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)
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
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 Machine Learning (Schultz, 2010):

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>Time to document and clean up</td>
</tr>
<tr>
<td>52%</td>
<td>Dealing with questions from users</td>
</tr>
<tr>
<td>44%</td>
<td>Not receiving attribution</td>
</tr>
<tr>
<td>40%</td>
<td>Possibility of patents</td>
</tr>
<tr>
<td>34%</td>
<td>Legal Barriers (ie. copyright)</td>
</tr>
<tr>
<td>-</td>
<td>Time to verify release with admin</td>
</tr>
<tr>
<td>30%</td>
<td>Potential loss of future publications</td>
</tr>
<tr>
<td>30%</td>
<td>Competitors may get an advantage</td>
</tr>
<tr>
<td>20%</td>
<td>Web/disk space limitations</td>
</tr>
</tbody>
</table>
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 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

Reproducible
Highly Discoverable
Re-usable
Going beyond automation

Organization
# Digital Astronomy in the Local Desktop

## Workflows to Access and Massage VO Data

### Capture Actions, Tasks, Dependencies, Provenance

### Improve Clarity and Reproducibility

<table>
<thead>
<tr>
<th>CIG</th>
<th>Vhel</th>
<th>e_Vhel</th>
<th>r_Vhel</th>
<th>Dist</th>
<th>MType</th>
<th>e_MType</th>
<th>OptAssym</th>
<th>r_MType</th>
<th>Bmag</th>
<th>e_Bmag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7299.0</td>
<td>3.0</td>
<td>1</td>
<td>96.9</td>
<td>5.0</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>14.167</td>
<td>0.271</td>
</tr>
<tr>
<td>2</td>
<td>6983.0</td>
<td>6.0</td>
<td>2</td>
<td>94.7</td>
<td>6.0</td>
<td>1.5</td>
<td>0</td>
<td>1</td>
<td>15.722</td>
<td>0.324</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>4.0</td>
<td>1.5</td>
<td>0</td>
<td>1</td>
<td></td>
<td>16.057</td>
<td>0.507</td>
<td>0.246</td>
<td>0.354</td>
</tr>
<tr>
<td>4</td>
<td>2310.0</td>
<td>1.0</td>
<td>3</td>
<td>31.9</td>
<td>3.0</td>
<td>1.5</td>
<td>0</td>
<td>1</td>
<td>12.818</td>
<td>0.424</td>
</tr>
<tr>
<td>5</td>
<td>7865.0</td>
<td>10.0</td>
<td>3</td>
<td>105.9</td>
<td>0.0</td>
<td>1.5</td>
<td>0</td>
<td>1</td>
<td>15.602</td>
<td>0.364</td>
</tr>
<tr>
<td>6</td>
<td>5164.0</td>
<td>9.0</td>
<td>2</td>
<td>68.5</td>
<td>5.0</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>14.445</td>
<td>0.325</td>
</tr>
</tbody>
</table>
Digital Science - Reproducibility and Visibility in Astronomy

Scientific Workflows

Digital Libraries of workflows may boost the use of the existing infrastructure of data (VO)
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
AstroTaverna: Create, annotate and run a workflow

http://amiga.iaa.es/p/290-astrotaverna.htm
AstroTavera: Create, annotate and run a workflow

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

<table>
<thead>
<tr>
<th></th>
<th>Low Res</th>
<th>High Res</th>
<th>Extreme Res</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td>4 Bytes</td>
<td>4B</td>
<td>4 Bytes</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>2,048 x 2,048</td>
<td>16MB</td>
<td>8,192 x 8,192</td>
</tr>
<tr>
<td><strong>Channels</strong></td>
<td>16,384</td>
<td>0.27TB</td>
<td>16,384</td>
</tr>
<tr>
<td><strong>Stokes &amp; Weighting</strong></td>
<td>1</td>
<td>0.27TB</td>
<td>1</td>
</tr>
</tbody>
</table>
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 $\approx 625\text{TB}$
The next generation of archives

Digital Science - Reproducibility and Visibility in Astronomy

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

<table>
<thead>
<tr>
<th>Archives should evolve from data providers into</th>
</tr>
</thead>
<tbody>
<tr>
<td>» Virtual Data providers</td>
</tr>
<tr>
<td>» Software Tasks providers</td>
</tr>
</tbody>
</table>

» Archives speaking Web Services

Astronomy of multi archives/facilities/wavelength
Interconnected and interoperable archives
» Virtual Observatory
» Software Tasks

The next generation of archives

Digital Science - Reproducibility and Visibility in Astronomy
The next generation of archives

Process should benefit of the same privileges acquired by data
Preserving the method ensures replication of final results at any moment
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
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

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