ALMA science data management

Felix Stoehr
ALMA Science Archive

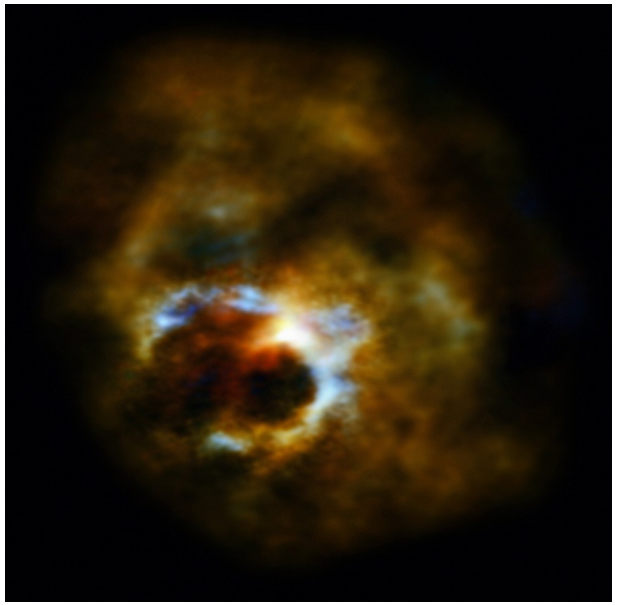
credit goes to the entire ALMA team

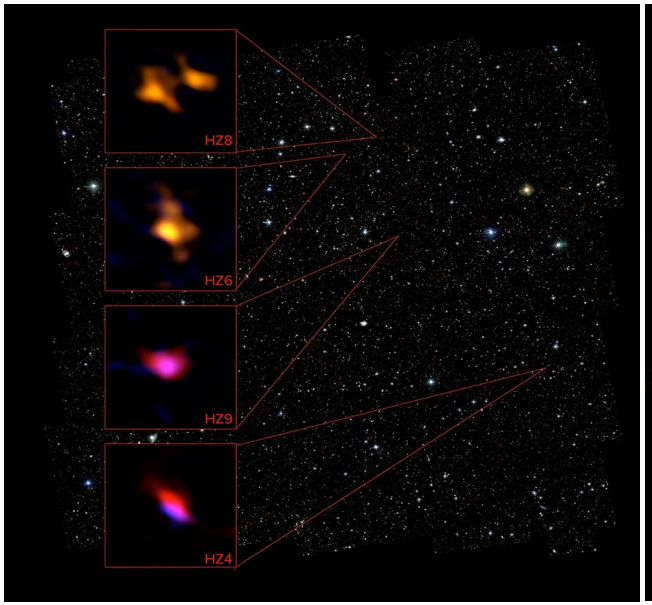
ALMA

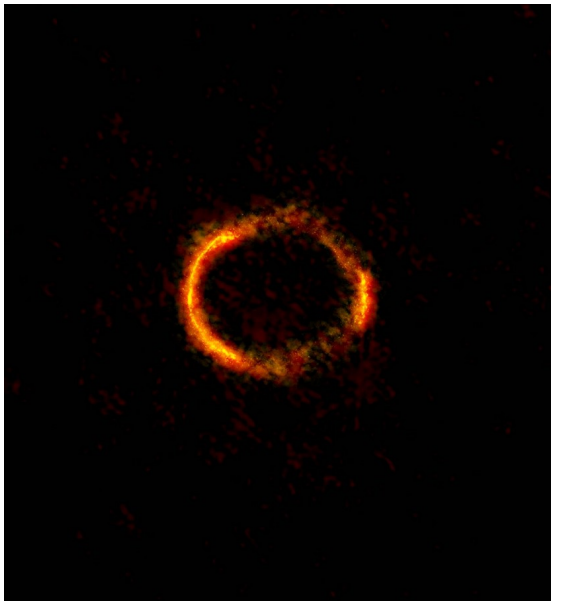
C. Malin

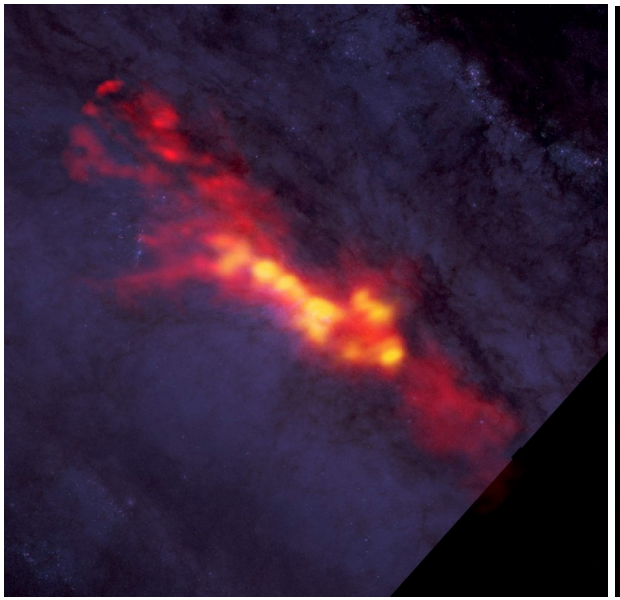
- built by ESO, NRAO, NAOJ in cooperation with Chile
- 66 antennas at 5000m elevation in the Atacama Desert
- interferometry at 84-702 GHz with 16km baselines
- full operations: 200TB/yr=6.6Mbytes/s
- Cycle 3: observing, call for Cycle 4: spring 2016

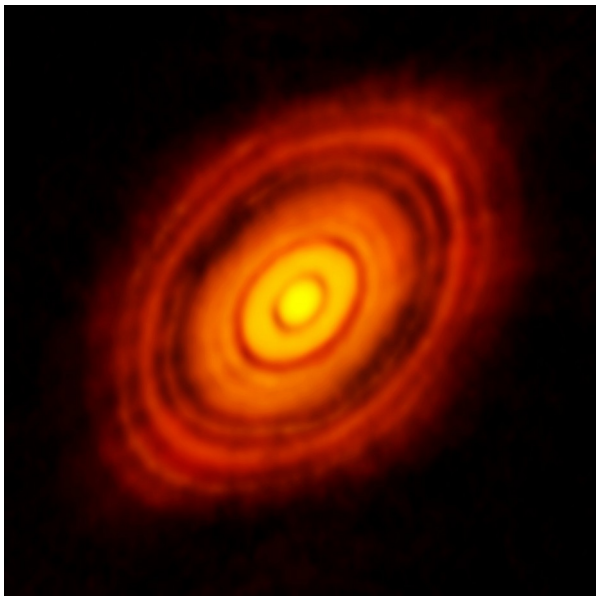
ALMA









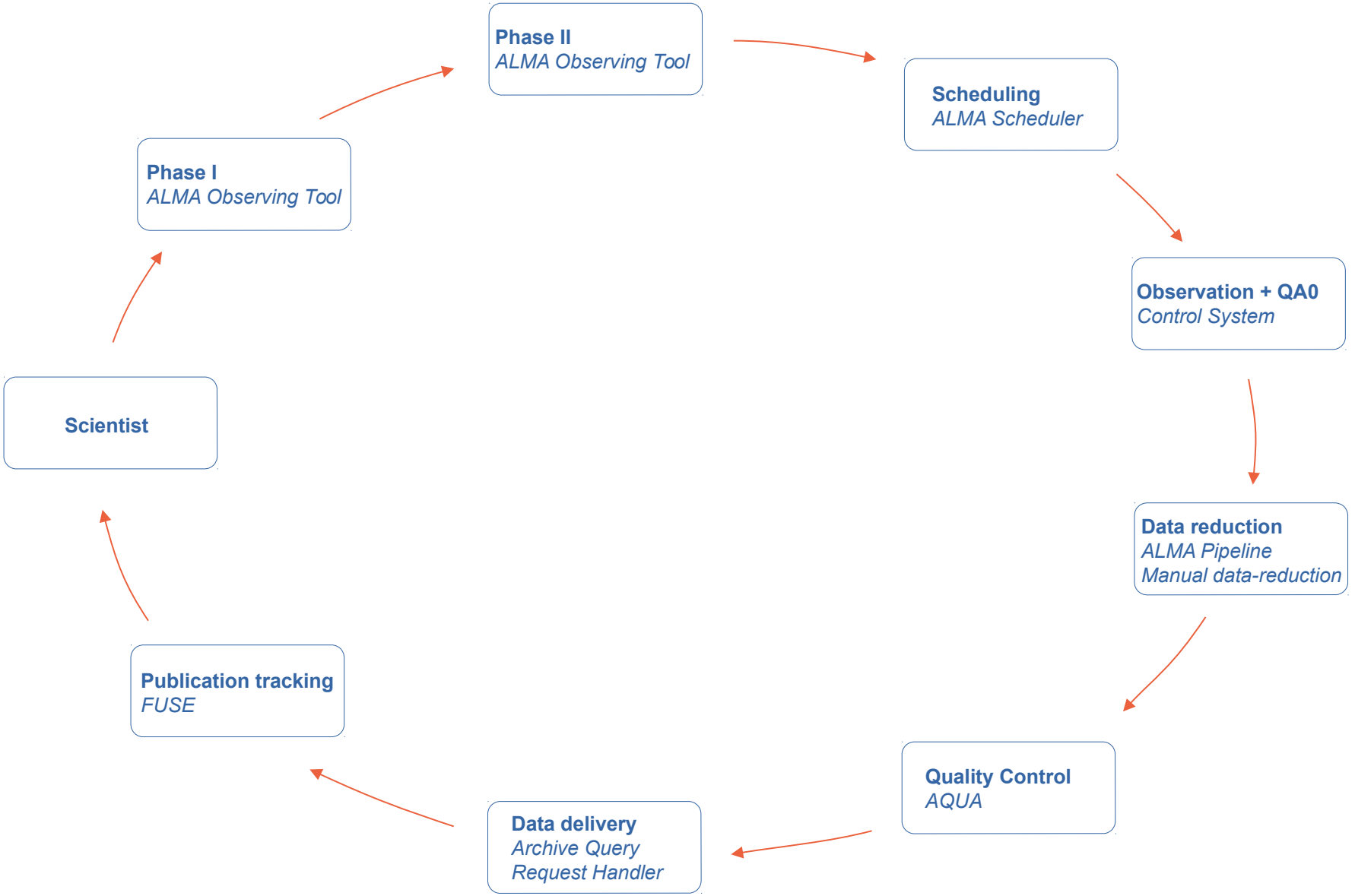


ALMA

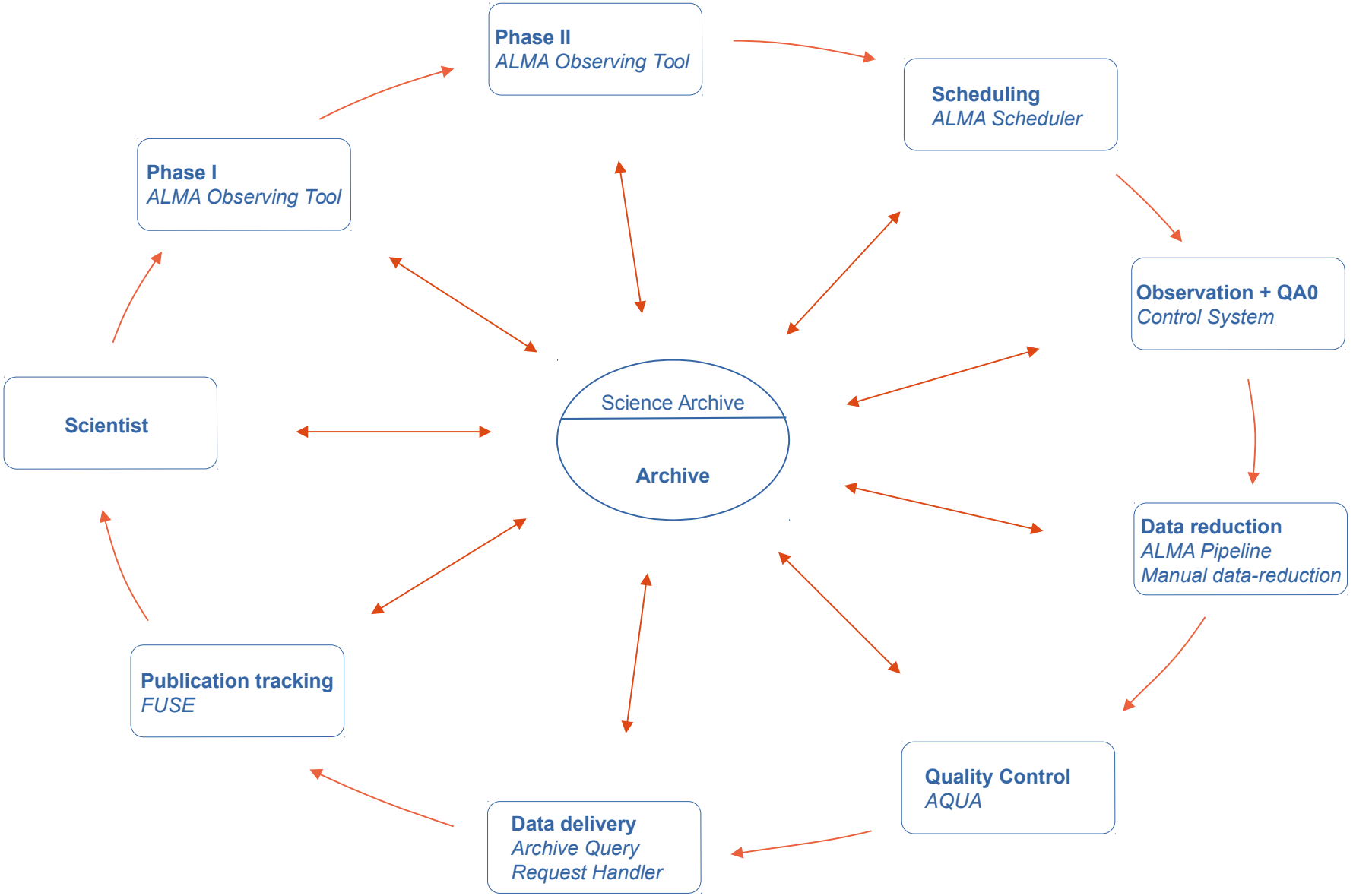
context

- demands of astronomers **scale faster than funding**
- collaboration over **continents** is required for the largest projects
- **challenges**
  - John Hibbard: Who does train the next generation of experts who have hands-on end-to-end expertise? Maybe there is not only room for smaller telescopes but astronomy needs those to train the next generation?
  - communication/organization

workflow


Phase II
ALMA Observing Tool
Scheduling
ALMA Scheduler
Observation + QA0
Control System
Data reduction
ALMA Pipeline
Manual data-reduction
Quality Control
AQUA
Data delivery
Archive Query
Request Handler
Publication tracking
FUSE
Scientist
Phase I
ALMA Observing Tool

workflow


Phase II
ALMA Observing Tool
Scheduling
ALMA Scheduler
Phase I
ALMA Observing Tool
Observation + QA0
Control System
Scientist
Science Archive
Archive
Data reduction
ALMA Pipeline
Manual data-reduction
Publication tracking
FUSE
Quality Control
AQUA
Data delivery
Archive Query
Request Handler

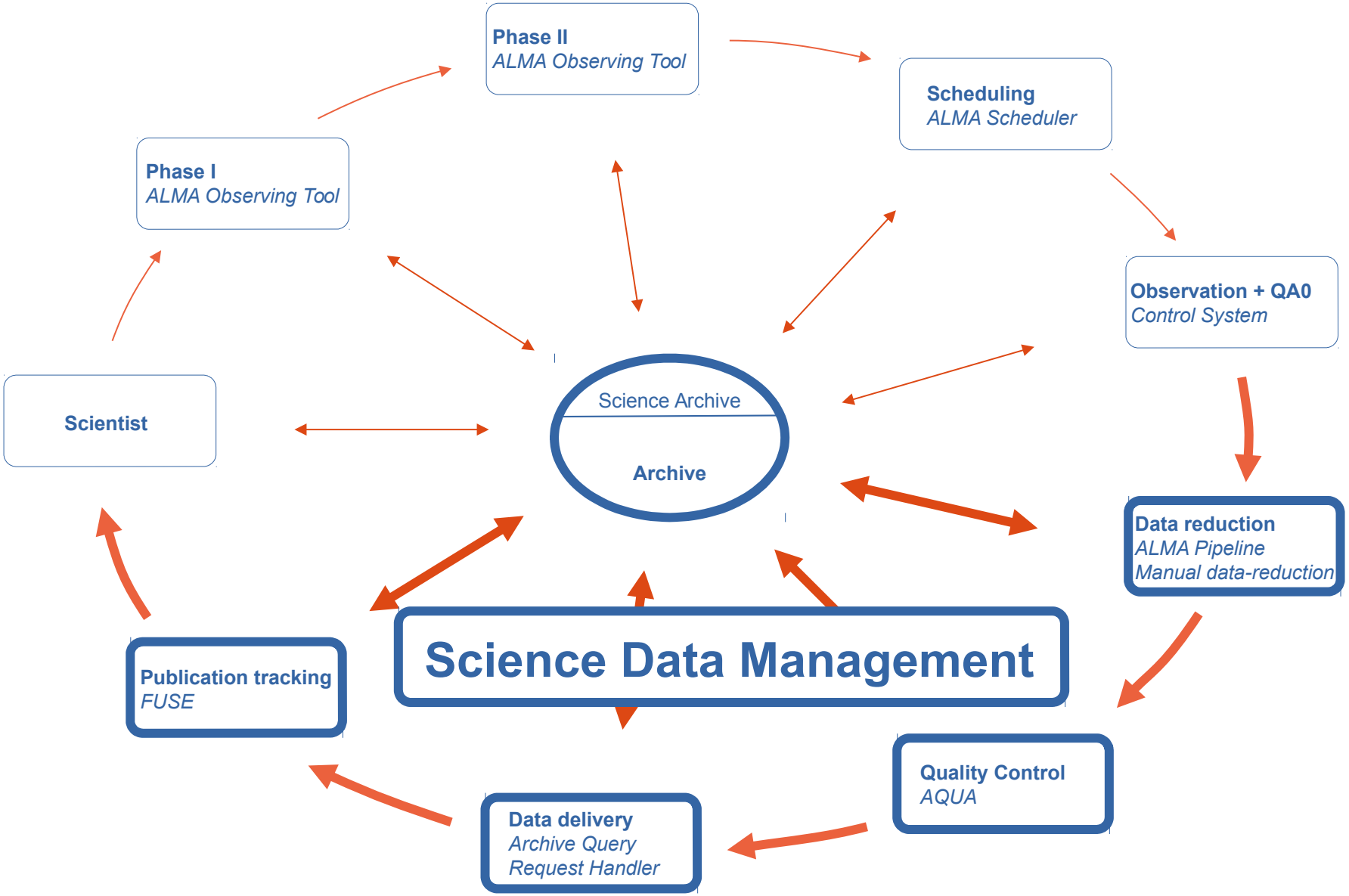
workflow


Phase II
ALMA Observing Tool
Scheduling
ALMA Scheduler
Phase I
ALMA Observing Tool
Observation + QA0
Control System
Scientist
Science Archive
Archive
XML
Data reduction
ALMA Pipeline
Manual data-reduction
Publication tracking
FUSE
Quality Control
AQUA
Data delivery
Archive Query
Request Handler

workflow


Phase II
ALMA Observing Tool
Scheduling
ALMA Scheduler
Phase I
ALMA Observing Tool
Observation + QA0
Control System
Scientist
Science Archive
Archive
JUST SAY NO!
Data reduction
ALMA Pipeline
Manual data-reduction
Publication tracking
FUSE
Quality Control
AQUA
Data delivery
Archive Query
Request Handler

science data management

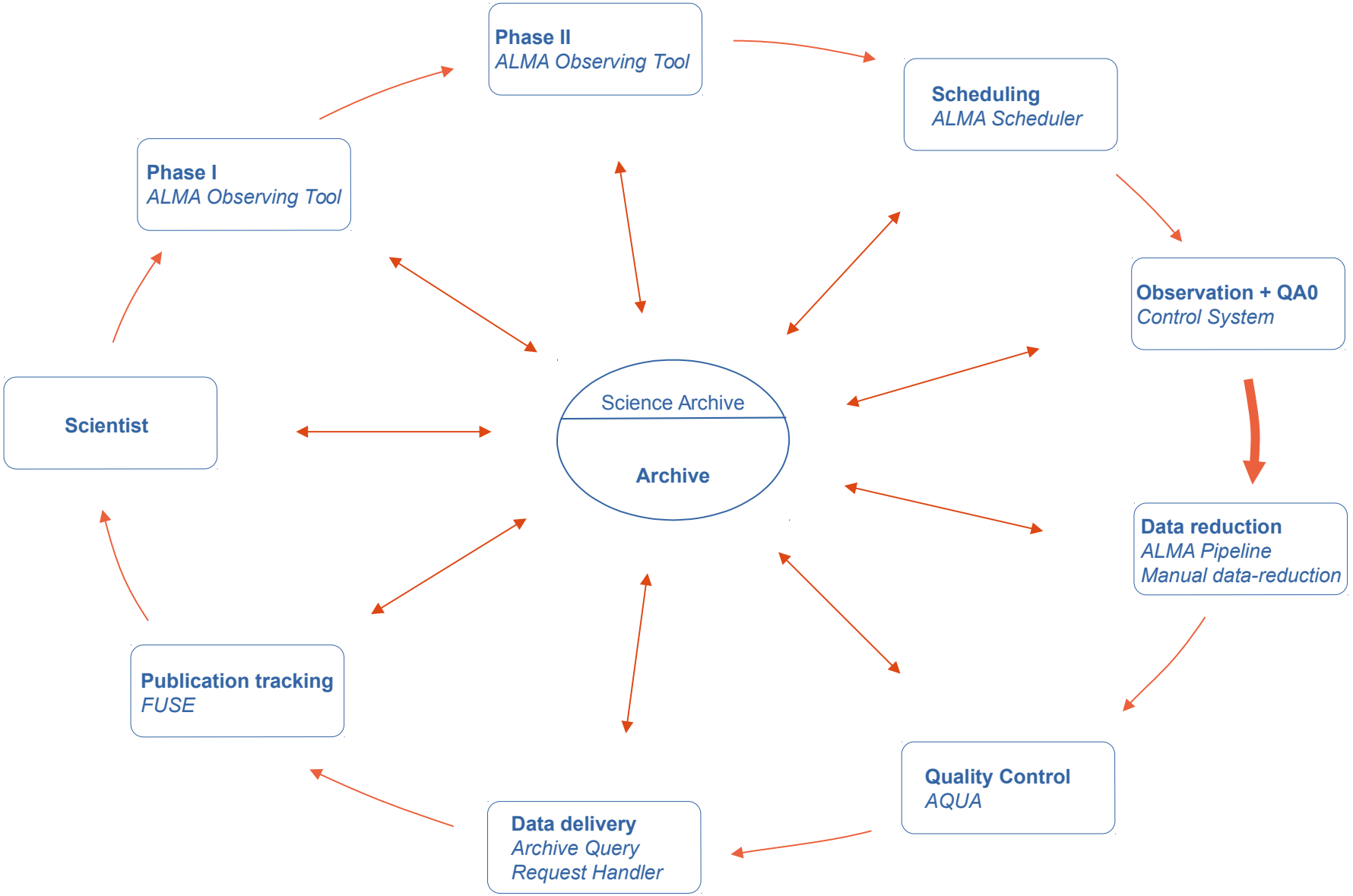


success

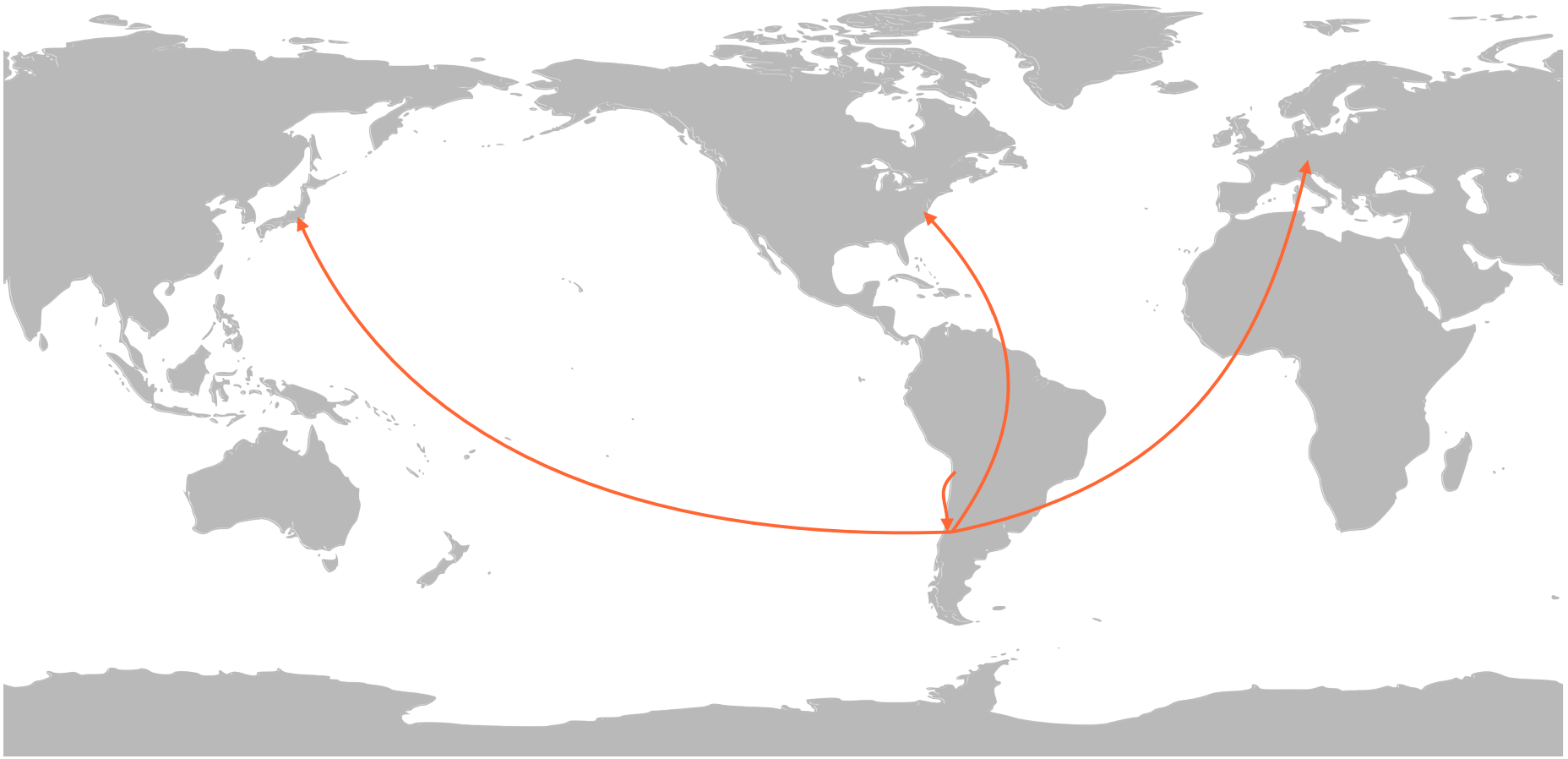
context

- **success = science output of the community**
- **someone else is responsible** for our success
- we have to create the best **end-to-end user-experience possible**

dataflow


Phase II
ALMA Observing Tool
Scheduling
ALMA Scheduler
Phase I
ALMA Observing Tool
Observation + QA0
Control System
Scientist
Science Archive
Archive
Data reduction
ALMA Pipeline
Manual data-reduction
Publication tracking
FUSE
Quality Control
AQUA
Data delivery
Archive Query
Request Handler

dataflow



dataflow

context

- ALMA will produce about the same amount of data in one year as ESO has produced in its first 50 years. (And ESO will, too!)
- LOFAR, MWA, Gaia, PanStars, LSST, SKA, Euclid, ELTs

- T. Tyson: “Astronomy is **transformed** from being a data-starved science to one where data is overabundant”
- multi-wavelength science: **less time** per wavelength regime
- **astronomers do not scale**: bytes/astronomer grow exponentially

- my prediction:
  - now: astronomers **compete for observing time**
  - future: observatories will **compete for astronomers** to work with their data

dataflow

context

- technical evolution will be ahead of astronomy
  - networks, storage, CPU
- however:
  - transfers, storage, processing will need to be **parallel**
- **code-to-data** will become more and more prominent

data reduction

data reduction

- science-grade data
- pipeline
- manual data-reduction
- first science analysis

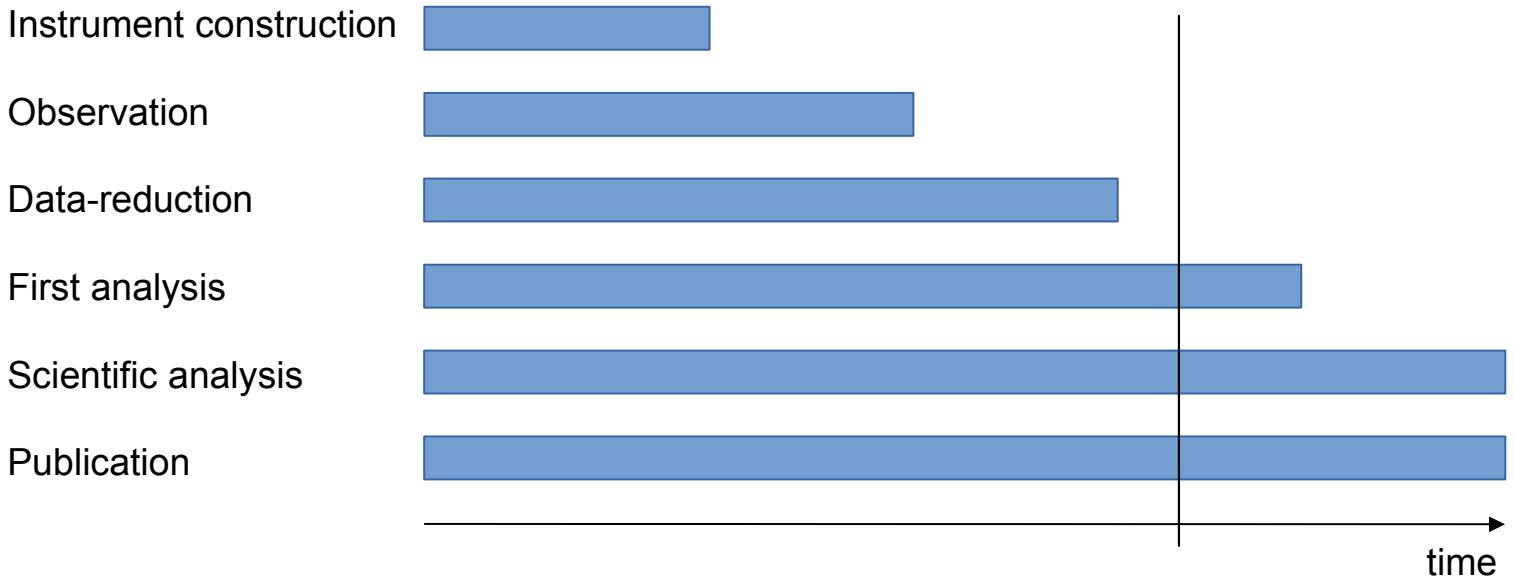
science-grade data

- ALMA policy: will provide **science grade products** for all raw data
- we develop a full **data-reduction package** CASA (python and C++)
- **extreme usersupporting**: face-to-face support from proposal preparation to science analysis in three contintents

science-grade data

context

- what do PIs do?



- workload is **shifting from PIs to observatories**
- more complex telescopes and more data: **pipelines and science-grade data-products** will become an integral part of the telescope/instrument design

science-grade data

context

- deep impact on how science is done
- astronomers will become consumers of data instead of being co-producers
  - good: astronomy makes best use of astronomers, its future rare resource. → The science output will increase.
  - risk: astronomers may not understand data limitations
- observatories will have a much **larger responsibility**
- in some areas: **data-mining** will gain importance
- it will be (even) easier to have an **archival career**

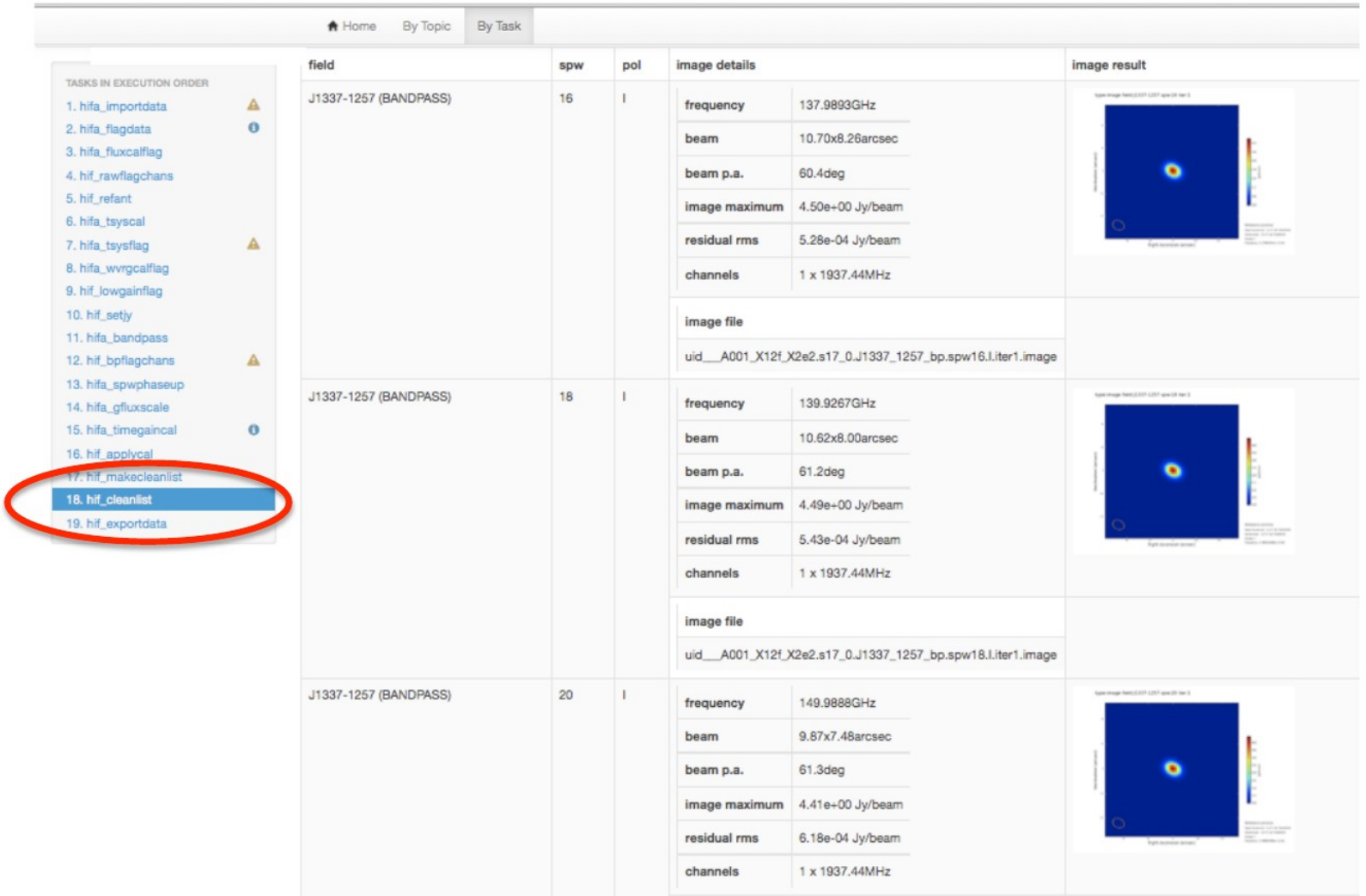
pipeline

- ALMA pipeline is **deployed with CASA**
- is **data-driven** and organized into **tasks**
- status: **calibration has been accepted**, imaging is in commissioning

- **reprocessing**? under discussion
- challenge: **backwards compatibility** of python code

WebLog: Calibrator Images

slides courtesy of Liz Humphreys



pipeline

context

- importance of pipelines for the success of a facility can not be overestimated

- pipeline as **integral part** of the design of an instrument

- challenges:
  - maintenance and **evolution** of the pipelines
  - reprocessing
  - Most of all:

    cost **very easily underestimated**

manual data-reduction

- ALMA's **deliver science-grade data products** principle is kept at all cost
- manual data-reduction JAO and the three ARCs until pipeline is ready
- some modes might need manual reduction also in the future
- benefits
  - helps train expert ALMA staff
  - gained knowledge can be used to improve the pipeline

manual data-reduction

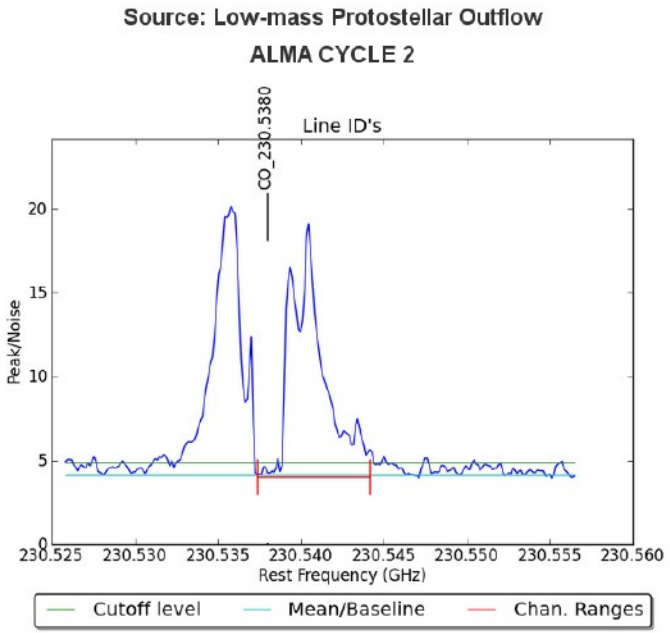
context

- very high cost (but also high quality)
- products can be tailored to the needs of the PI
- not complete/homogeneous products
- reprocessing is essentially infeasible

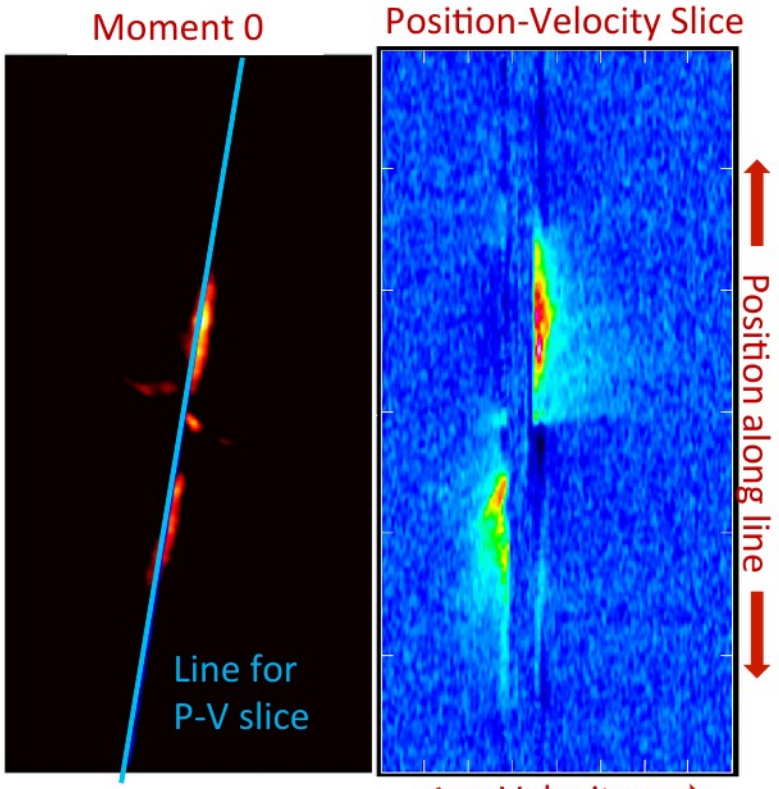
first data analysis

- ADMIT (ALMA data mining toolkit) development program
- CASA add-on, will run at JAO on all products

plots courtesy of ADMIT team



Spectrum: peak emission over noise per channel



Identified lines

frequency	formula	transition	velocity	fwhm	startchan	endchan
230.53800	CO	2-1	1.200E+01	3.503E+00	100	155

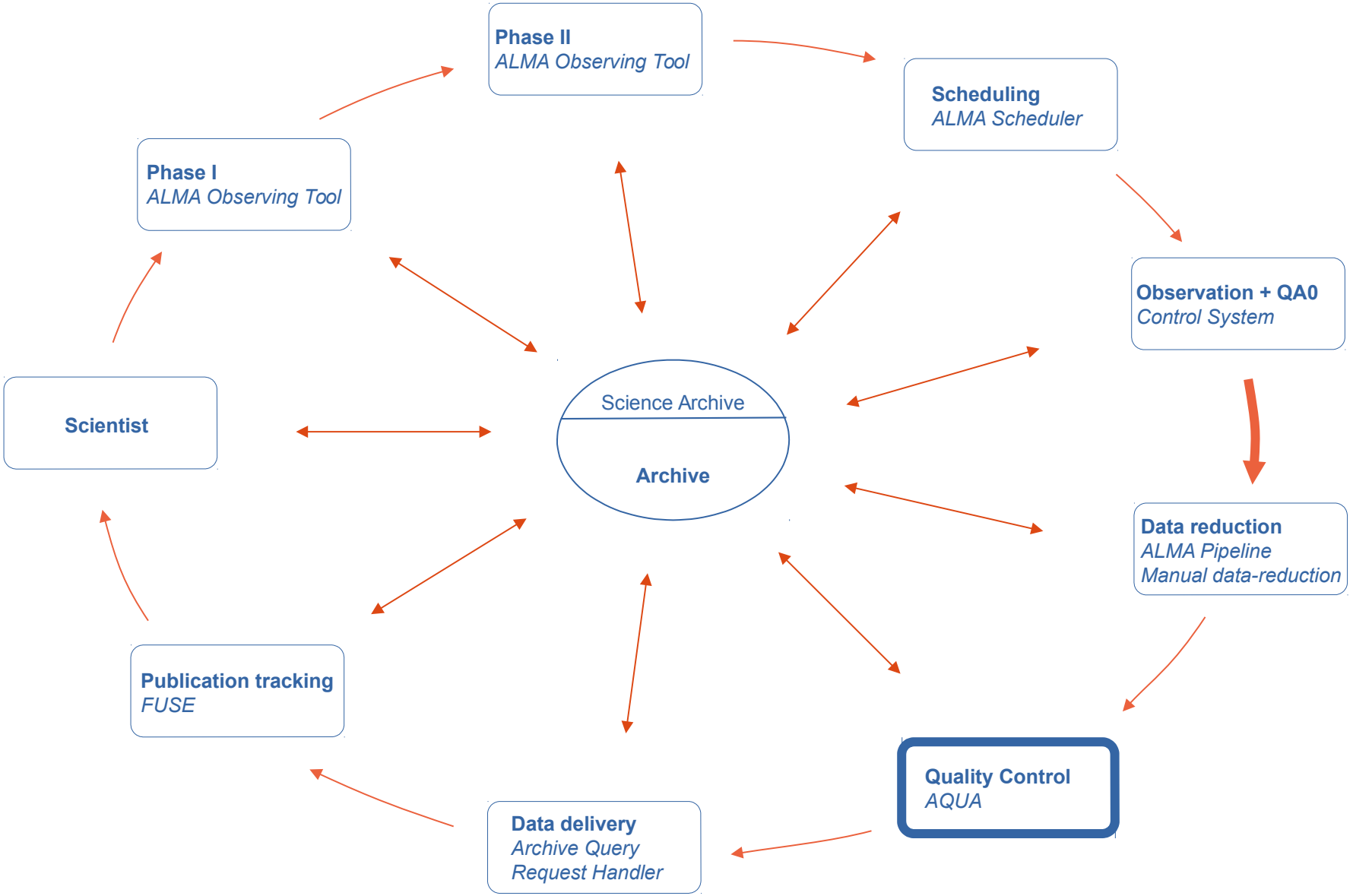
first data analysis

context

- so far most facilities describe their **observations**
- next frontier: first data-analysis allows to describe the **science content** of those observations

- PI ground-based facilities are “following” space facilities and surveys
- again: more workload and responsibility shift from scientists to observatories

quality control

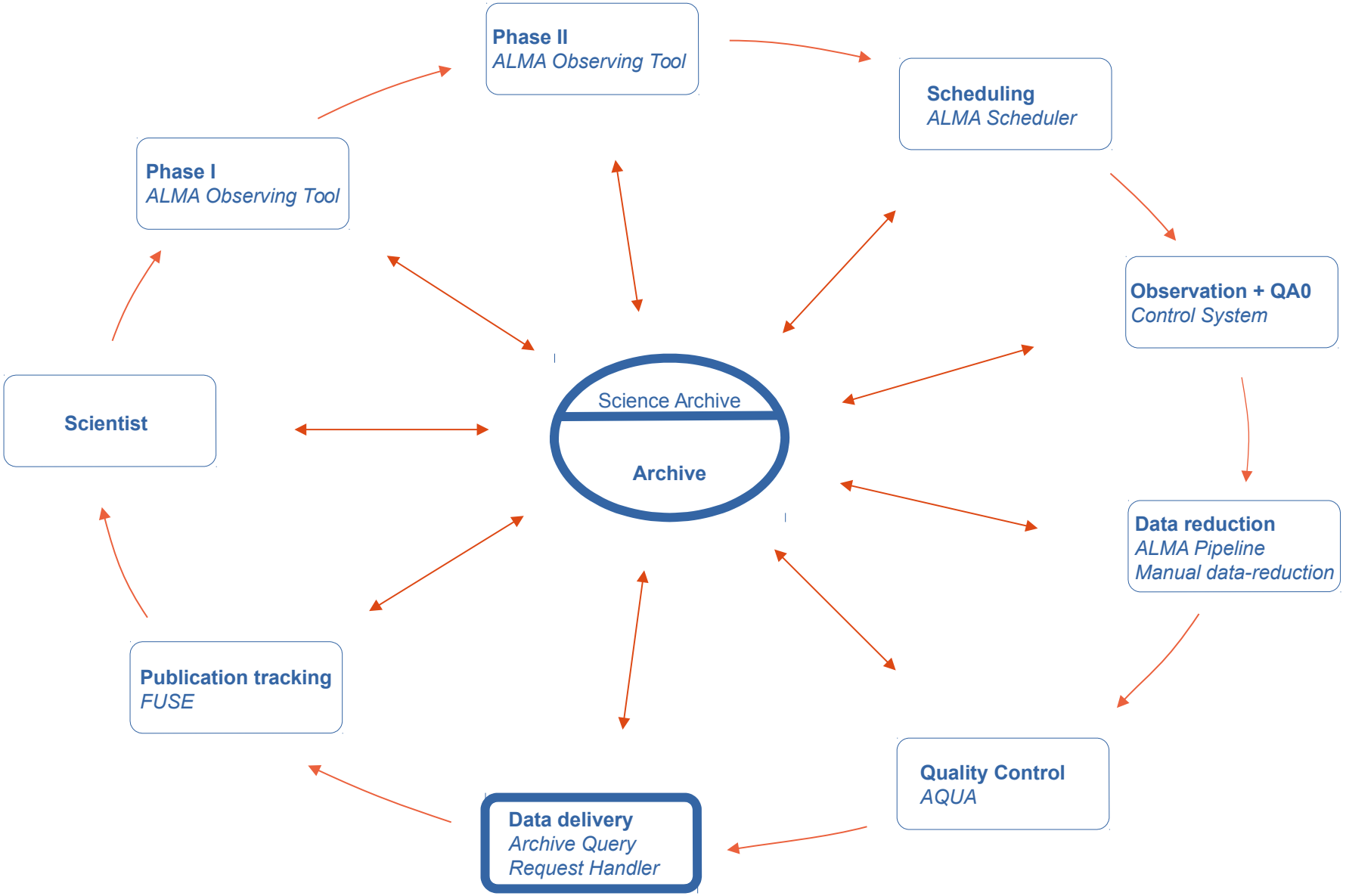


quality control

- ALMA quality assessment
  - QA0: raw data quality assessment after one observation
  - QA1: trend analysis of the facility
  - QA2: data product quality assessment for one dataset using pipeline weblog
  - QA3: analysis of problems reported by users after data delivery

- QA0 and QA1are expected to be automatized
- QA2 and QA3 will remain tasks of a real person

science archive



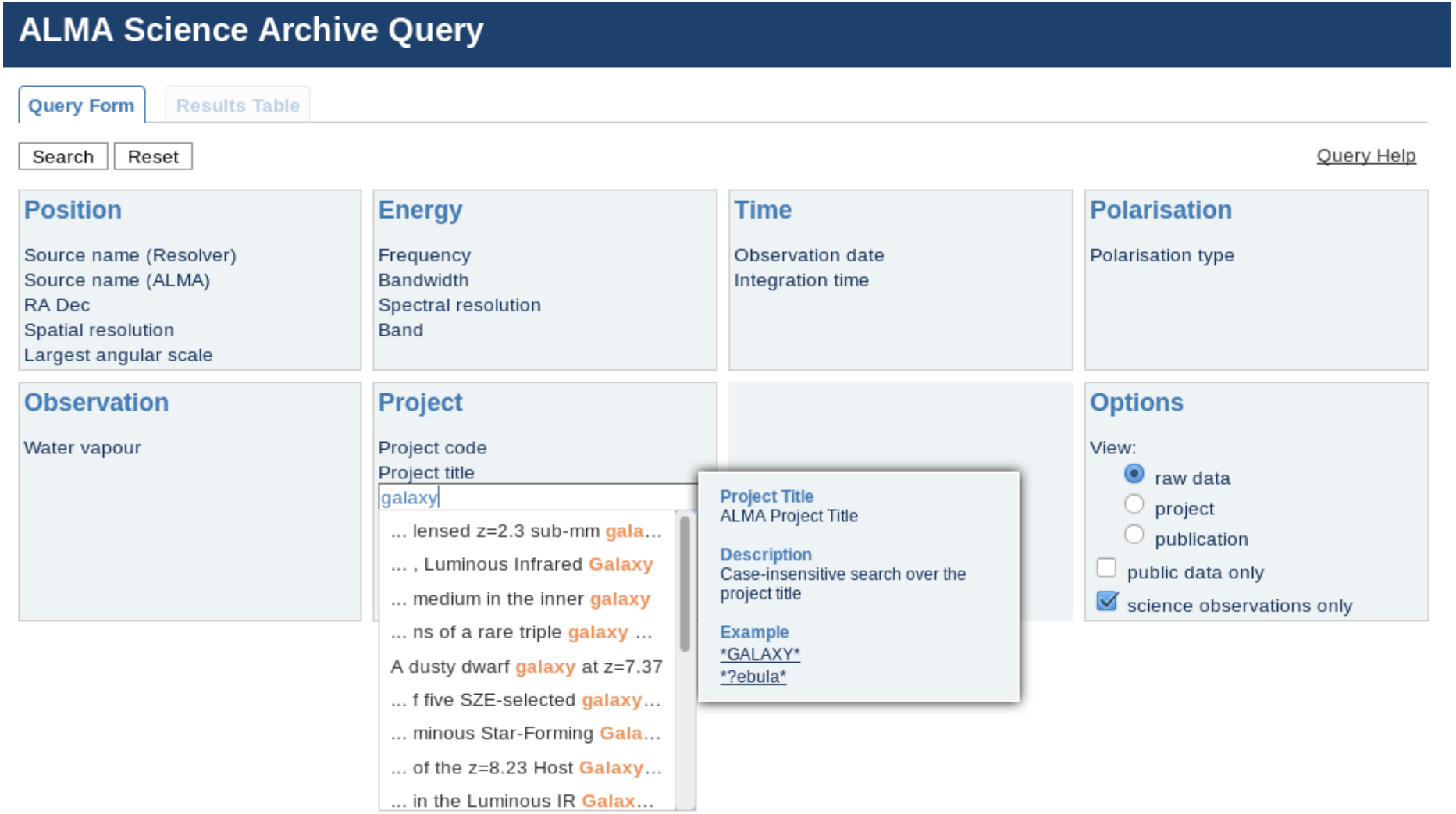
science archive

- complete, correct, consistent, (homogenized)
- speak the **language of the astronomers.** Query by **physical concepts**

> **DANIEL DURAND TEST** of science archives:
> You know have one, when you can find your own data
> by only giving physical parameter constraints

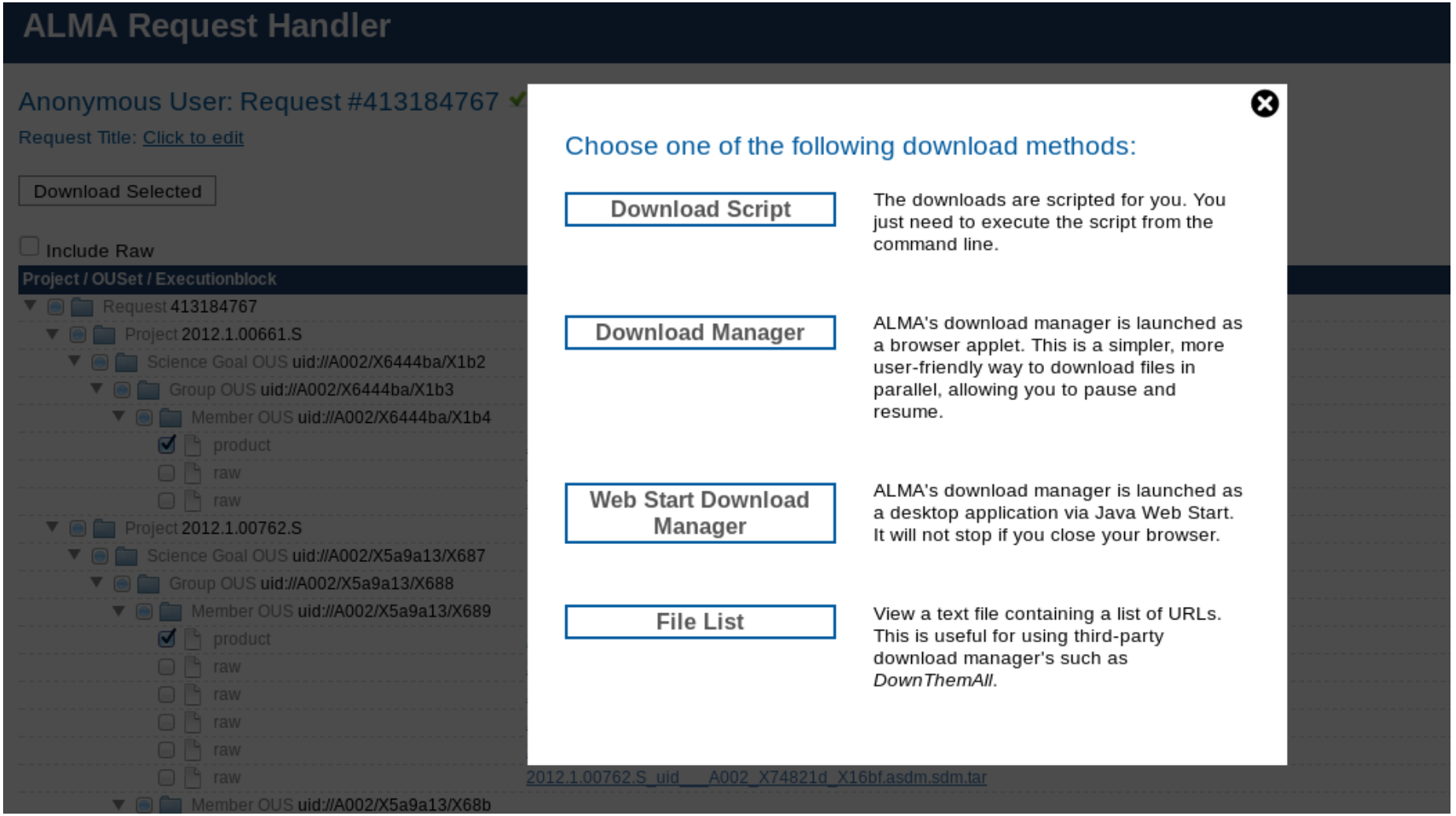
- reduce **interaction cost**
- unscope search: give access to the full **n-dimensional parameter space**

science archive



- future: products, CARTA, PPI, HiPS+AladinLite, VO (ObsCore, SIAPv2, TAP)

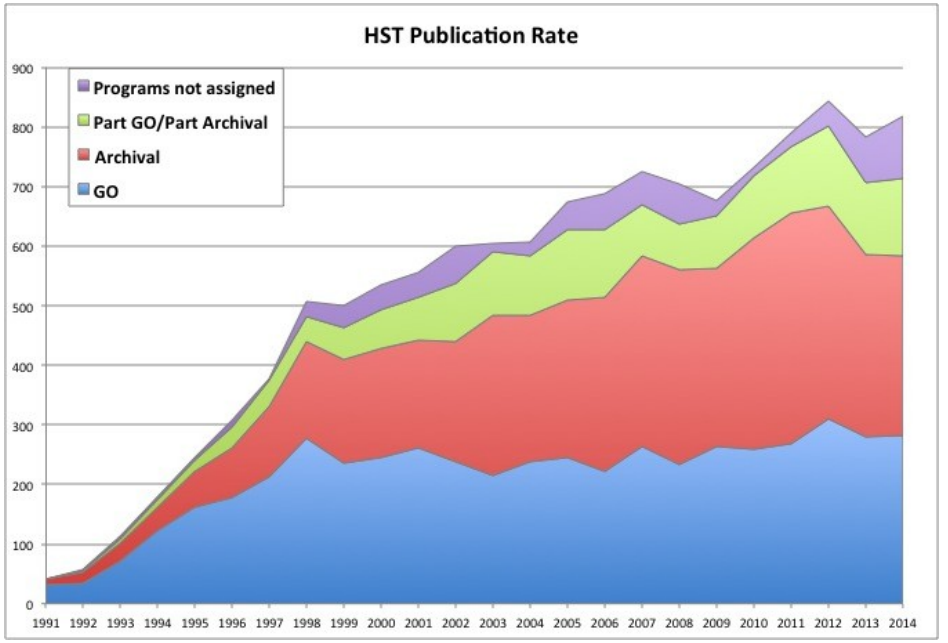
science archive



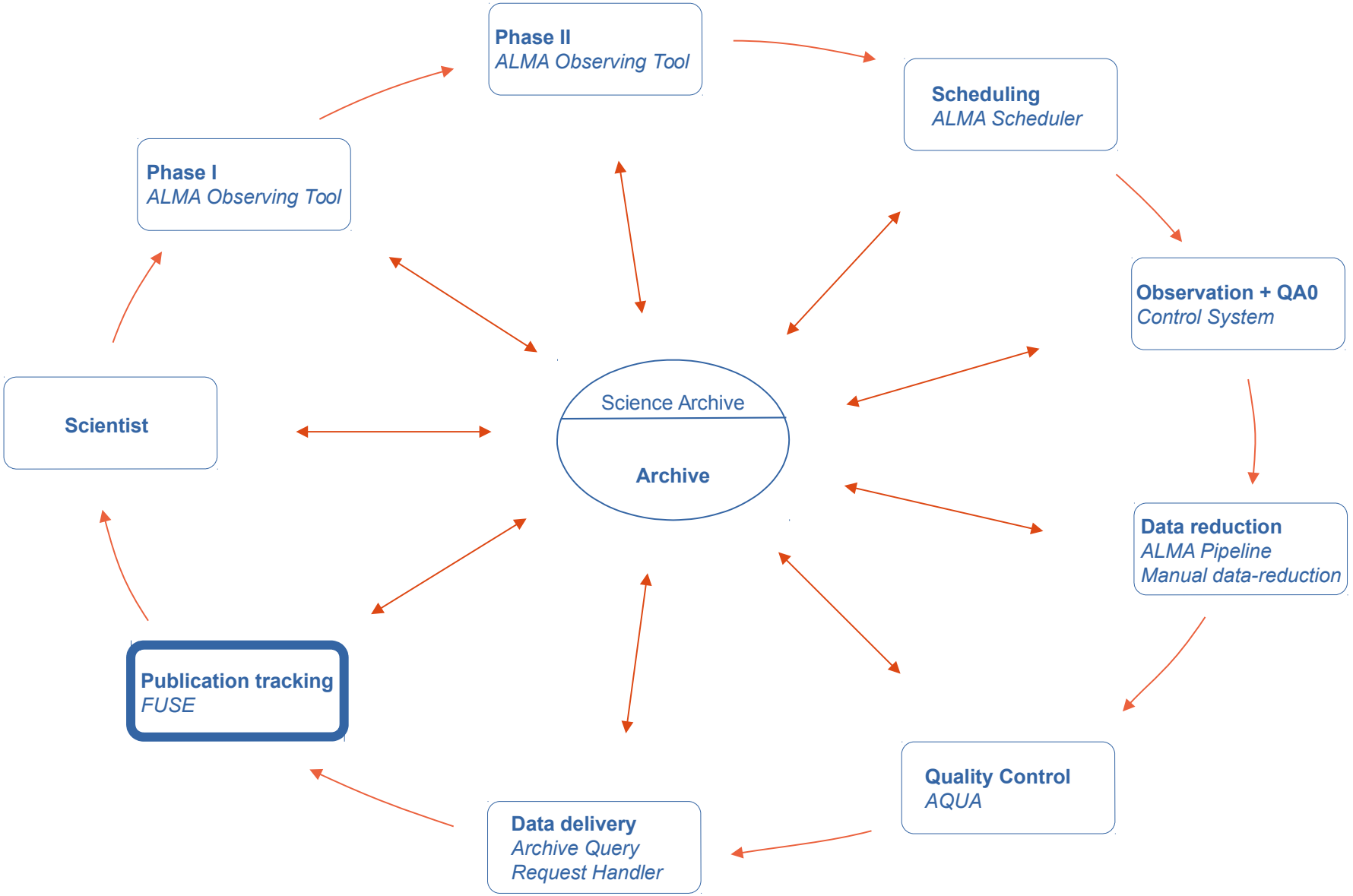
science archive

context

- success? → best possible **end-to-end user-experience**
- creating a good Science Archive **helps maximise the scientific return**
- great **return-for-investment** ratio
- interoperable: VO

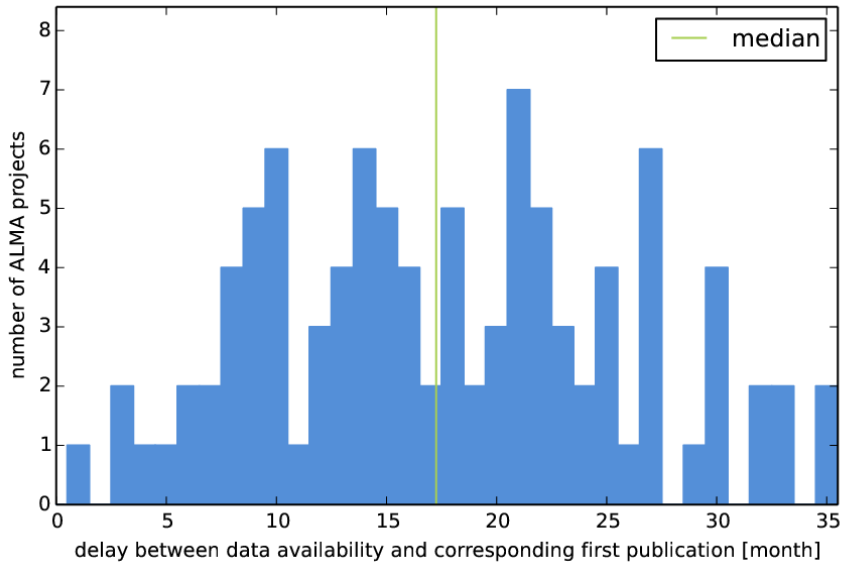


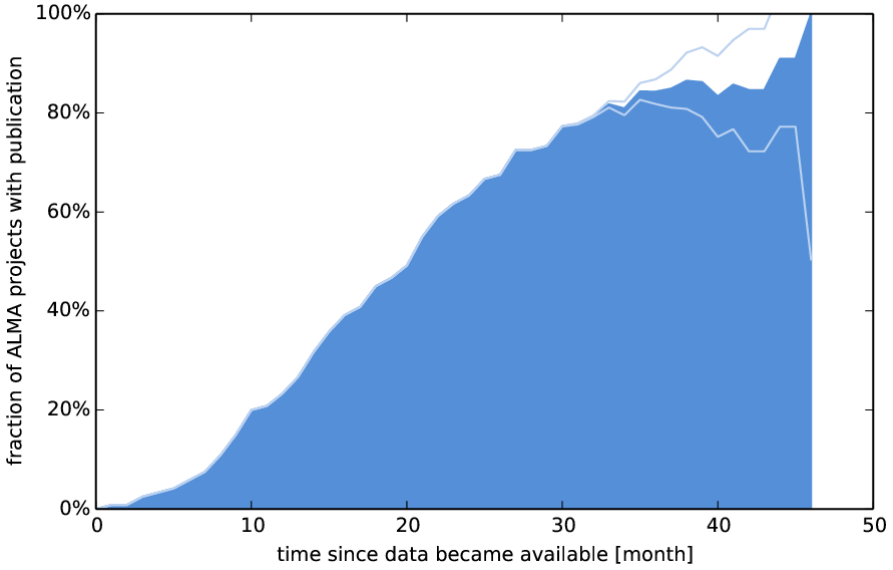
publications



publications

- ALMA policy: users **must** cite the data they use
- ESO/NRAO/NAOJ libraries **track** publications
- we check arXiv and contact PIs if needed before articles go into press





publications

ALMA project status questionnaire

ALMA has great interest in providing data and support in a way that enables PIs and the community at large to advance science and to publish their findings in astronomical journals.

While we attempt to provide data that is useful for expert and non-expert PIs, we may not be reaching this goal effectively in some cases.

According to our records, it has been two years since the end of the proprietary period of your project, and we have been unable to identify a related publication.

We would like to use this occasion to ask you to help us provide better data and support for yourself and future users of ALMA, by giving us your feedback on this very short, anonymous questionnaire.

* **Please describe the status of your project.**
Choose one of the following answers

- ○ There is a publication. (Please tell us below which journal we need to add monitoring)
- ○ A publication is in press.
- ○ A publication is in preparation.
- ○ Only part of the requested project was actually observed. The data were not complete enough for the expected science.
- ○ The project was complete, but the quality of the data was not good enough.
- ○ The project was complete, and the quality of the data was good enough but we had problems producing the data products.
- ○ Although the data were complete and good enough, the expected science was not contained in the data. For example, the experiment was a detection experiment.
- ○ Although the data were complete, good enough and the expected science was contained in the data, the scientific field had moved on in the meantime. For example a competitor has already published similar results.
- ○ Although the data were complete, good enough and still relevant, no effort was available (any more) to analyze the data.
- ○ Although the data were complete and good enough, still relevant and effort was available we have to wait for data from other facilities. (Please tell us which ones)
- ○ Personal reasons.
- ○ Other:

We would very highly appreciate any other comments or suggestions you might have for us with respect to the survey above but also with the general view to improving the publishable quality of ALMA data as well as our support to PIs and the community at large.

Submit

Exit and clear survey

publications

context

- publication tracking is
  - essential to **measure the success** and
  - a crucial tool to **improve operations**

to make sure that the investment really gets converted into science

summary

- ALMA's has a modern science data management concept
- goal: **science-grade products** for all raw data
- extreme **user-support**
- so far ALMA is **very successful**
- “**transformational facility**”
- still **a lot of work** is required to develop ALMA to its full potential

summary

context

- **success** = maximizing the **end-to-end user experience**
- **astronomers** not data will be the rare resource
- more **work and responsibility** will shift **to observatories**
- the way we do science is changing

- This will be **golden times for astronomers**