LOFAR data: from archive to arXiv

Alexandar Mechev, Leiden University
The LOFAR
The Science

Different window
Complimentary Science
The Array

7000 Antennas
1900 km baselines
10 Tbps raw data
80 Gbps correlated
10 - 240 MHz
The Survey

Map of northern sky
3000+ observations
8TB each
>20 PB
-Hacky Playground?
-Stable Workflow?

https://tinyurl.com/ADASS19-apmechev-4
The Dream
“Science Ready”

Not ‘one and done’
Requires iteration
Questions harder than answers
The Reality
LOFAR Data

- Large data sets
- Extensive archive
- Development pace
- Small playground!

Not just being FAIR!
Fighting EVIL
What is EVIL?

- Esoteric
- Versionless
- Irreproducable
- Laborious
We fight the Esoteric

- Define Scientific Workflows
- Make it runnable ‘at home’

We fight the Versionless

- Use (singularity/docker) images
- Version and even test them!
We fight Irreproducibility

- Well defined data lineage
- Trivial re-processing

We fight Laboriousness

- Automate processing
- Automate fault detection
The Solution
Processing setup

- Distributed Processing
- Apache Airflow
  - Automate testing
- Infrastructure independent jobs
  - ‘self-contained/defined’
  - Can use clouds
- Or Jupyter

ponents: tinyurl.com/ADASS19-apmechev-5
Archive -> arXiv

- Data triggers
- Pipeline launch (NL-grid)
  - Diagnostics -> HTTP (CouchDB)
- Data delivery (http/macaroons)
- Imaging (Wherever)
- Images published

- Reproducible: Just run it again
The Future
Successes

>1000 Datasets
Two Archive locations
As fast as Observing

Easy to Use
Easy to parallelize
Reproducible!

Challenges

The users problem
The authentication

The interfaces
The processing resources

The Questions
What good are computers? They can only give you answers.

-Picasso

That’s why we have scientists.
Thank you!

With the help of:

**Leiden University**
- Huib Intema
- Aske Plaat
- Timothy Shimwell
- Huub Rottgering
- Frits Sweijen

**SURF{Sara**
- Natalie Danezi
- Raymond Oonk
- Coen Schrijvers

**ASTRON**
- Zheng Meyer-Zhao
- Yan Grange

github.com/apmechev/{GRID_LRT,AGLOW}