Open Science and reproducibility

Julián Garrido Sánchez (@tetrarquis)
Instituto de Astrofísica de Andalucía (CSIC)

Co-authors: L. Verdes-Montenegro, S. Sánchez, S. Luna, M. G. Jones

SKA data challenges workshop, Bologna
2 October 2019

OPEN

Knowledge is open if anyone is free to access, use, modify, and share it

OPEN KNOWLEDGE FOUNDATION

A WORLD WHERE KNOWLEDGE CREATES POWER FOR THE MANY, NOT THE FEW.
THIS IS THE WORLD WE CHOOSE.

Open Definition, Open Knowledge Foundation, https://okfn.org
Open Science, European Commission, 2017, doi: 10.2777/75255

OPEN

Knowledge is open if anyone is free to access, use, modify, and share it

OPEN KNOWLEDGE FOUNDATION

A WORLD WHERE KNOWLEDGE CREATES POWER FOR THE MANY, NOT THE FEW.
THIS IS THE WORLD WE CHOOSE.

Open Science represents an approach to research that is collaborative, transparent and accessible

Open Definition, Open Knowledge Foundation, https://okfn.org
Open Science, European Commission, 2017, doi: 10.2777/75255

The normative system of Science

Norm	Counternorm
Communality	Secrecy
Open Sharing	Closed
Universalism	Particularism
Evaluate research on own merit	Evaluate research on reputation
Disinterestedness	Self-interestedness
Motivated by knowledge & discovery	Treat science as a competition
Organized skepticism	Organized dogmatism
Consider all new evidence, even against one’s prior work	Invest career promoting one’s own theories, findings

credit: @pcmasuzzo

THE NORMATIVE SYSTEM OF SCIENCE

Norm	Counternorm
Communality	Secrecy
Open Sharing	Closed
Universalism	Particularism
Evaluate research on own merit	Evaluate research on reputation
Disinterestedness	Self-interestedness
Motivated by knowledge & discovery	Treat science as a competition
Organized skepticism	Organized dogmatism
Consider all new evidence, even against one's prior work	Invest career promoting one's own theories, findings
QUALITY	QUANTITY
Collaborate	Compete

credit: @pcmasuzzo

RESEARCH ASSESSMENT

doi: 10.2777/75255

Top-Down

Research Assessment

Research output

- Research activity
- Publications
- Datasets and research results
- Open source
- Funding

Research process

- Stakeholder engagement /citizen science
- Collaboration and interdisciplinarity
- Research integrity
- Risk management

Service & leadership

- Leadership
- Academic standing
- Peer review
- Networking

Research impact

- Communication and dissemination
- IP (patents, licenses)
- Societal impact
- Knowledge exchange

Teaching and supervision

- Teaching
- Mentoring
- Supervision

Professional experience

- Continuing professional development
- Project management
- Personal qualities

Open Science Career Assessment Matrix

doi: 10.2777/75255

@tetrarquis

RESEARCH ASSESSMENT

March 2017

European Commission

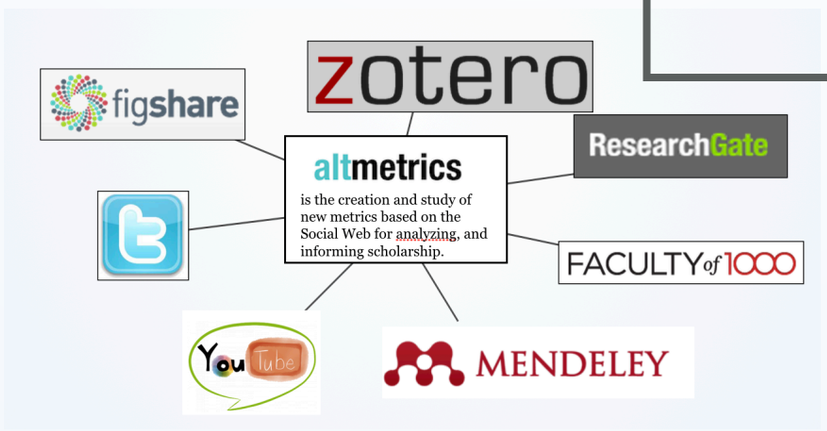
Next-generation metrics:
Responsible metrics and evaluation for open science

Report of the European Commission Expert Group on Altmetrics

Not just citation of articles, various forms of social media shares, web-downloads, any other measure of the Q and impact of research outcomes

Community driven

- DORA declaration
- Metric Tide
- Leiden Manifesto
- etc



European Commission

April 2018

Mutual Learning Exercise

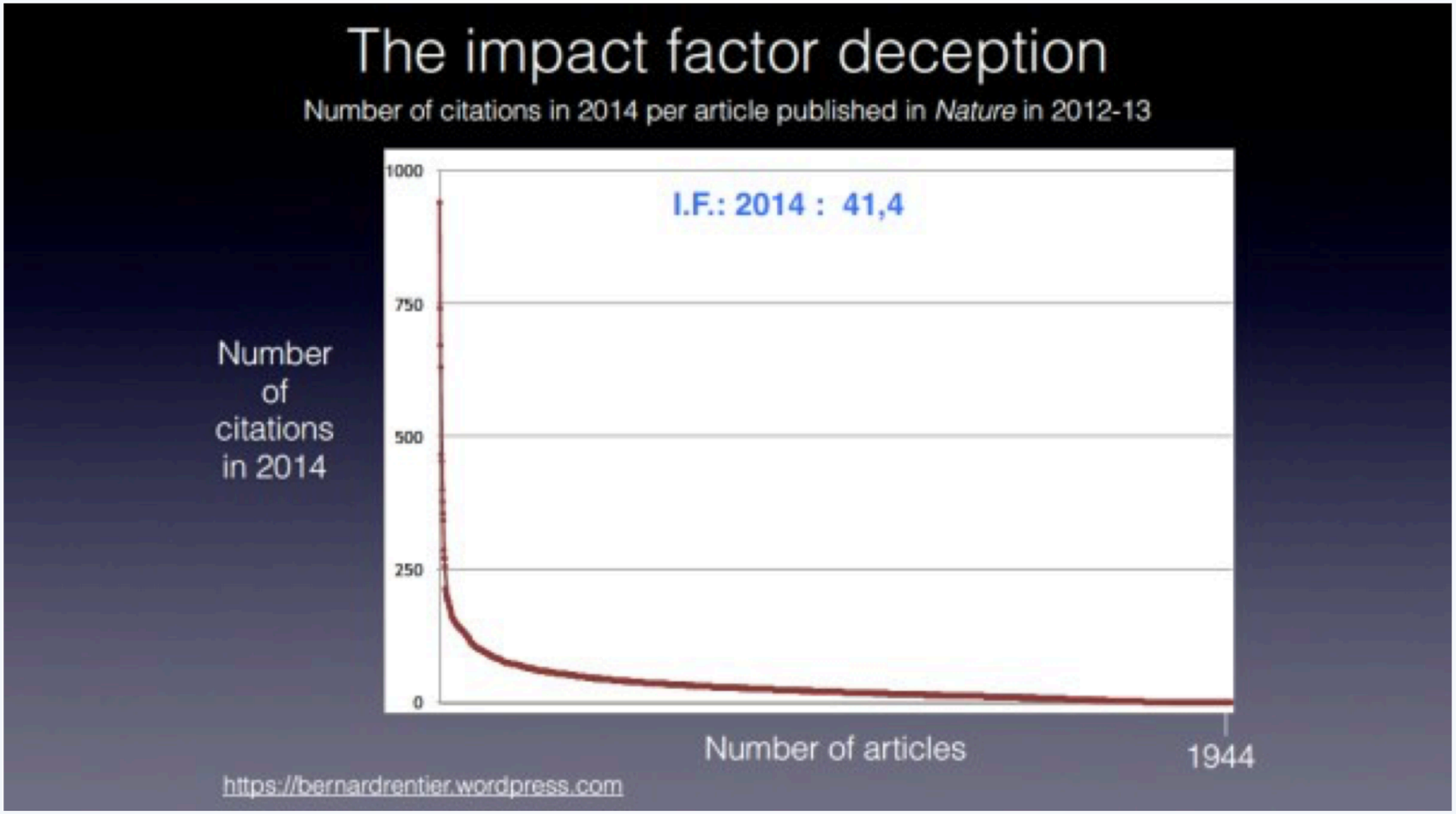
Open Science: Altmetrics and Rewards

Horizon 2020 Policy Support Facility

Thematic Reports:
Types
Use in the context of Open Science
Incentives and Rewards
Strategies, Experiences and Models
Final Report

RESEARCH ASSESSMENT

- A study was carried out on all 1,944 articles published in Nature in 2012 and 2013.
- Cites in 2014.
  - 280 (14.4%) do account for half of the total citations
  - 214 (11%) get 0 or 1 citation



doi: 10.2777/75255

cOALition S

What is cOALition S?

On 4 September 2018, a group of national research funding organisations, with the support of the European Commission and the European Research Council (ERC), announced the launch of cOAlition S, an initiative to make full and immediate Open Access to research publications a reality. It is built around Plan S, which consists of one target and 10 principles.

cOAlition S signals the commitment to implement, by 1 January 2020, the necessary measures to fulfil its main principle:

"By 2020 scientific publications that result from research funded by public grants provided by participating national and European research councils and funding bodies, must be published in compliant Open Access Journals or on compliant Open Access Platforms."

Plan S: Built on strong principles

- **No** publication should be locked behind a **paywall**
- OA must be **immediate,** i.e. no embargo periods
- **No copyright transfer**; publication under a **CC BY license** by default
- **Transparency** about pricing and contracts
- Funders commit to support **publication fees** at a **reasonable** level
- **Multiple routes** to OA compliance
- Commitment to assess research outputs based on their **intrinsic merit** and NOT venue of publication

Credit: Neil Jacobs, Making full and immediate Open access a reality. 2019

cOALITION S

Alignment of Open Access policies

National funders

FWF Der Wissenschaftsfonds.

ACADEMY OF FINLAND

Agence Nationale de la Recherche ANR

Science Foundation Ireland Sfi For what's next

INFN Istituto Nazionale di Fisica Nucleare

Luxembourg National Research Fund

NWO Netherlands Organisation for Scientific Research

The Research Council of Norway

NATIONAL SCIENCE CENTRE POLAND

arrs SLOVENIAN RESEARCH AGENCY

FORTE: Swedish Research Council for Health, Working Life and Welfare

FORMAS

UK Research and Innovation

NSTC

VINNOVA Sweden's Innovation Agency

Charitable and international funders

wellcome

BILL & MELINDA GATES foundation

World Health Organization

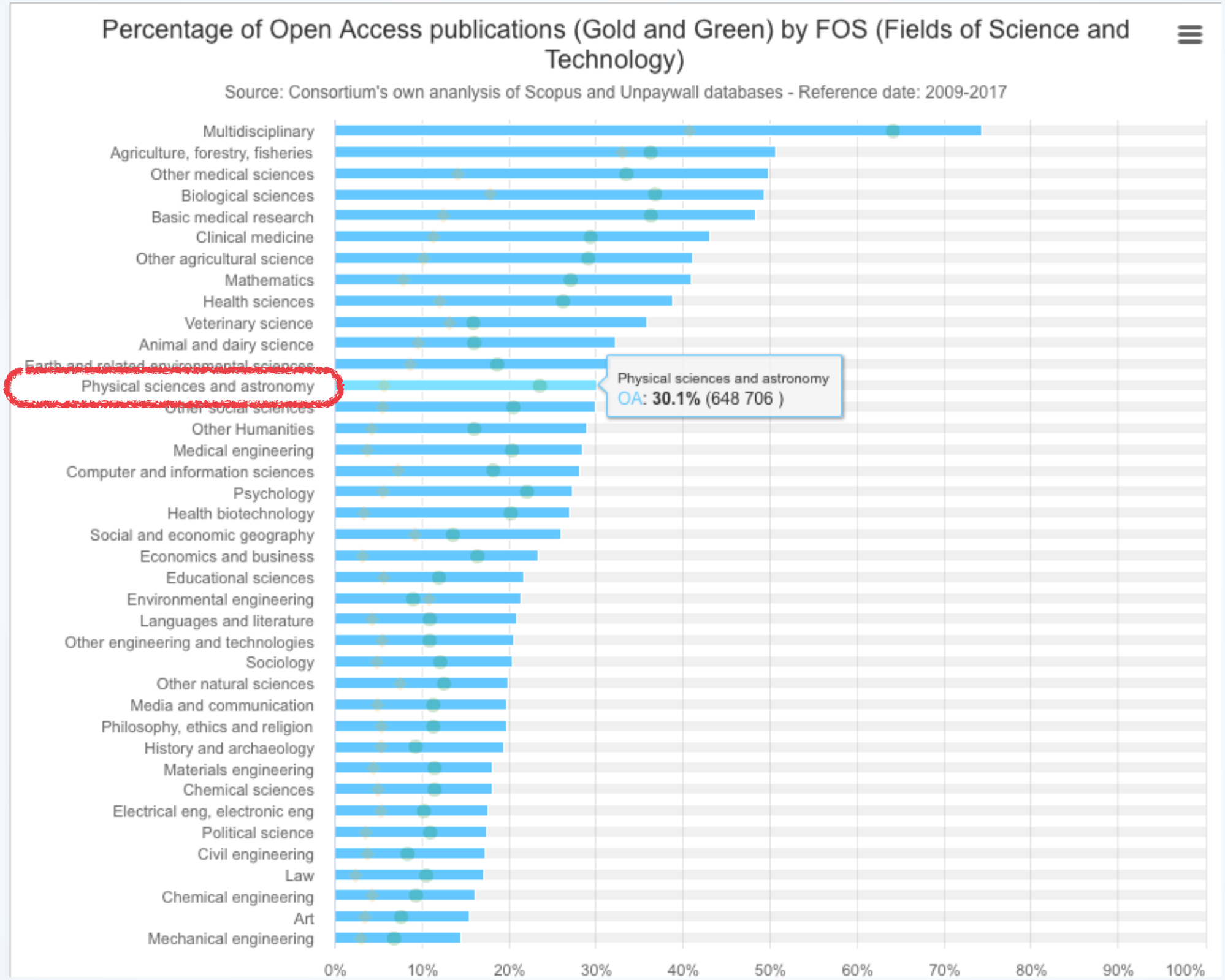
TDR For research on diseases of poverty UNICEF • UNDP • World Bank • WHO

asap Aligning Science Across Parkinson's

European funders

erc European Research Council

OPEN ACCESS PUBLICATIONS

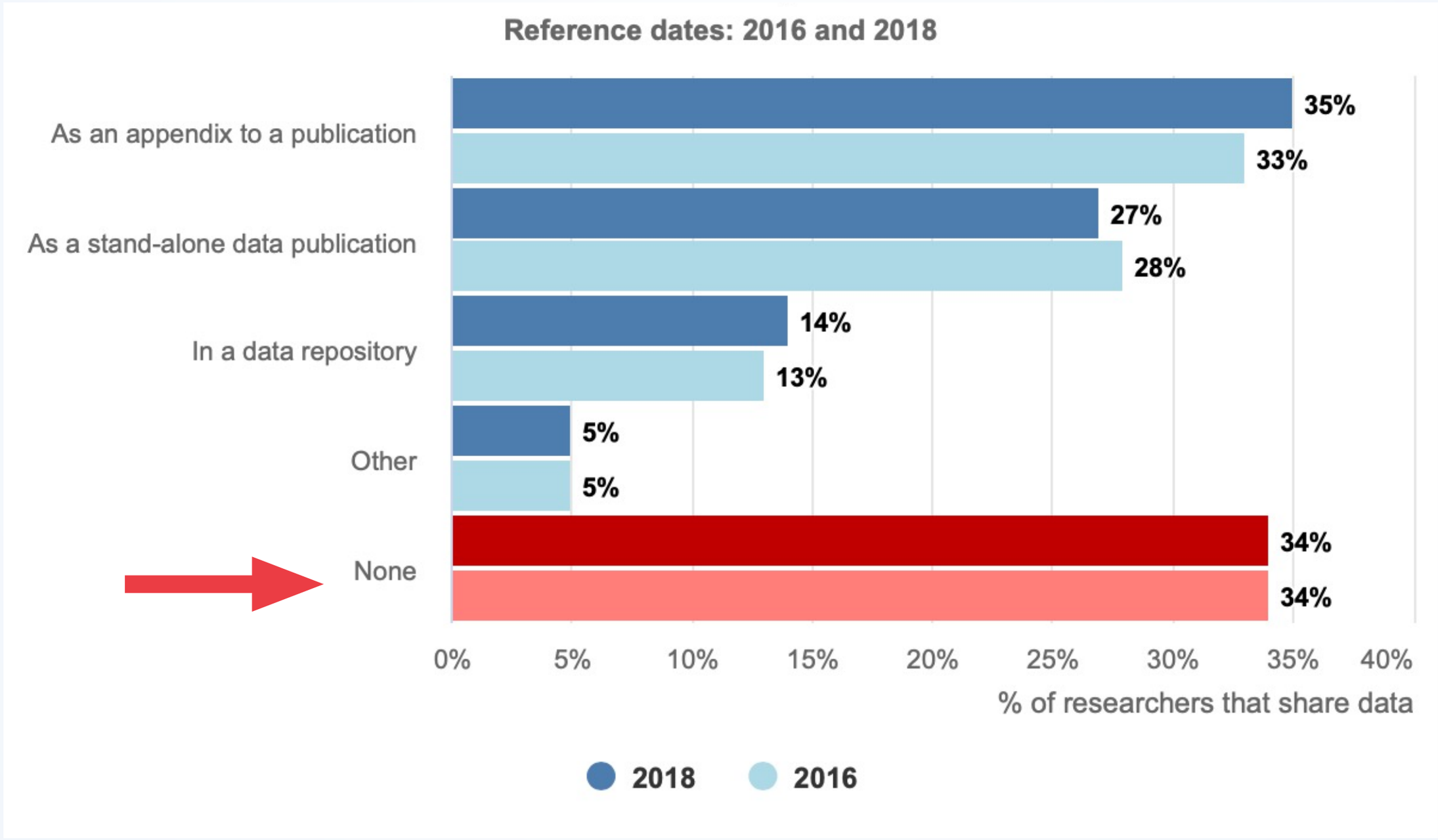


@tetrarquis

Source: Open science monitor

ATTITUDE TOWARDS DATA SHARING

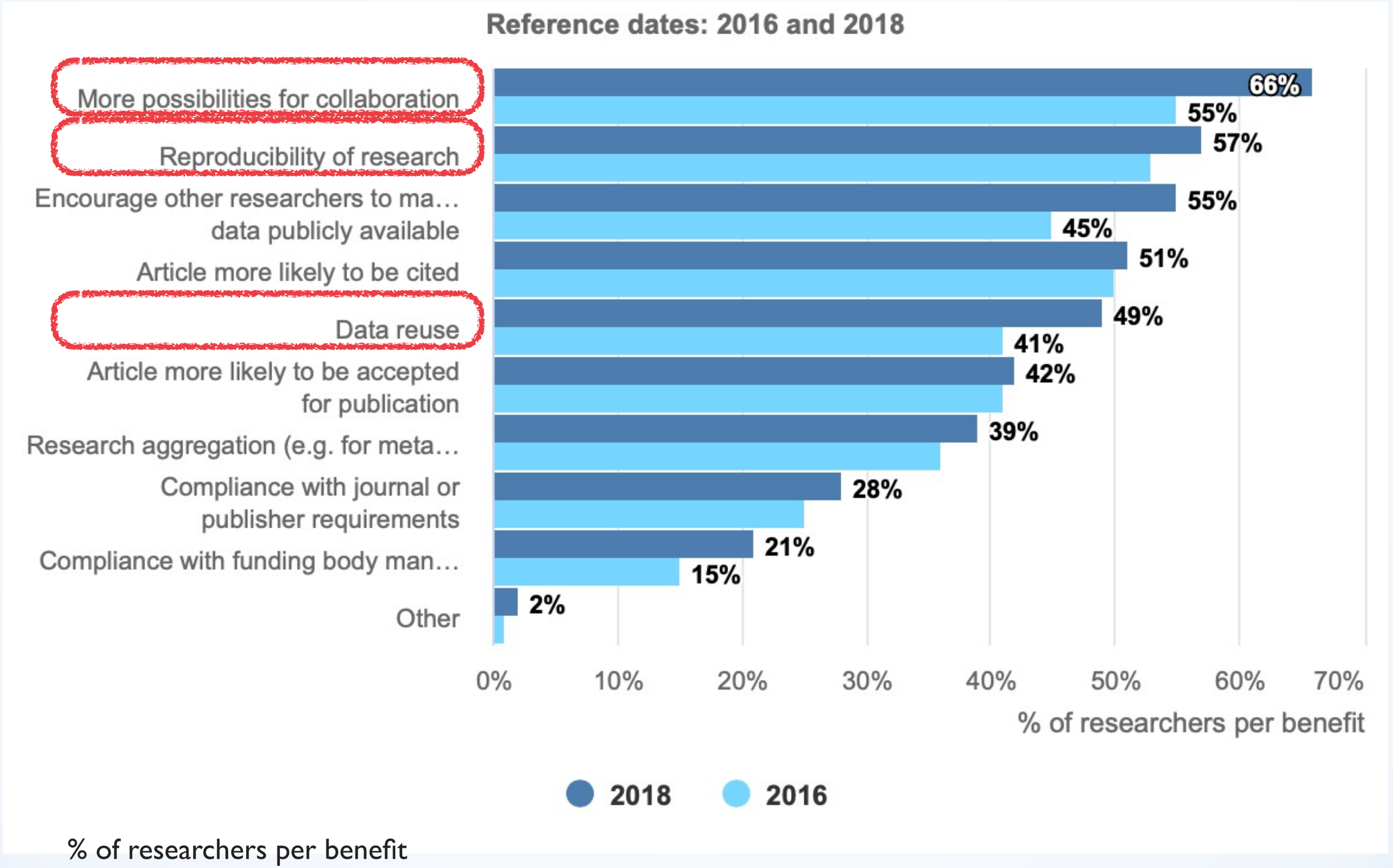
% of researchers that share data



Source: Open science monitor

ATTITUDE TOWARDS DATA SHARING

Benefits of sharing research data

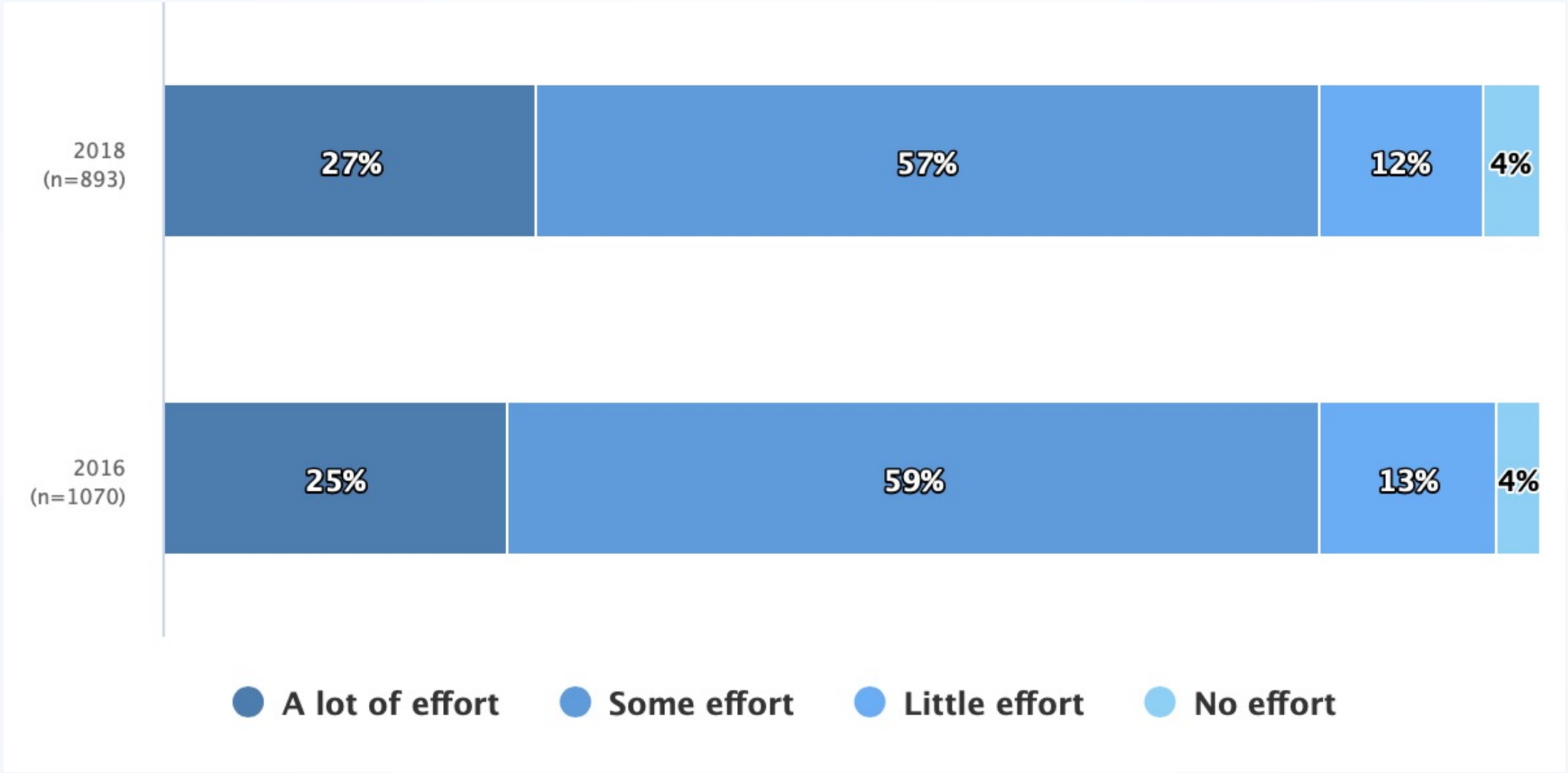


% of researchers per benefit

Source: Open science monitor

ATTITUDE TOWARDS DATA SHARING

Effort required to make research data available and re-usable by others



% of researchers per about of effort

Source: Open science monitor

The AMIGA sample of isolated galaxies. XIII. The HI content of an almost "nurture free" sample

Show affiliations

Jones, M. G.; Espada, D.; Verdes-Montenegro, L.; Huchtmeier, W. K.; Lisenfeld, U.; Leon, S.; Sulentic, J.; Sabater, J.; Jones, D. E.; Sanchez, S.; Garrido, J.

Context. We present the largest catalogue of HI single dish observations of isolated galaxies to date, as part of the multi-wavelength compilation being performed by the AMIGA project (Analysis of the interstellar Medium in Isolated GAlaxies). Despite numerous studies of the HI content of galaxies, no revision focused on the HI scaling relations of the most isolated L_* galaxies has been made since Haynes & Giovanelli (1984, AJ, 89, 758).
Aims: The AMIGA sample has been demonstrated to be almost "nurture free", therefore, by creating scaling relations for the HI content of these galaxies we will define a metric of HI normalcy in the absence of interactions.
Methods: The catalogue comprises of our own HI observations with Arecibo, Effelsberg, Nançay and GBT, and spectra collected from the literature. In total we have measurements or constraints on the HI masses of 844 galaxies from the Catalogue of Isolated Galaxies (CIG). The multi-wavelength AMIGA dataset includes

Open Access

Check if your publisher allows self archiving

- Check the journal policy on the publisher's website

SHERPA/RoMEO

RoMEO Colour	Archiving policy
Green	Can archive pre-print *and* post-print or publisher's version/PDF
Blue	Can archive post-print (ie final draft post-refereeing) or publisher's version/PDF
Yellow	Can archive pre-print (ie pre-refereeing)
White	Archiving not formally supported

Journal:	**Astronomy and Astrophysics** (ISSN: 0004-6361, ESSN: 1432-0746)
RoMEO:	This is a RoMEO **green** journal
Paid OA:	A paid open access option is **available** for this journal.
Author's Pre-print:	✓ author **can** archive pre-print (ie pre-refereeing)
Author's Post-print:	✓ author **can** archive post-print (ie final draft post-refereeing)
Publisher's Version/PDF:	✓ author **can** archive publisher's version/PDF
General Conditions:	• On author's personal website or institutional website or OAI compliant website • Some journals require an embargo for deposit in funder's designated repositories (see journal) • Publisher's version/PDF may be used (see journal) • Must link to publisher version • Publisher copyright and source must be acknowledged • Non-commercial
Mandated OA:	Compliance data is available for 4 funders
Paid Open Access:	Charges and discounts for hybrid Open Ac
Copyright:	Example Policy - Preprint servers / ArXiv - Example Copyright Policy
Updated:	03-Oct-2018 - **Suggest an update for this record**
Link to this page:	http://sherpa.ac.uk/romeo/issn/0004-6361/

Based on: Correia & Principe, 2019

Welcome to Dissemin

Dissemin detects papers behind pay-walls and invites their authors to upload them in one click to an open repository.

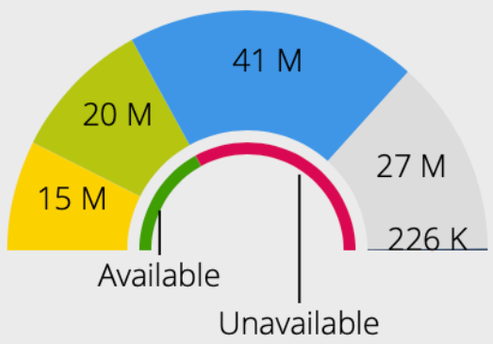
Try any author name

Advanced search

Green open access

Many researchers do not use their right to make their papers freely available online, in addition to the paywalled version offered by traditional publishers.

This forces libraries to buy overpriced electronic subscriptions to journals, when they can afford them at all.



- Available from the publisher 15,369,718
- Available from the author 19,877,776
- Could be shared by the authors 40,713,451
- Unknown/unclear sharing policy 27,413,492
- Publisher forbids sharing 226,134

Open repositories

Uploading your papers on your own webpage is not enough. Such copies are less stable and harder to find than documents uploaded to well-indexed repositories.

Dissemin searches for copies of your papers in a large collection of open repositories and tells you which ones cannot be accessed.

Open Access

Welcome to Dissemin

Dissemin detects papers behind pay-walls and invites their authors to upload them in one click to an open

Papers

← 1 2 3 4 →

Page 1 of 4

Advanced search

PDF

Julián Garrido, Ignacio Requen

Semantic model for flood

Download jh.iwaponline.com

PDF

Jesús C. Echeverría, Pablo de V

A fiber-optic sensor to de
porous silica xerogel film

Download www.researchgate

Julián Garrido, Ignacio Requen

Towards summarizing kn

Upload

PDF

Kristina M. Hettne, Katherine Wolstencroft, Khalid Belhajjame, Carole A. Goble, Eleni Mina, Harish Dharuri, Lourdes Verdes-Montenegro, David De Roure, Julián Garrido, Marco Roos

Best Practices for Workflow Design: How to Prevent Workflow Decay

Towards summarizing knowledge: Brief ontologies

Journal article published in 2012 by Julián Garrido, Ignacio Requena

Full text: Unavailable

Publisher: Elsevier

Preprint: archiving allowed. Upload

Postprint: archiving allowed. Upload

Published version: archiving forbidden. Upload

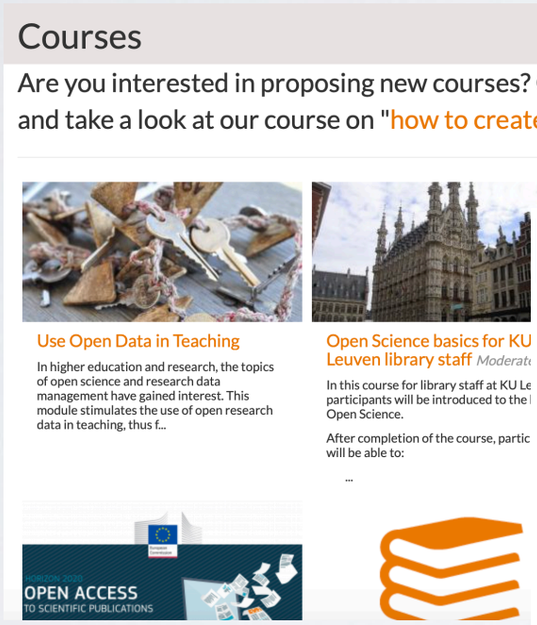
Policy details (opens in a new window). *Data provided by* SHERPA/RoMEO

Contact authors Contact

Open Science - Training

Fostering Improved Training Tools For Responsible Research & Innovation

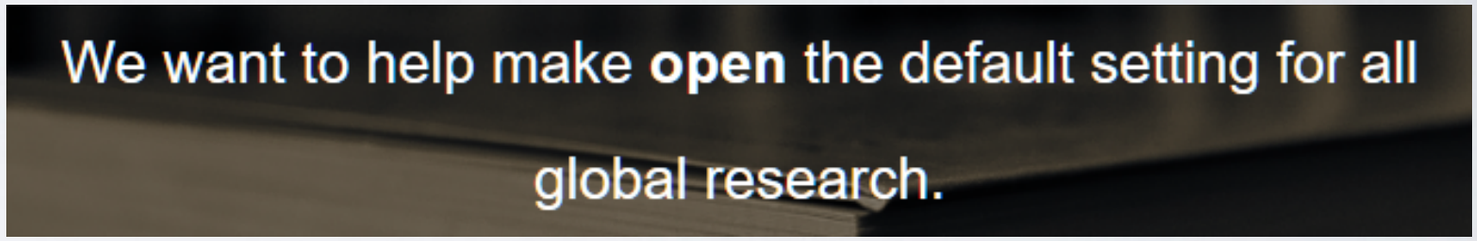
- FIT4RRI maintains a collection of RRI and Open Science training materials on the FOSTER portal.



https://www.fosteropenscience.eu

Open Science - Training

Massive Open Online Course (and Community!)
https://opensciencemooc.eu/



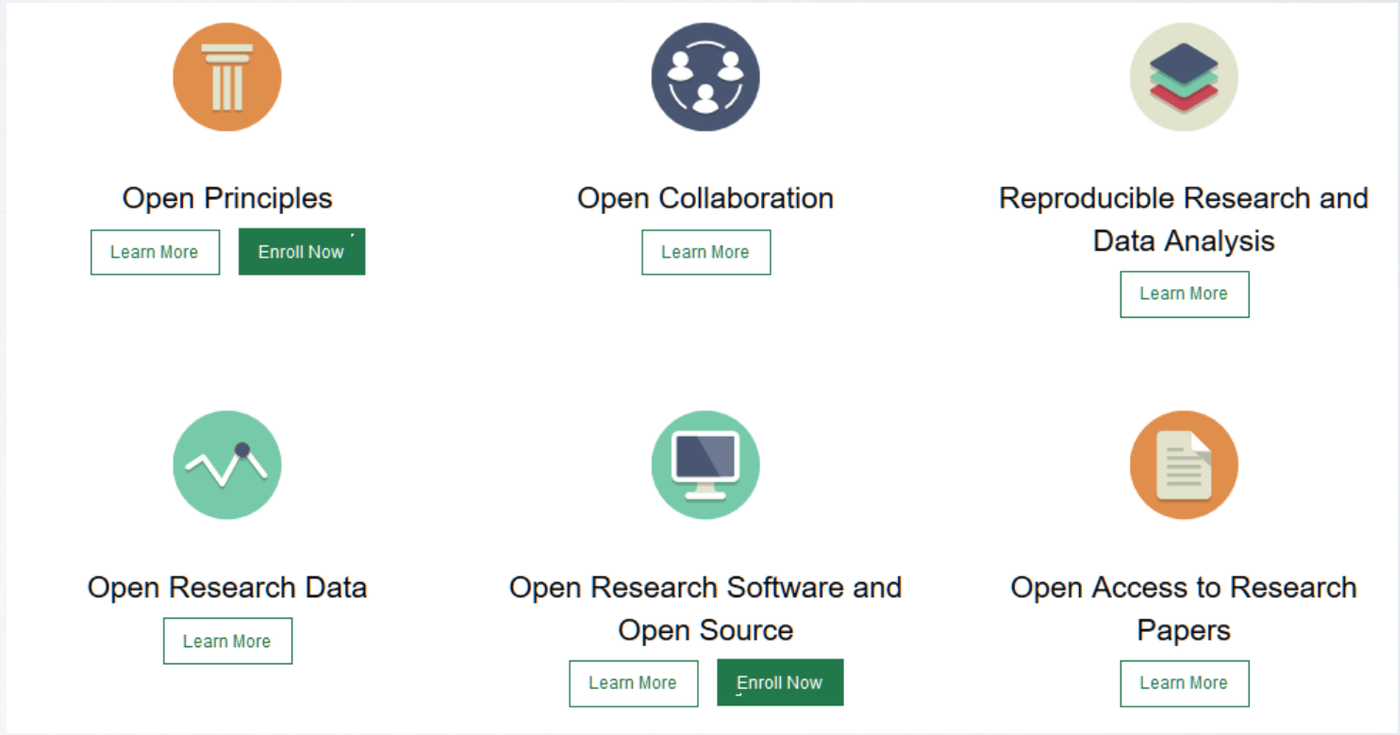
Education, training, support.

Empowerment and leadership.

Shifting power dynamics.

Building a global community.

Massive-scale engagement.



@OpenScienceMOOC

info@opensciencemooc.eu

https://github.com/OpenScienceMOOC

OPEN SCIENCE MOOC
FREE | OPEN | LEARNING

credit: @pcmasuzzo

Open Science, always?

Ethics and limitations

- As Open as possible and as close as necessary
  - Protect results for commercial and industrial exploitation
  - …

GDPR and Anonymized Data

- Data describing personal information is the basis for scientific research in various fields.
- Collecting and processing personal data has been recently regulated by the General Data Protection Regulation for all EU citizens.
- The data management community has proposed data anonymization techniques to allow Open Science.

https://amnesia.openaire.eu

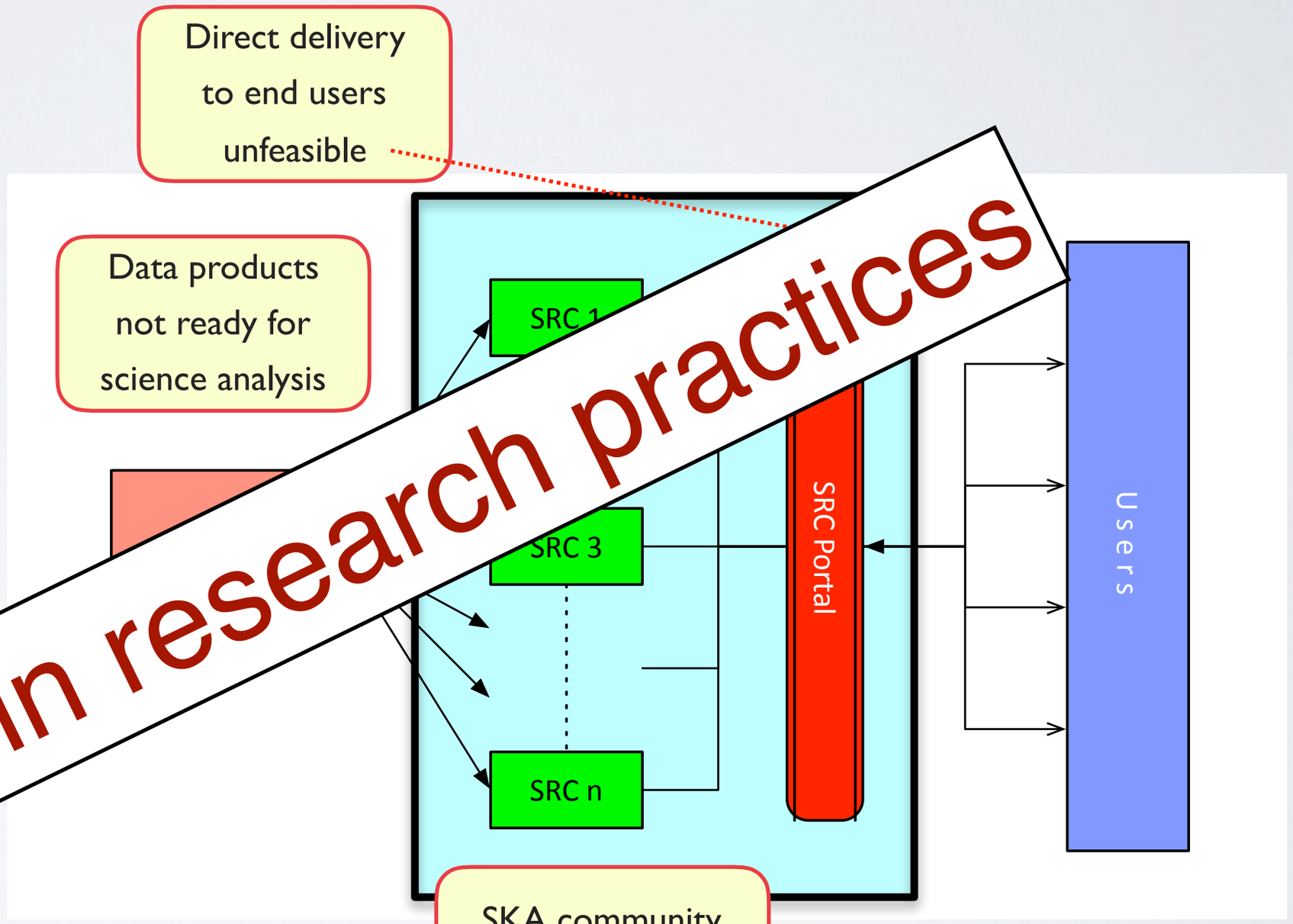
SKA Regional Centres

SKA Organization

The SKA Regional Centre Network

Access to data products, tools and processing power to generate and analyse Advanced Data Products (ADPs)

The core of SKA science



Global shift in research practices

The SRCs as Open Science Hubs

Open Science implementation will facilitate sharing data, resources and tools across the SKA community through the SRCs.** The methods can be verified, reused, repurposed, so accelerating discovery and transfer of knowledge

Users = scientists = we want to follow the Scientific method

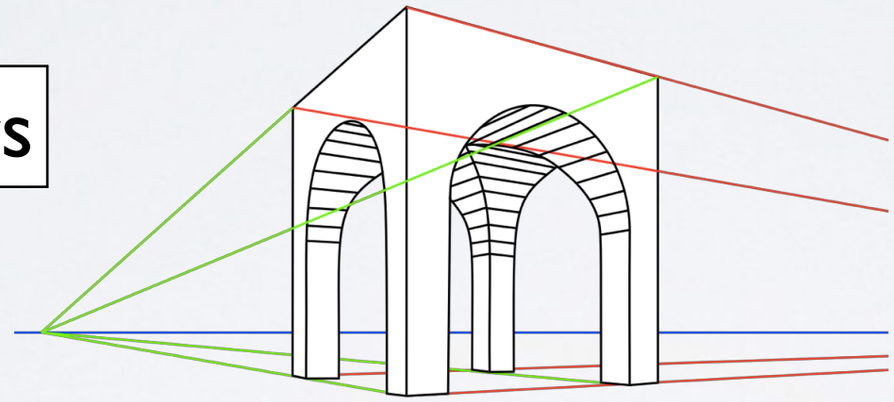
** Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks. Its implementation at the SRCs will facilitate sharing data, methods, resources and tools across the community, enabling verification, reusability and repurpose.

PERSPECTIVES

Implementation of Open, reproducible science is challenging, even more in this new framework:

NEW ROLES → NEW PERSPECTIVES

Individual users



Evaluators/
Funding agencies

Publishers

KSP teams

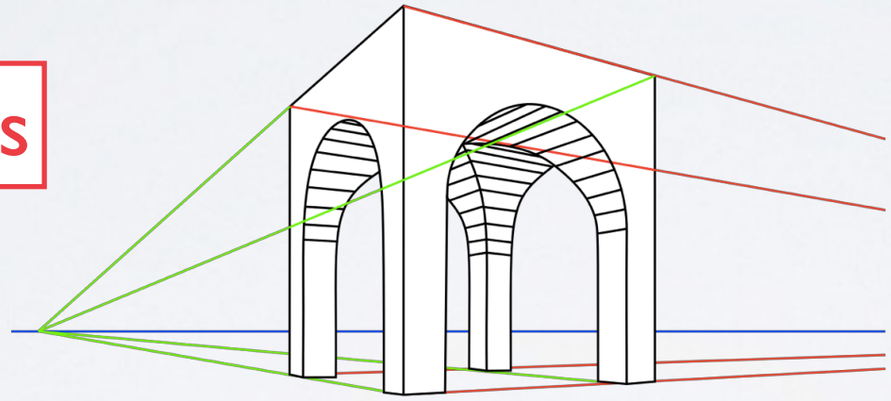
Service
providers

MY MOTIVATION

Implementation of Open, reproducible science is challenging, even more in this new framework:

NEW ROLES → NEW PERSPECTIVES

Individual users



Evaluators/
Funding agencies

Publishers

KSP teams

SKA SWG

Service providers

SKA SQUARE KILOMETRE ARRAY Spain

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

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

ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

SKA SRCCG & SRCSC

A proto-SRC at IAA

SKA Organization

SRC PROTOTYPE AT IAA-CSIC

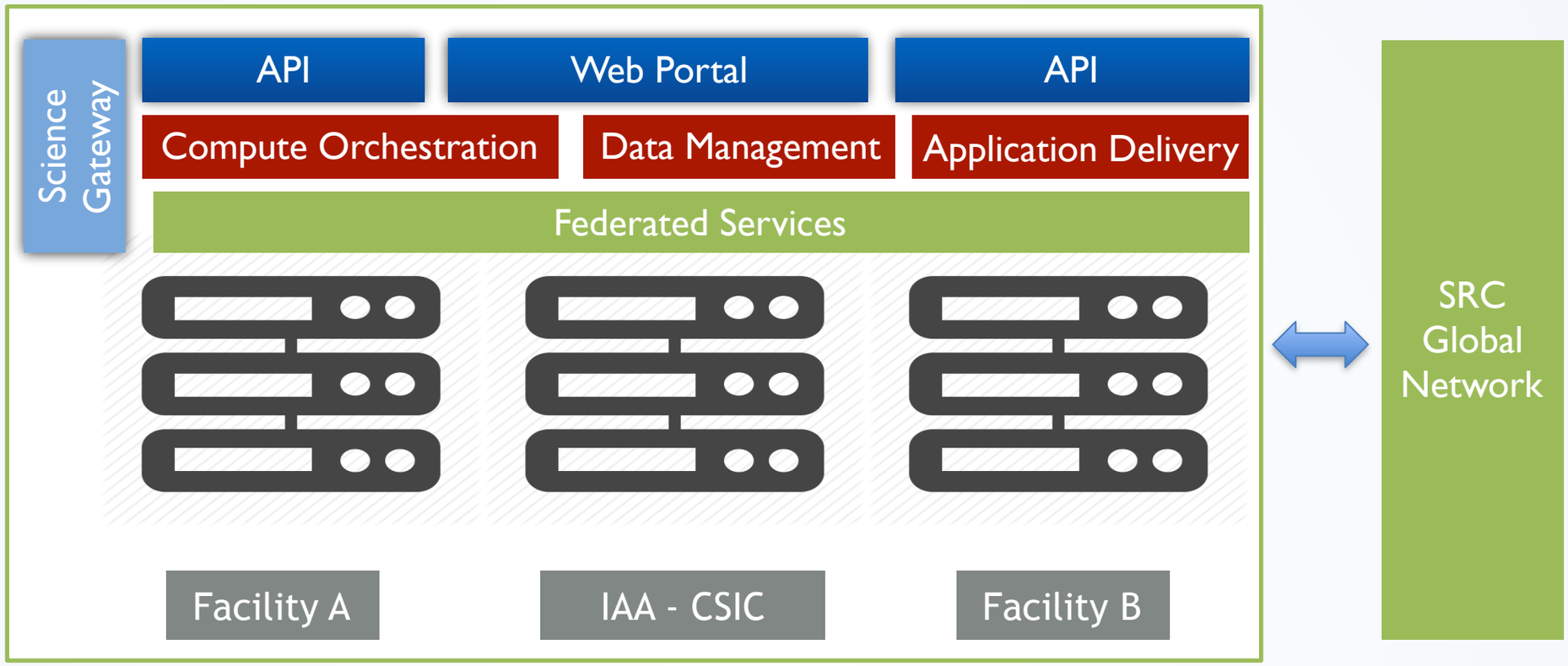
Centre of Excellence "Severo Ochoa"(S.O.) accreditation from the Spanish Ministry of Science that acknowledges the Spanish centres that carry out cutting-edge research.

- A prototype of SKA SRC included in the IAA S. O. scientific programme.

Objectives:

- Support IAA members participation in
  - SKA precursors /pathfinders
  - SKA Data Challenges
- Embrace Open Science Principles: Data-Intensive and Reproducible Research for the SKA Regional Centres
- Partnership with national HPC facilities / experts in computational science
- Collaborate with other SRC initiatives
- Innovation in analysis techniques, new algorithms

SRC PROTOTYPE AT IAA-CSIC



Science Analysis Platform

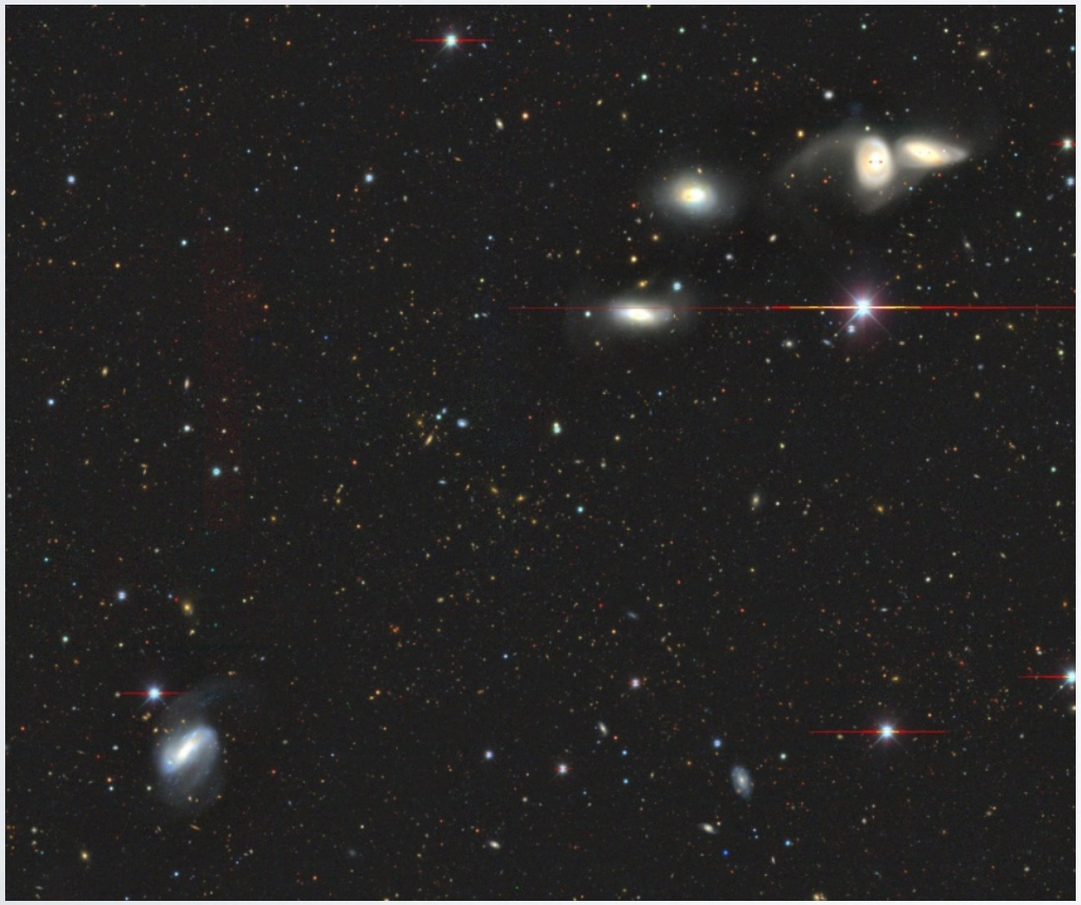
- Identify technical specifications from the use case requirements
- Set-up of the associated IAA computing/storage resources
- Partnership with national computing facilities
- Collaborations with e-Infrastructures and other SRCs prototype initiatives
- Provide a Science Gateway

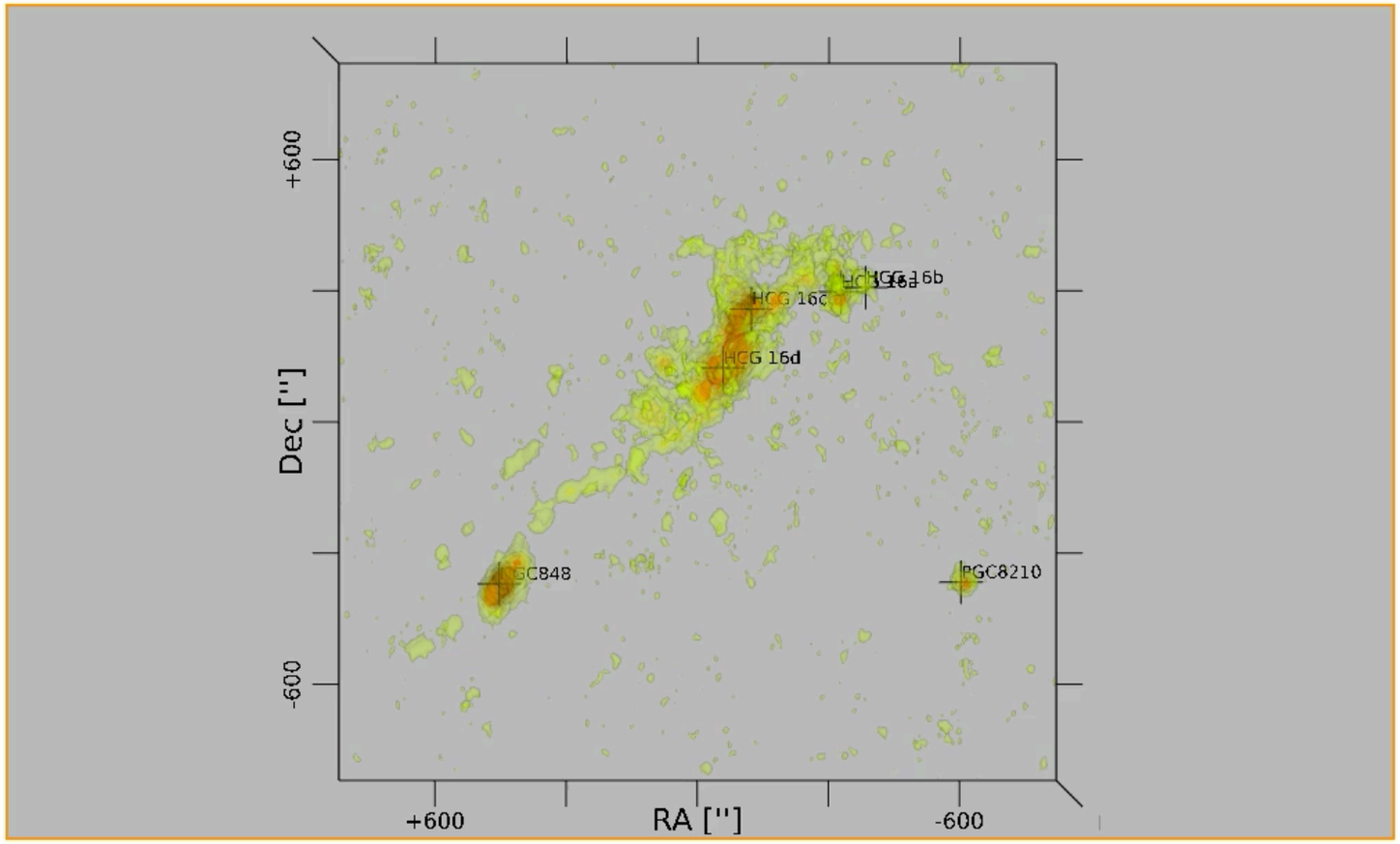
Open Science in a real scientific experiment

http://littlepandabears.blogspot.com/2018/09/galaxies.html
https://searchengineland.com/figz/wp-content/seloads/2014/08/lab-test-experiment-ss-1920.jpg

Specific Example: HI in HCG 16

- HCG 16 is complex compact group with starburst galaxies, AGN, tidal tails, etc. The main goal of this project is to is to study the HI content of the group and to determine which on-going processes are causing it to change.
- **Collaborators:** L. Verdes-Montenegro, A. Damas, S. Borthakur, M. Yun, A. del Olmo, J. Perea, B. Williams, D. Lopez Gutierrez, F. Vogt, S. Luna, J. Román, J. Garrido, S. Sanchez, J. Cannon & P. Ramírez





Viewpoints: Reset View | R.A. - Dec. | V - Dec. | V - R.A. | Perspective View | Next View

HI layers: 12-sigma | 9-sigma | 6-sigma | 3-sigma

BEING FAIR

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:

- **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.

BEING UN-FAIR

Common astronomy examples of un-FAIR practices:

- The **raw** data are in an **archive** but the final, **reduced data** and images are only publicly available in the paper **PDF**.
- The final data are "available", but you **need to request** them by email.
- There are some **scripts** for processing the data on a server somewhere, but no one remembers how to run them.
- The code is on **github**, but good luck trying to install/execute it.

I'm not pointing fingers here, we are all guilty of these things, myself included. We need to improve as a community.

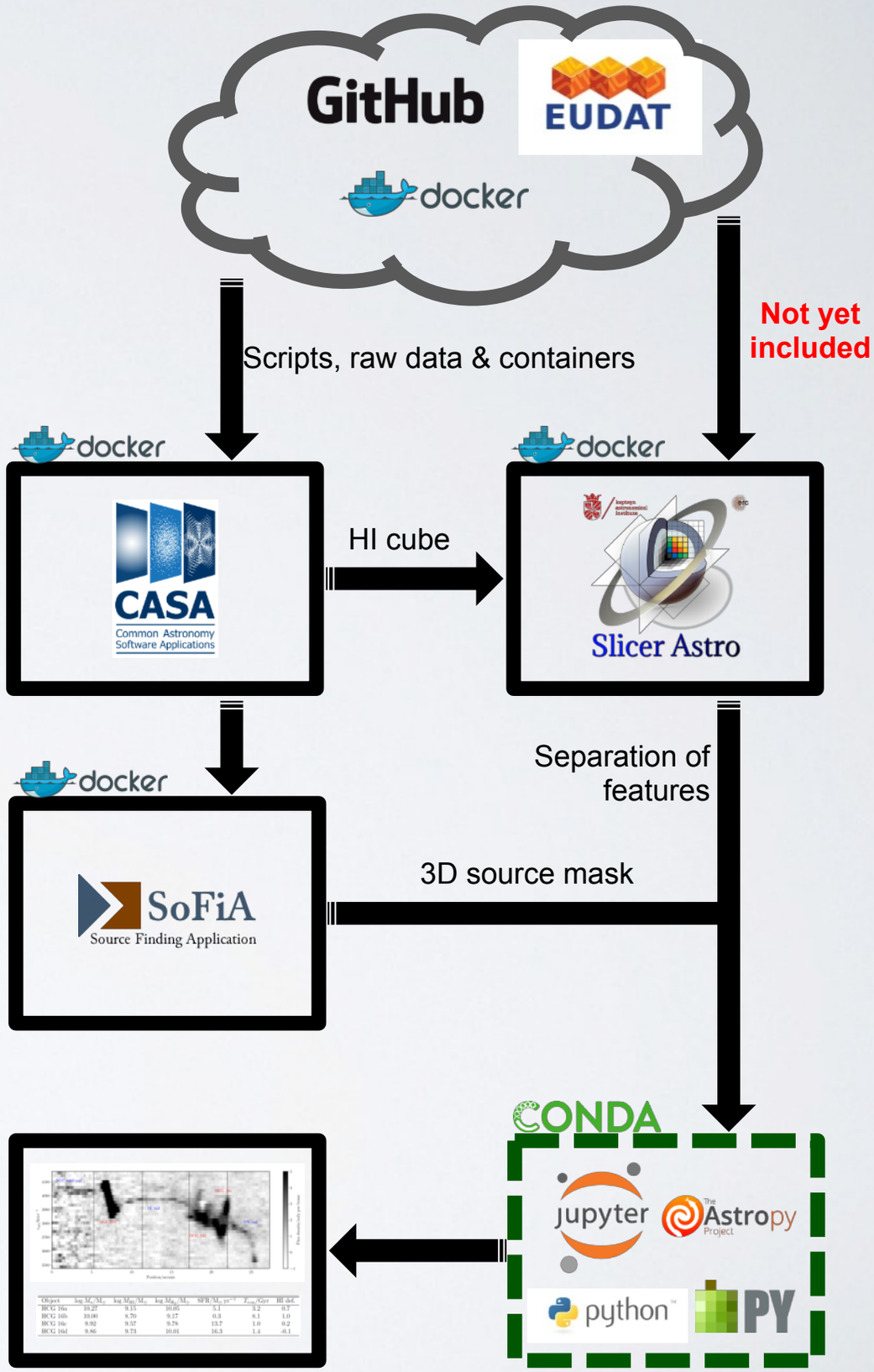
Our Workflow

- FAIR focuses on the **data,** we want to go beyond this and include also the **methods**.

- It is executed entirely within **Docker** containers and **Conda** environments. So it can be run on any platform with Docker and Conda, using a single bash script.

- The code and data are publicly available in **github** and **EUDAT**.

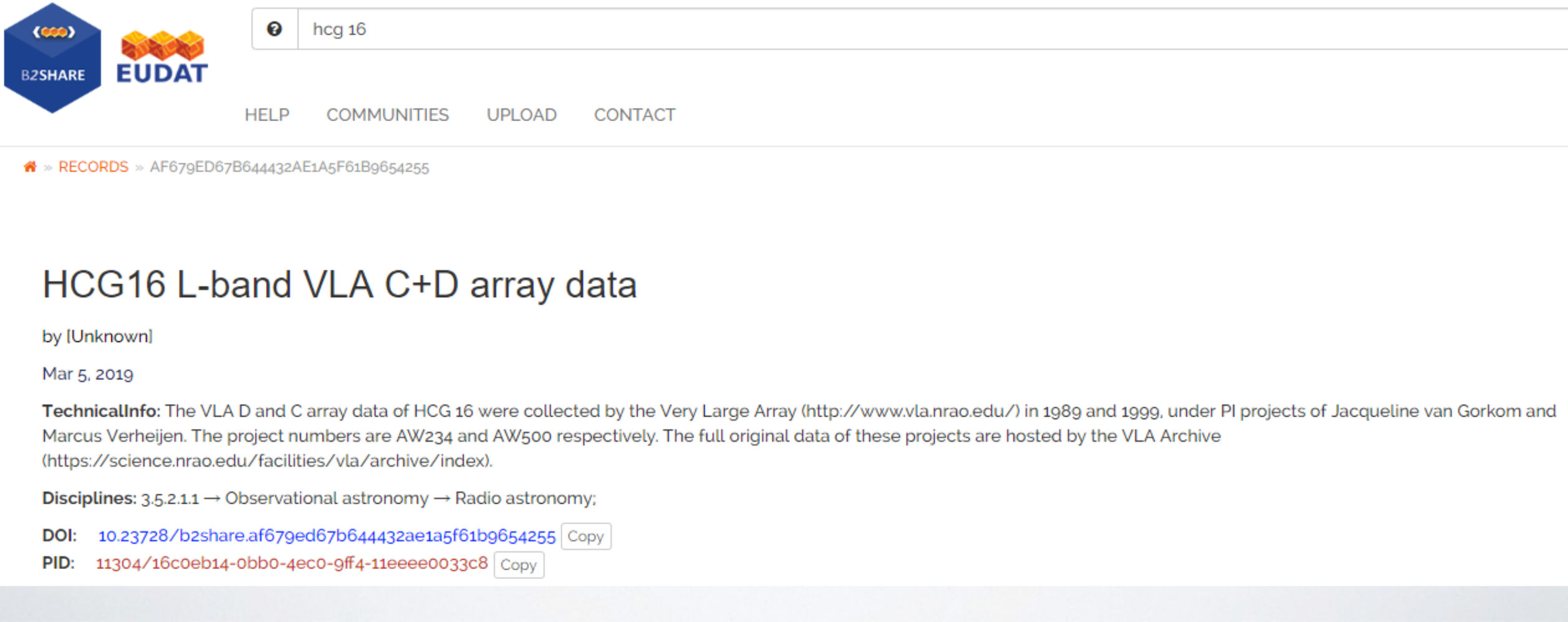
- The workflow can also be executed in EOSC



Our Workflow

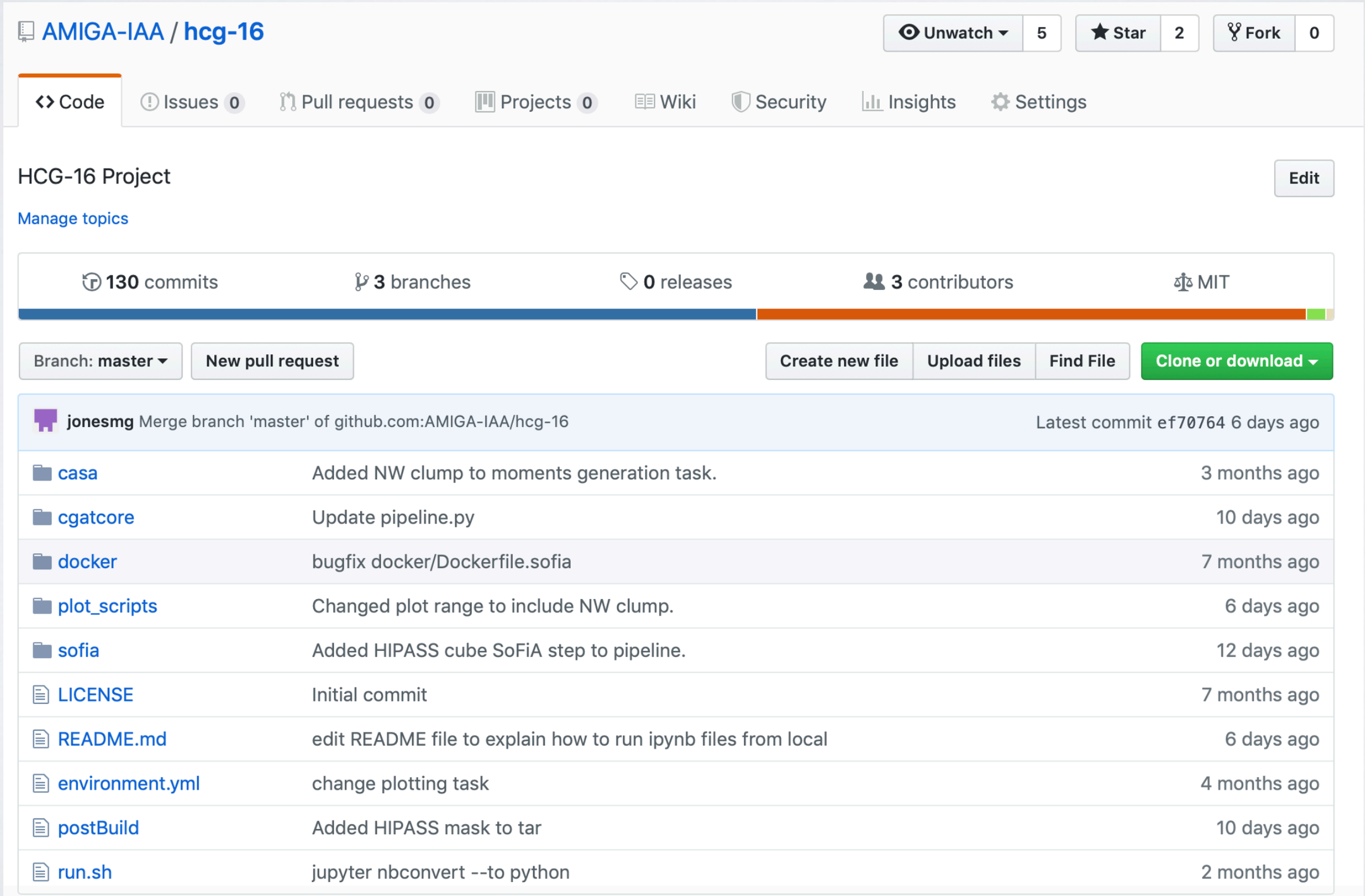
The raw data are hosted on a the EUDAT service, which provides:

- Cloud storage
- Persistent identifiers (DOI)
- Access (can download with wget)
- Basic metadata and search functionality



Our Workflow

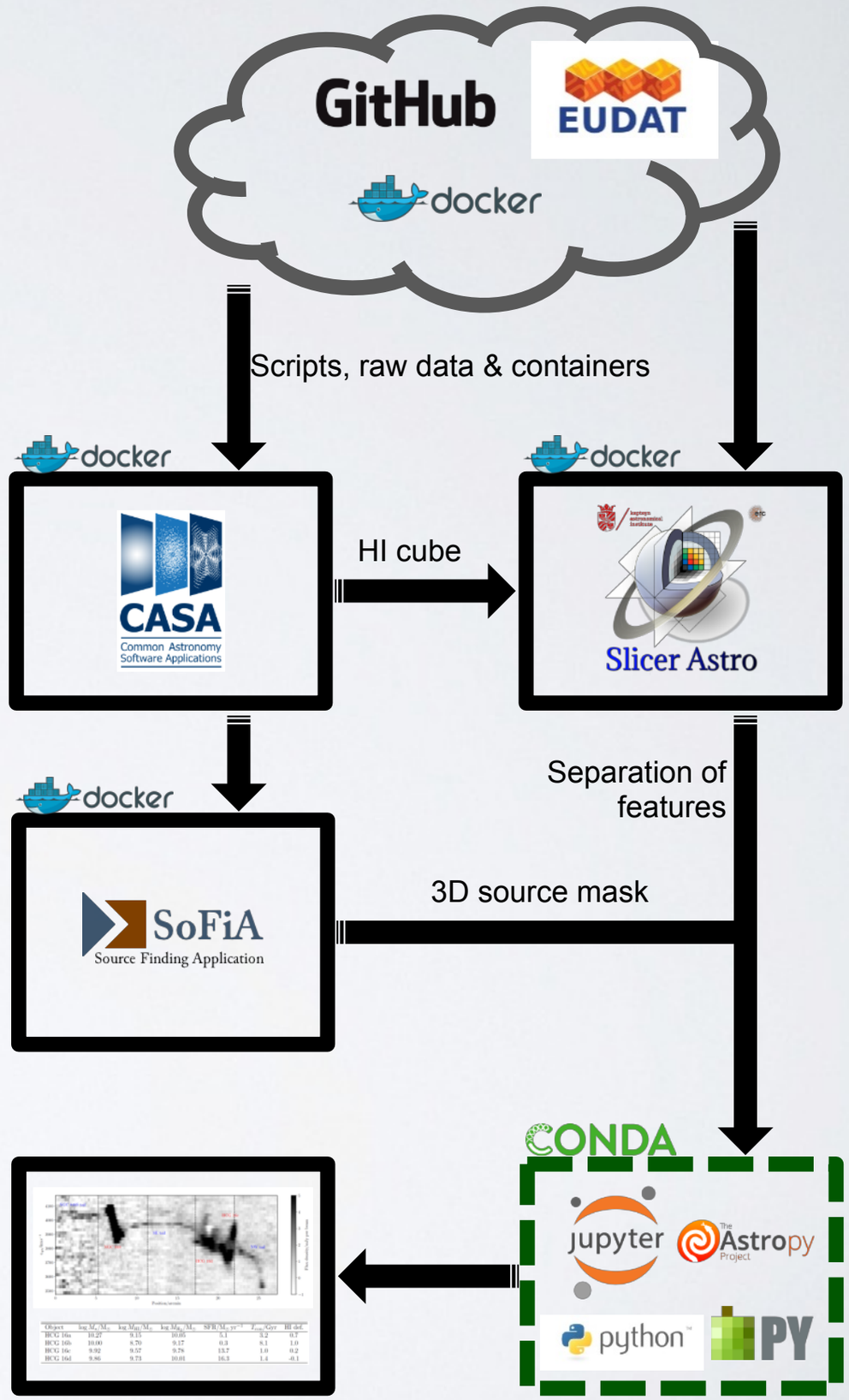
- All the code for the all of the workflow from raw data to final plots is stored in github and is openly accessible.



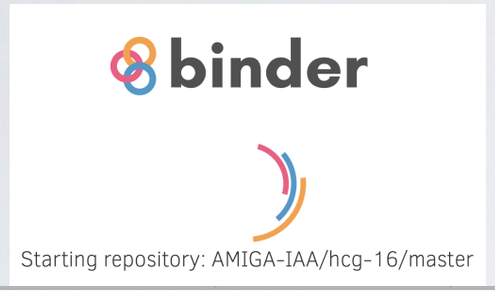
https://github.com/AMIGA-IAA/hcg-16

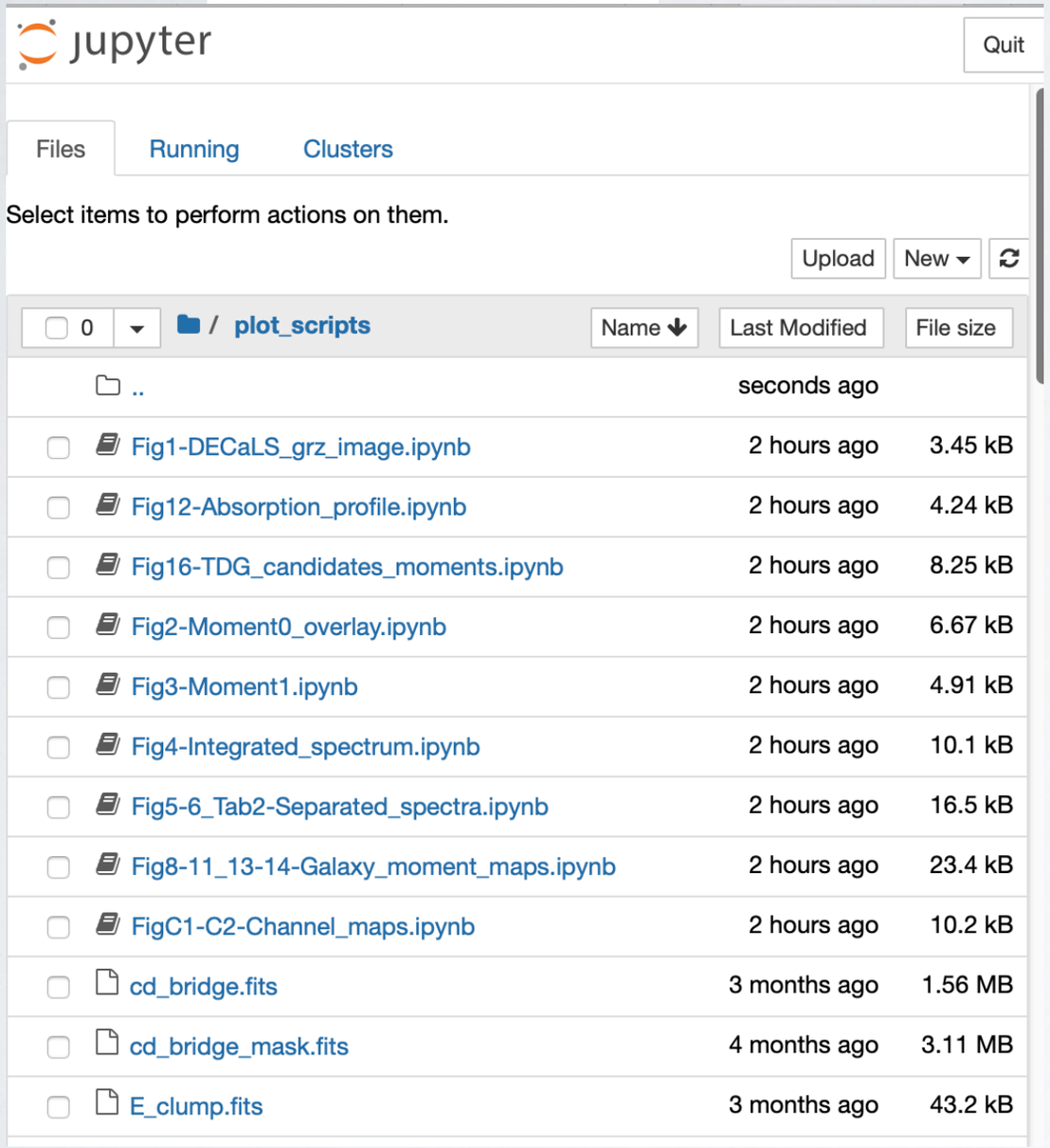
Our Workflow

- run.sh will do automatically the following steps:
  - download and install conda
  - download and install cgatcore, a workflow management system
  - construct a conda python environment with which to run the code
  - download the source code
  - download the input data
  - run the pipeline



Reproducible Figures





https://mybinder.org/v2/gh/AMIGA-IAA/hcg-16/master

Figure 2. HCG 16 HI moment zero map and overlay

In [ ]:
```
import matplotlib,aplpy
from astropy.wcs import WCS
from astropy.io import fits
from general_functions import *
import matplotlib.pyplot as plt
```

In [ ]:
```
font = {'size'   : 14, 'family' : 'serif', 'serif' : 'cm'}
plt.rc('font', **font)
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['lines.linewidth'] = 1
plt.rcParams['axes.linewidth'] = 1

#Set to true to save pdf versions of figures
save_figs = True
```

The files used to make the following plot are:

In [ ]:
```
moment0_casa = 'HCG16_CD_rob2_MS.mom0.pbcor.fits'
moment0_sofia = 'HCG16_CD_rob2_MS_mom0.fits'
r_image_decals = 'HCG16_DECaLS_r_cutout.fits'
```

1. A moment 0 map of HCG 16 generated using a simple 3σ threshold in each channel (made with CASA). This file was generated in the *imaging* step of the workflow, which is described in the script imaging.py.
2. A moment 0 map of HCG 16 generated using 3.5σ mask made with SoFiA after smoothing over various kernel sizes. This file was generated in the *masking* step of the workflow. The SoFiA parameters file which makes this file is HCG16_CD_rob2_MS.3.5s.dil.session.
3. An *r*-band DECaLS fits image of HCG 16. This file was downloaded directly from the DECaLS public website. The exact parameters defining the region and pixel size of this images is contained in the pipeline.yml file.

Take Away

- Define Conda environments or containers and your co-authors will be able to run your code (but not only them).
- Release vs publish:
  - Code: GitHub, bitbucket
  - Papers, Documentation, data, …
  - Zenodo (DOIs,)
- Publish:
  - Papers: pre-prints in arxiv, open repositories and journals
  - Code: GitHub, bitbucket
  - Data: Astronomy archives (e.g. CDS) vs open Repositories
- CV and career assessment:
  - Your CV can be something more than a list of papers.
  - Altmetrics as supplement (not a replacement) to highlight research products that might otherwise go unnoticed

CONCLUSIONS

- SKA will be a game changer in the way we do science
  - Large international teams
  - Don't move the data
  - New methods to share computational resources
  - Sharing the tools (reinventing not affordable), doing reproducible science
  - Reproducibility in the Big Data era: Data providers moving to service providers

CONCLUSIONS

- SKA will be a game changer in the way we do science
  - Large international teams
  - Don't move the data
  - New methods to share computational resources
  - Sharing the tools (reinventing not affordable), doing reproducible science
- Reproducibility in the Big Data era: Data providers moving to service providers

Will we forget about reproducibility since we need to "**efficiently**" exploit large datasets?

@tetrarquis