes[1].set_ylim([-1, 0])
axes[1].set_xlim([1, -1]);
```

Below the code, two plots are shown. The left plot is a time-series graph with the x-axis ranging from 0 to 10 and the y-axis from -2.0 to 2.0. It features two oscillating lines: a red line labeled 'pendulum1' and a blue line labeled 'pendulum2'. The right plot is a phase space graph with the x-axis from 1.0 to -1.0 and the y-axis from -1.0 to 0.0. It shows a red parabolic curve at the top and a blue trajectory that oscillates between the curve and the bottom of the plot.

At the bottom left, the Wakari logo is visible. At the bottom right, a caption reads: 'Simple animation of the pendulum motion. We will see how to make better animation in Lecture 4.'



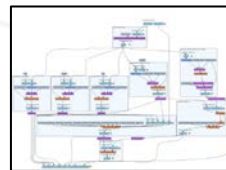
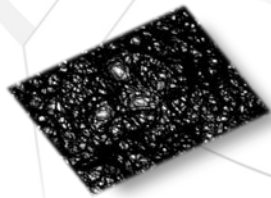
# Digital Science - Reproducibility and Visibility in Astronomy Research Objects

Similar Initiative to ESO Telbib

## ADSLabs

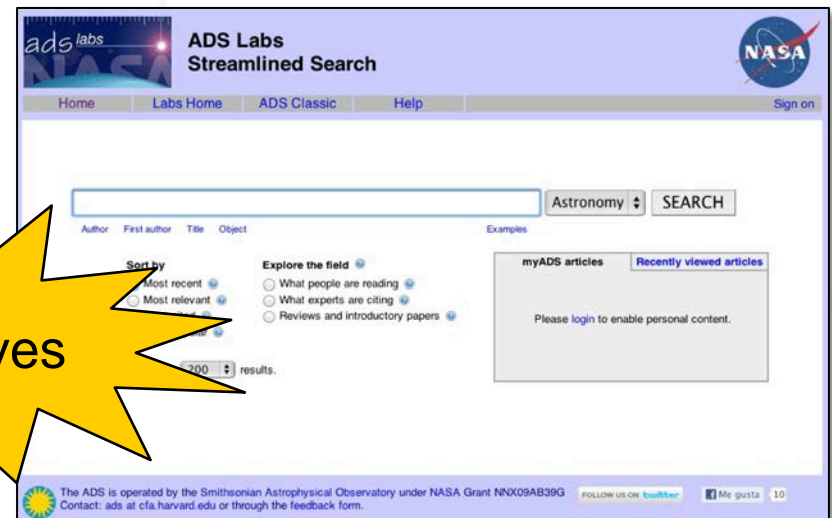
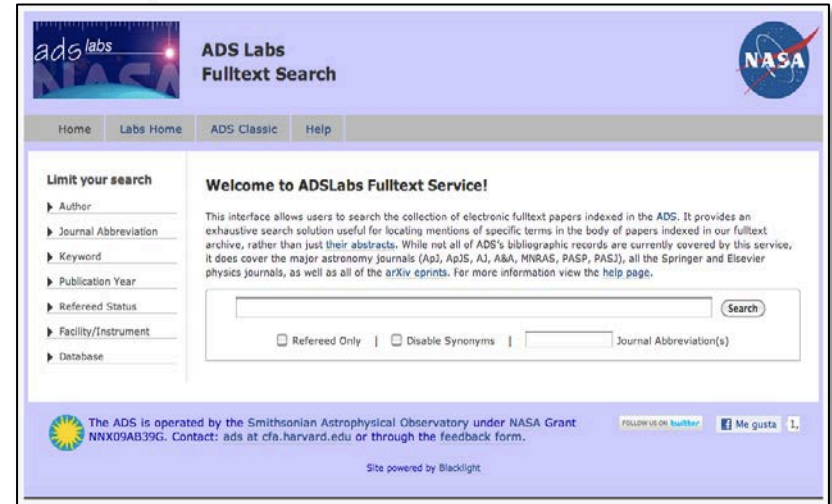
### ADO Linked Components

- » 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

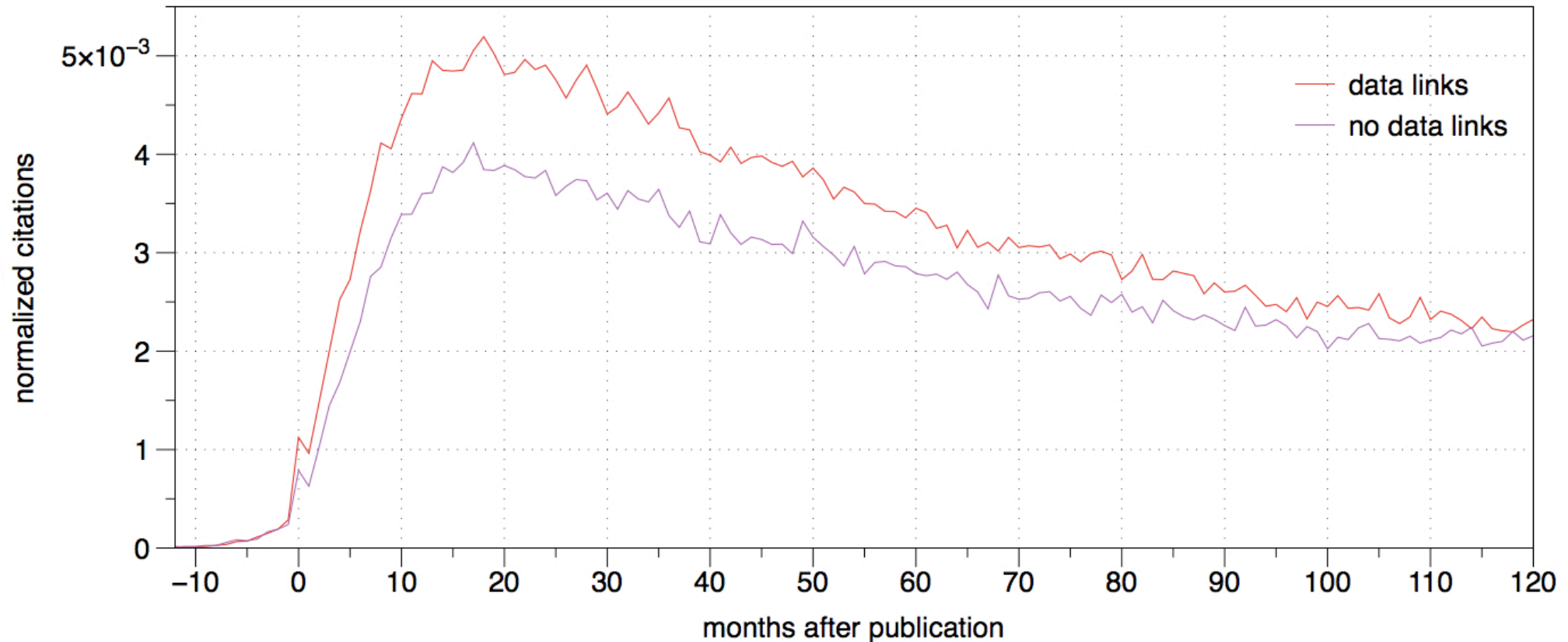
<http://labs.adsabs.harvard.edu/>



## The Incentive

Papers with data links are cited more than those without

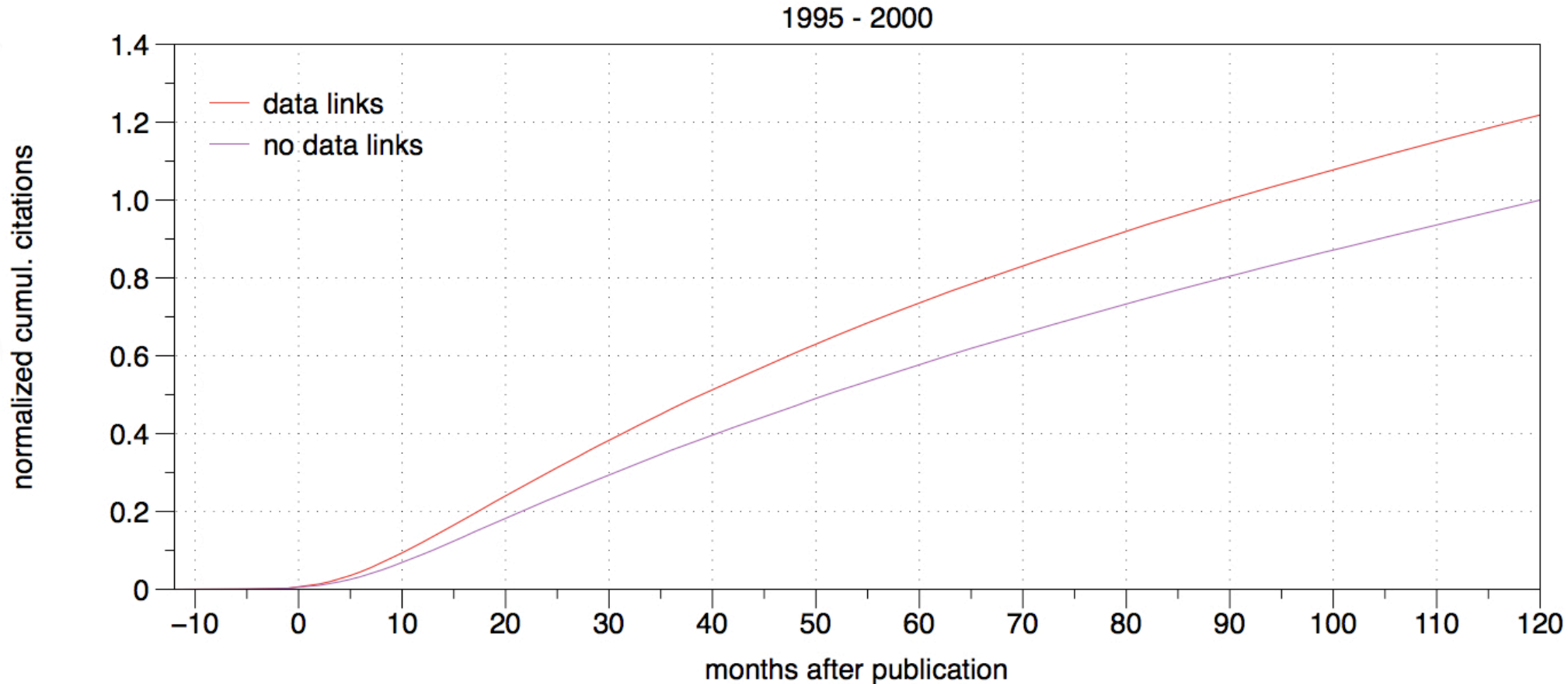
1995 - 2000



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.

# Digital Science - Reproducibility and Visibility in Astronomy

## 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

