Scientific reproducibility in the Spanish SKA Regional Centre prototype

Susana Sánchez Expósito, Lourdes Verdes-Montenegro, Julian Garrido, Sebastian Luna, Javier Moldon, Laura Darriba, Michael Jones

Instituto de Astrofísica de Andalucía (CSIC)

28th November 2018

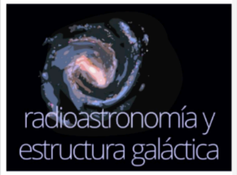
PRESENTATION

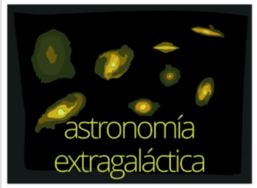
http://www.iaa.csic.es

IAA in a Nutshell

CSIC - Physical Science and Technologies

Research





Technology

Space Instrumentation

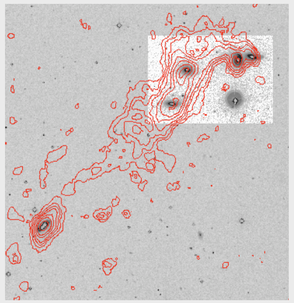
Ground-based Instrumentation

Observatories

AMIGA group

http://amiga.iaa.es

Analysis of the Interstellar Medium of Isolated Galaxies



Hickson Compact Group 16
Verdes-Montenegro+. 2001

Science

Multi-wavelength studies of galaxies in two extreme environments: isolated galaxies and dense compact groups.

Technology

Data interoperability, distributed computing

AMIGA, involved in SKA at different levels:

- Coordination of the Spanish participation in SKA
  - * Spain joined SKA in June 2018
- Science: HI SKA Science Working Group
- Technological:
  - SKA Science Data Processor.
  - SKA Regional Centres:
    - **SRC Coordination Group**
    - **SRC Steering Committee**
    - **AENEAS**: Design of a federated European Science Data Centre for SKA

Advanced European Network of E-infrastructures for Astronomy with the SKA

EXCELENCIA SEVERO OCHOA

SRC PROTOTYPE AT IAA-CSIC

Involvement of IAA teams in SKA prec/path:

MeerKAT

ASKAP

- WSRT-Apertif
- e-MERLIN
- JVLA/EVLA:
- GMRT
- Arecibo
- Parkes

Objectives

Develop the technical capabilities and expertise to get accredited as an SRC

Strengthen scientific activities related to SKA and its precursors Prepare for KSPs

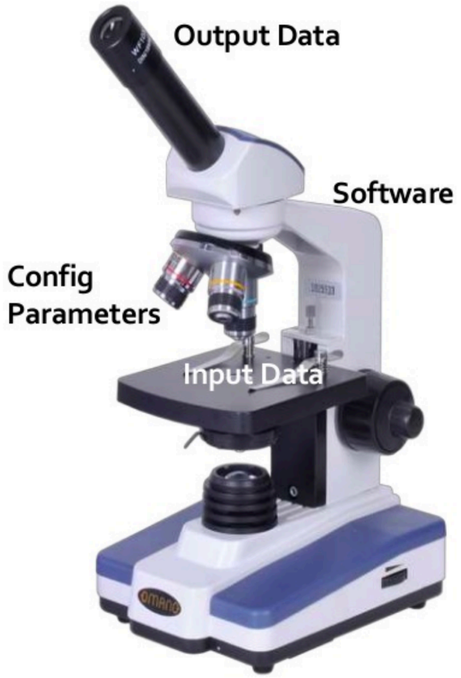
Enable Open Science and FAIR principles

Transversality/wavelength agnostic

EXCELENCIA SEVERO OCHOA

Scientific Reproducibility in Computational Research

Witnessing "Datascopes"



Experiment

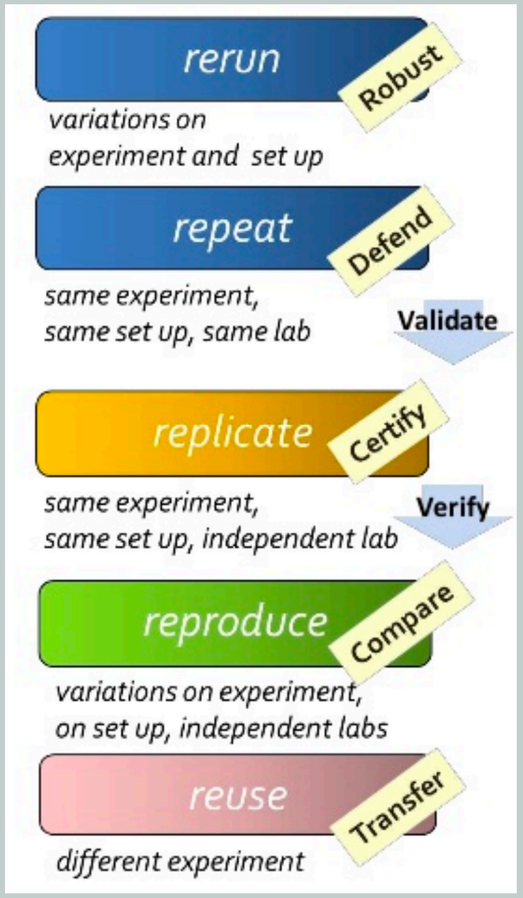
Materials
datasets, parameters, algorithm seeds

Methods
techniques, algorithms, spec. of the steps, models

Setup

Instruments
codes, services, scripts, underlying libraries, workflows, , ref resources

Laboratory
sw and hw infrastructure, systems software, integrative platforms computational environment



What is Reproducibility? The R Brouhaha. Carole Goble. 2016.
https://www.slideshare.net/carolegoble/what-is-reproducibility-gobleclean

Need:

- Input, intermediate and result data
- Scripts, software tools involved in the data analysis
- Enough metadata describing both data and method
- Human readable annotations
- Provenance
- Computing / Execution environment

Availability

↓

FAIR- ability

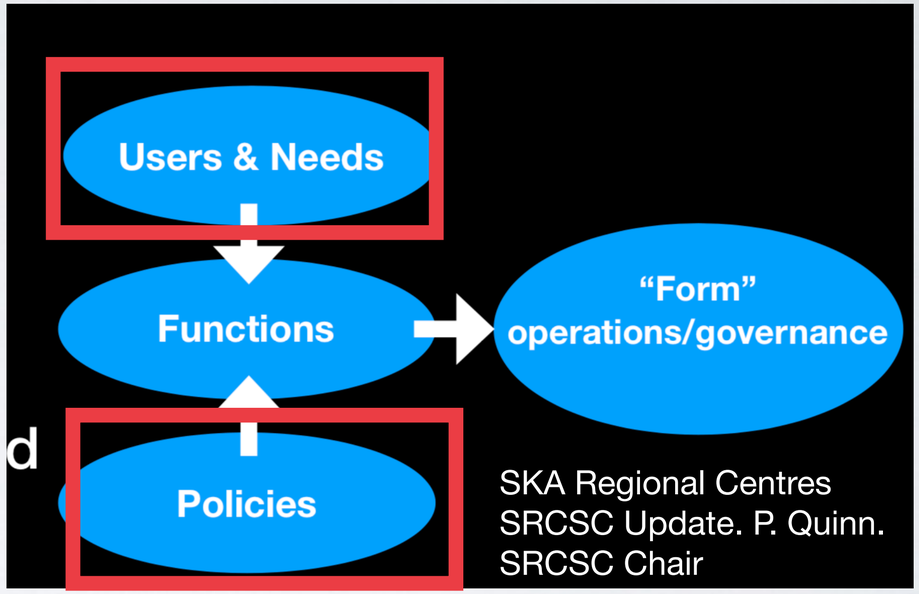
FAIR and Open Science

Availability → *FAIR- ability*

FAIR (www.go-fair.org) is a multi-disciplinary bottom-up initiative to make scientific data reusable. The FAIR principles state that scientific data should be:

Scientific codes/ workflows also need to be FAIR!

- **Findable:** Data have sufficient metadata and unique, persistent identifiers in a searchable database.
- **Accessible:** Data is stored in trusted/standard repository. Metadata and data can be understood by machines/people.
- **Interoperable:** Metadata use a standard language, external connections to other data/resources are qualified.
- **Reusable:** Data have sufficient provenance information and clear licenses.



IDENTIFYING SRC FUNCTIONALITIES FOR SUPPORTING SCIENTIFIC REPRODUCIBILITY FROM TWO PERSPECTIVES(*)

(*) See more about perspectives in: *Why Open Science at SRCS?* L. Verdes-Montenegro et al.
https://zenodo.org/record/2636349#.Xdx8IiUo90s

"SKA-LINK PERSPECTIVE"

SKA-Link: combining knowledge to pioneer Big-Data solutions for SKA Regional Centres

http://amiga.iaa.es/p/330-SKA-Link.html

- 2-year project funded by CSIC i-Link program
- PI: L. Verdes-Montenegro (IAA-CSIC)
- Overall aim:

Set of **best practices** that assist the SKA community to successfully exploit SKA data in a FAIR way, with an emphasis on the use of technologies that facilitate the reproducibility

List of Best Practices (draft)

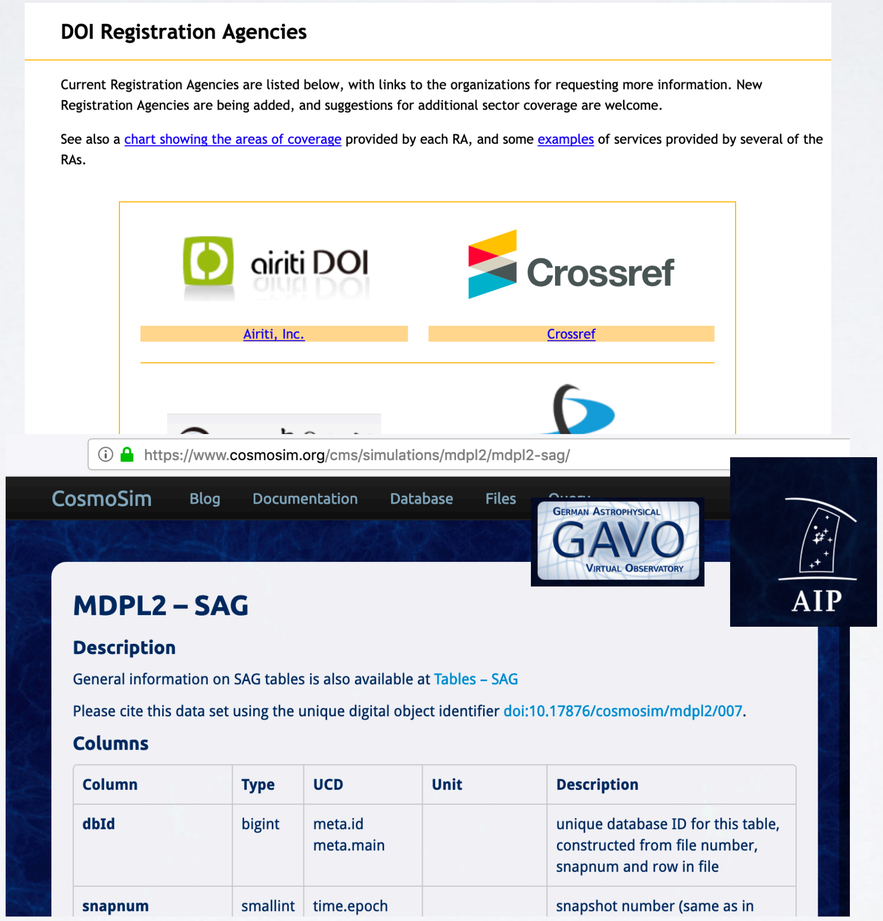
- Unified access for the SRCs Network
- **SKA data should fulfil IVOA standards**
- Publish and share SKA studies and results as Research Objects
- **Provide Persistent Identifiers**
- Ensure SKA data and workflows are released under clear usage licenses
- **Automatic capture according to standards of workflow provenance**
- Certification process
- SRCs should be staffed by skilled people experienced in data science and data stewardship (Data stewards are in charge of ensuring data can be accessed and reused over time, providing documentation or taking care of the technology changes)
- Provide assessment frameworks to review/validate SKA studies
- Gather the metric indicators to track the adoption of scientific methods in SRCs

UNIQUE AND PERSISTENT IDENTIFIERS

Findable: Data have sufficient metadata and unique, persistent identifiers in a searchable database.

- Make data/tools citable, findable and traceable in the **long-term**

- **Enable connections among input data, result data, tools**, etc. involved in an experiment as well as with the **authors** (ORCID).



It would be part of the data curation to be carried out by the SRCs:

– Provide SKA data (ODP & ADP) with unique Persistent IDentifiers (e.g. DOIs).

– Register SKA data PIDs in Registry Agencies.

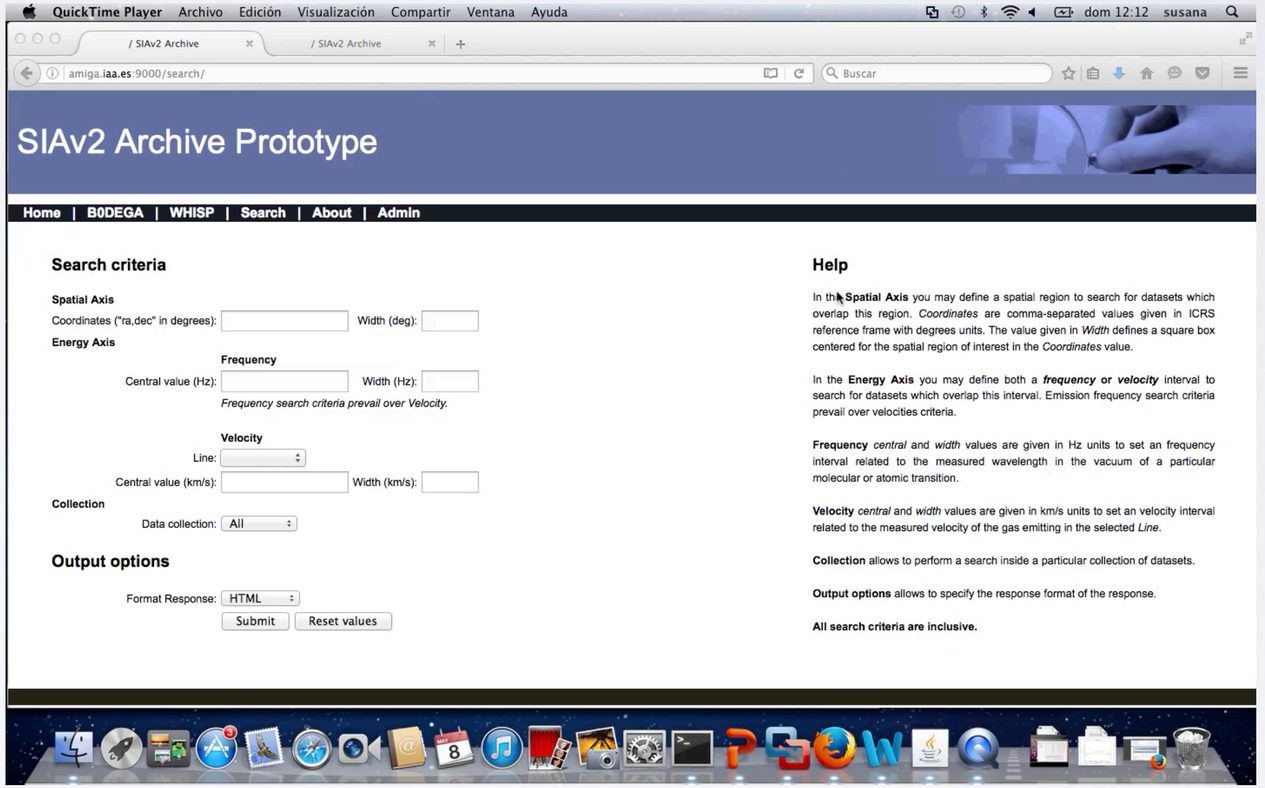
– Keep DOI metadata updated (e.g link to the data products location)

IVOA STANDARDS

Accessible: Data is stored in trusted/standard repository. Metadata and data can be understood by machines/people.
Interoperable: Metadata use a standard language, external connections to other data/resources are qualified.

- IVOA standards:
  - **SIAv2**: queries for images /cubes by parameters (e.g DPType=cube, POS,BAND)
  - **SODA / DataLink**: CutOut services
  - **HiPS**: Discovery.

- Standards for discovering/accessing visibility data?



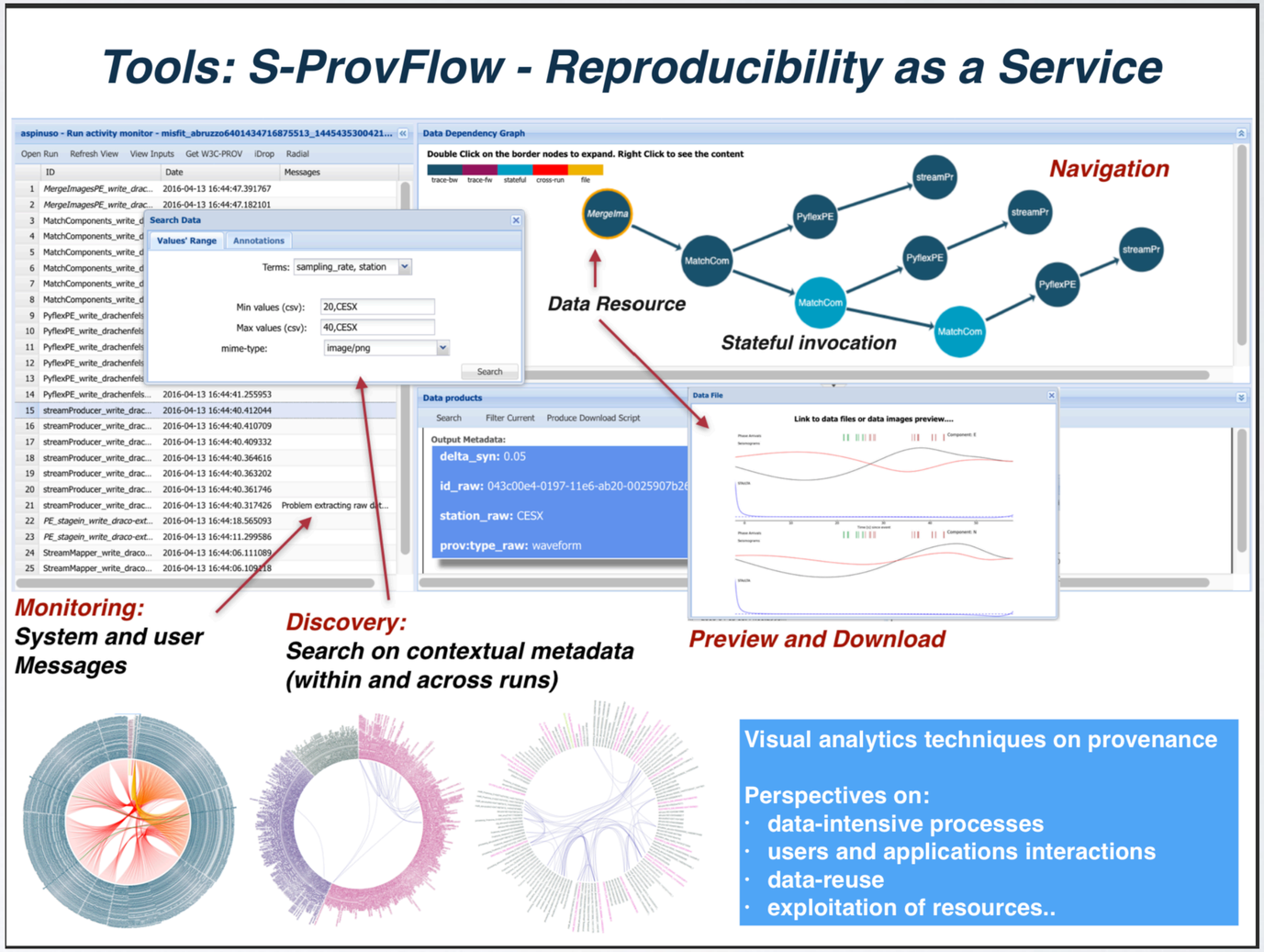
AMIGA development used as a Prototype for the IVOA SIAv2 protocol. J.E. Ruiz 2014

AUTOMATIC CAPTURE OF WORKFLOW PROVENANCE AND ACCORDING TO STANDARDS

Metadata of the workflow execution:

- Input / intermediate / output data
- Configuration parameters
- User executing the workflow
- Software environment: libraries, OS version
- Etc.

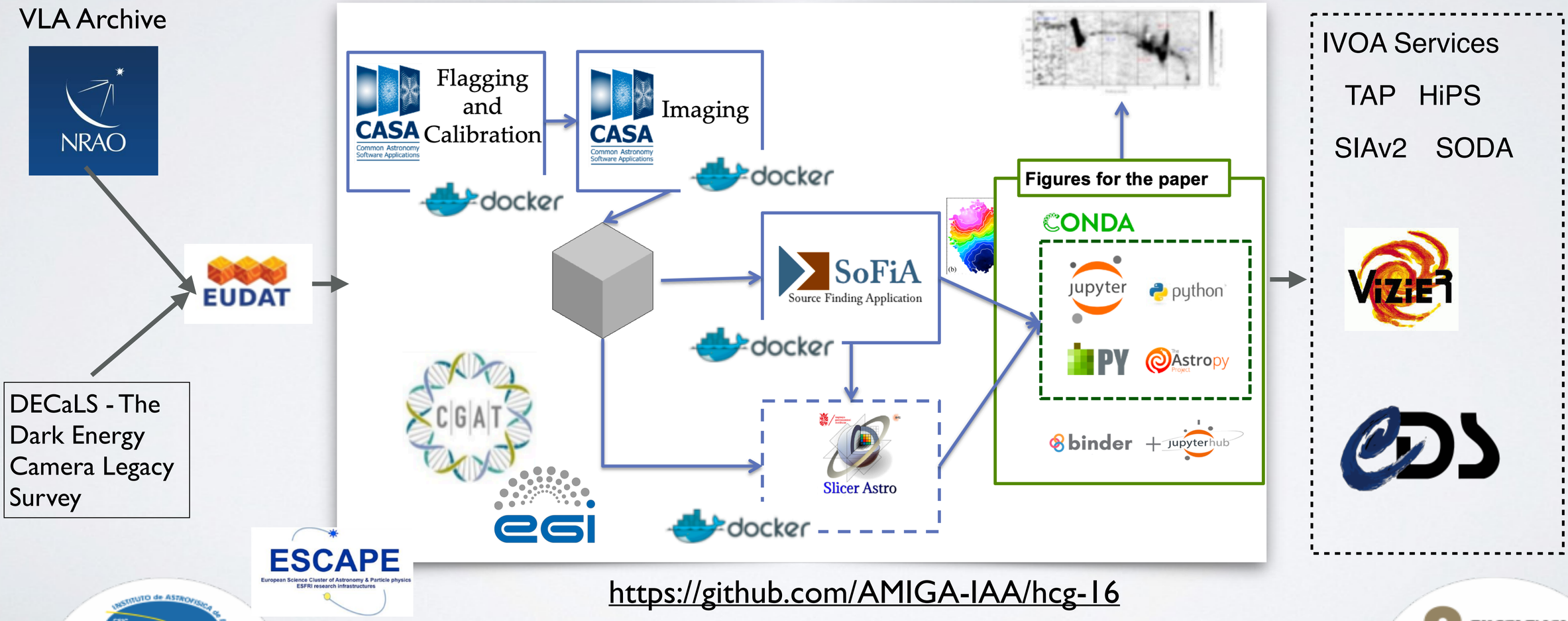
(*) Collaboration with GMV company on this topic. Innoglobal project: "Supporting Open Science in the SKA by collecting Provenance from its Science Data processor"



A. Spinuso (Royal Netherlands Meteorological Institute). SKA-Link Kick-off meeting, April 2017, Granada.

USER PERSPECTIVE

"Evolution of compact groups from intermediate to final stages: A case study of the HI content of HCG 16".
***M.G. Jones**, et al. Accepted in A&A. 2019. https://arxiv.org/abs/1910.03420*

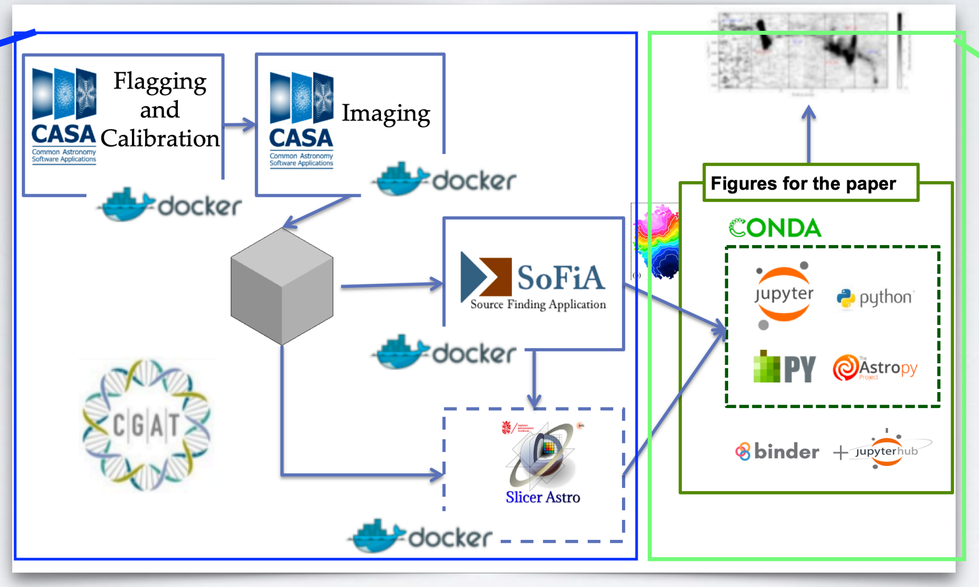


https://github.com/AMIGA-IAA/hcg-16

SEARCHING FOR THE MOST SUITABLE TOOL FOR EACH WORKFLOW STEP

Containers + Workflow Manager

- Prone to batch mode
- Computing efficiency
- Easy software deployment
- Avoid problems with software dependencies



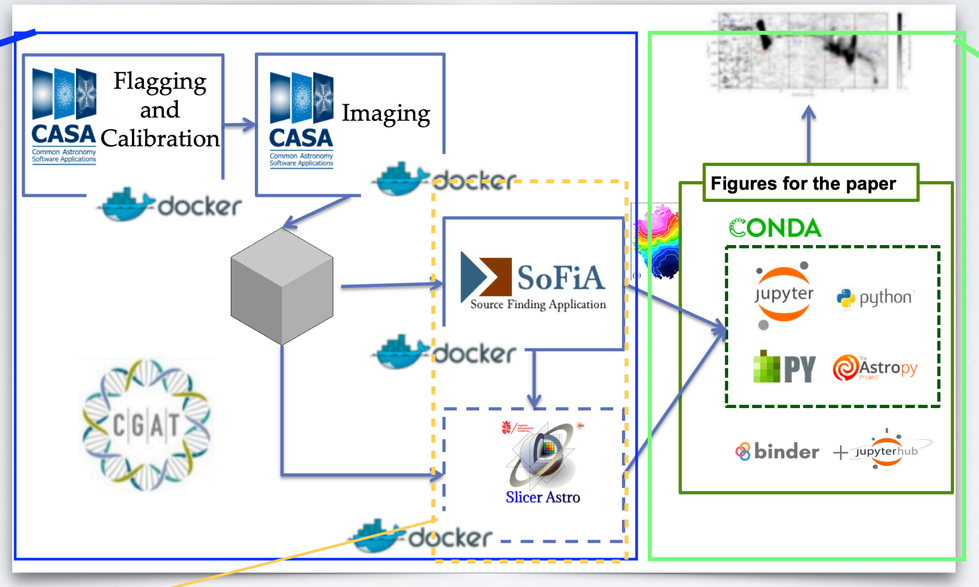
Notebooks

- Interactivity
- Better understanding of the analytical tasks
- Speed "trial and error" cycle
- Reproducing paper figures/plots
- Easy verification by colleagues
- Increase of re-usability

SEARCHING FOR THE MOST SUITABLE TOOL FOR EACH WORKFLOW STEP

Containers + Workflow Manager

- Prone to batch mode
- Computing efficiency
- Easy software deployment
- Avoid problems with software dependencies



Can all/most processing be done in an unsupervised (batch) mode or is manual interaction/checking essential?

- **Supervised**:
  - Multiscale deconv. of strongest sources
  - Source-finding, masking may require human inspection in complicated cases (interacting galaxies)
- **Unsupervised:**
  - Combination of cubes from individual pointings
  - Mosaicking
  - Generation of cutouts, minicubes, moment maps

Perspective from the HISWG on SKA Data Processing and Science Analysis. L.Verdes-Montenegro. AENEAS meeting

Notebooks

- Interactivity
- Better understanding of the analytical tasks
- Speed "trial and error" cycle
- Reproducing paper figures/plots
- Easy verification by colleagues
- Increase of re-usability

THANKS!