Kindling Community Science Tool Development: Final Report

Adam K. Leroy

Goals: In Fall 2013, we were awarded an ALMA development study grant, "Kindling Community Science Tool Development: A Test of the Python Third-Party Model." The goal of the grant was to support a focused workshop to figure out how utilities for data cube analysis could be created and shared among the community. We aimed to test an approach in which community astronomer developers leveraged third-party python packages (numpy, scipy, astropy) to produce tools to analyze data cubes. Our goals were to test the viability of the model, explore and discuss the interplay with the NRAO-led CASA data reduction package, seed development by developing some basic functionality, and produce documentation that would be useful for starting developers.

Workshop: This workshop was help from 7 – 18 April, 2014 in Charlottesvile, Virginia. The workshop included community experts and representatives of the CASA team, the AstroPy team[1], the yT[2] team, and the ALMA development programs CARTA and ADMIT. The workshop took place at NRAO's Edgemont Road (ER) facility, mostly in room 230.

Preparation: The NRAO IT department (especially Josh Malone and Tracy Halstead) set up computing resources in ER 230 that allowed the visitors to access to these test data. They also created several machines that provided a clean environment on which to test software. Their hard work paid off and the computing environment worked very well. Lyndele von Schill handled administrative preparations, including accommodation, plane flights, and catered lunches for the first couple of days, and breakfasts for the first few days.

Monetary Expense: the monetary expense involved in the project was entirely to support the travel of the participants to the workshop. This included plane tickets, accommodation, and board. The list of external attendees who were supported is: Chris Beaumont (Harvard), Adam Ginsburg (ESO), Jeffrey Kern (Socorro), Thomas Robitaille (MPIA), Erik Rosolowsky (U. Alberta), Sam Skillman (Stanford), Peter Teuben (U Maryland), John Zuhone (Goddard).

NRAO Personnel Participation: Lyndele von Schill contributed administrative support before and during the workshop. Tracy Halstead and Josh Malone contributed several days of IT expertise and setup. Adam Leroy contributed time to organizing the workshop before it occurred. Adam Leroy and Drew Brisbin prepared a large set of data that demonstrated the scientific and technical challenges posed by ALMA data. Jeffrey Kern, Crystal Brogan, Adam Leroy, Todd Hunter, Drew Brisbin, and Remy Indebetouw all participated in portions of the workshop. Adam Leroy carried out administrative follow-up afterwards.

Contents of Workshop: The workshop consisted of a mixture of group discussions, presentations by representatives of other software groups, and time spent coding or brainstorming in small groups or alone. The interface with CASA was a main focus of the first week. The second week saw the production of a basic framework for data cube analysis in python. A suggested framework for community participation was drafted by the group.

Output: The workshop produced the following output:

---

[1] AstroPy is a large community effort to develop python tools to support astronomy.

[2] yT is the leading software package for simulation visualization in astronomy.

1. An online organization at http://radio-astro-tools.github.io/ to serve as a central repository for community-created software. This is a space where community developers can post material with varying degrees of rigor. It also includes the ability to make feature requests and version control.

2. A large and growing (a year later) package, hosted off this same website to handle spectral data cubes (the key data unit for ALMA) in a python environment. This is written with the testing and documentation standards of the astropy project.

3. Some documentation and tutorial content (https://github.com/radio-astro-tools/tutorials) related to these served in the same location.

4. A high level suggested community vision and a suggested mode of interfacing with CASA, to be summarized in a NAASC memo.

5. A detailed approach for interfacing these software tools with CASA in a way that does not add responsibilities to the CASA team. The script for this is served off the repository above.

6. Experience for the yT group, which is primarily theoretical, interfacing with radio observers. Substantial progress was made interfacing ALMA data with yT tools, which has influenced subsequent yT development. An interface to the yT software package within the data cube object described above.

Of particular note, Adam Ginsburg (a postdoc at ESO) has made substantial progress on the infrastructure discussed in (1) and (2) since the workshop. He has presented posters on these results at the ALMA conference in Tokyo and the AAS meeting.

Study Findings and Recommendations: the high level goal of a development study is to explore an area for future work. In this sense, the workshop found the following:

1. For current realistic data sets, python analysis that leverages numpy and scipy appears viable. Memory mapping techniques available in python may help keep this a viable technique for very large data.

2. The "soft" advantages of this approach are substantial because a large amount of community effort is now going toward python-exposed tools. Though most of these do not directly affect ALMA, they add momentum.

3. Though the interface is not yet totally trivial, moving ALMA data into the yT simulation analysis environment is possible and offers a way to get access to a very large set of existing analysis functionality without rebuilding it from the ground up. The code package produced by the workshop includes methods that make this much easier and members of the yT team have continued to be involved.

4. A major shortcoming of the third party approach is the interconnected web of packages and libraries. Installation, versioning, and dependencies all make bundling of code challenging. Including third party software with CASA or introducing CASA dependencies into astropy (or a similar package) would incur a substantial workload. This is a strong argument against bundling or a vision where CASA ships

astropy or vice versa. On the other hand, the incurred cost is that installation (and so use) is more difficult and the burden is shifted to the user if things are not shipped together. This is a general python issue.

5. The largest software package produced at the workshop is aimed at providing the underlying infrastructure (a data object) that higher level routines would use. In this sense, it is both the necessary pre-requisite to kickstart this kind of development and represents a "language" that subsequent development would use (they would adopt this data model or their compatibility would suffer). In this sense, it can be thought of as a potential skeleton for future work and also as the most challenging piece of code.

6. The workshop also demonstrated a path for related packages that are written at a higher level. It produced a tool to make position-velocity cuts and a tool to handle radio beam conventions. Both are served off of the same public space and are written at a higher level than the data cube object (these are astronomer-level, while the data cube required something closer to developer-level).

7. Since the workshop, development has continued mainly via the efforts of Adam Ginsburg, a postdoc at ESO with help from Chris Beaumont and the yT team (where that interface is concerned). Ginsburg is able to contribute because this is part of his functional allocation and he considers this a worthwhile science investment. A team of equivalent contributors would be the key for something like this to gain momentum, but they do not really exist. Realistically, this cannot be senior staff or faculty with other responsibilities. It must either be coding-oriented postdocs or staff whose assigned responsibility is developing these tools. In that sense engaging Dr. Ginsburg was a major success of the project but a negative finding is that this tool development is not yet so high level that it can be done in the background as astronomers carry out science.

8. Similarly, tutorial development is something that appears to take allocated person-power. If this is not an assigned job, it is not something that astronomers do naturally as part of their science. The tutorials produced at and around the workshop are useful, but have not been spontaneously expanded in the last six months. Like code development, this is an area where assigning staff functional resources may be the best way to see progress (if it is judged important).

9. Community interest (visible as issues filed and emails) is slowly ramping up on the project and remains high a year later. This specific project certainly still represents a viable avenue for tool development. Python is undoubtedly the future for astronomical coding and something like this path is a natural and productive way for the community to maximize science output from ALMA data.