Sustainable Open-Source Tools for Sharing and Understanding Data

Dr. Ted Habermann, The HDF Group

Forward

A real-world example of open source sustainability from the stewards of a widely used open-source software tool might help us understand some aspects of the sustainability challenge and focus our discussion.

Introduction

Access to scientific results depends, at the most fundamental level, on the data format and tools available across disciplines, programming languages, and computing platforms for accessing those data. Once they are accessed, understanding the data depends on metadata that are closely integrated with the data at every level. The HDF Group has been solving these fundamental data storage, access, and understanding problems for scientists and analysts in many disciplines and sectors for many years (Habermann, 2014). We have provided open-source software and tools that have a well-proven track record of success and a growing commitment from the open-science community (Collette, 2015). Sustaining this support depends on integrating the expertise of the creators and maintainers of HDF (The HDF Group) with this burgeoning open-source community.

HDF (The Hierarchical Data Format) is a fast, flexible, and scalable data storage and access platform, providing free and open source data solutions for government, academia and the private sector. HDF technologies support cutting-edge research in climate science and earth observations, particle and plasma physics, seismology, environmental, planetary, biomedical and geodetic sciences among many others. Many research facilities rely on HDF for data preservation and sharing. Organizations with mission-critical systems depend on HDF. These users and organizations are the HDF community, and an important source of new and innovative uses of, and sustainability for, the HDF libraries and tools.

Several broadly successful open-source ecosystems are centered around HDF. In the Python world, these include h5Py, PyTables, Pandas and PyDAP. Searches of gitHub reveal over 1000 repositories built around these tools. Unfortunately, sustainability of these critical tools is threatened by recent changes: the primary developers and/or moderators of these tools have recently had to decrease their commitments to maintaining these tools due to changes in their employment.

The HDF Group (www.hdfgroup.org, THG) created and has maintained HDF for nearly 30 years, first as part of the National Center for Supercomputer Applications at the University of Illinois, and then as an independent, non-profit organization. The mission of the group is "to ensure the sustainable development of HDF (Hierarchical Data Format) technologies and the ongoing accessibility of HDF-stored data." The recent changes described above will leave significant gaps in many open-science communities that are using Python and HDF. We propose to address those gaps by 1) streamlining existing code bases for major python HDF interfaces (pyTables, h5Py, Pandas and pyDAP), and 2) providing on-going leadership, moderation, and community building for these tools.

Streamlining the HDF Python Code Base

Multiple approaches emerge in many development environments and become difficult to maintain over the long-term, so minimizing the amount of code to be maintained is a well-established best practice for long-term sustainability. The scientific python community recognized this problem with multiple HDF interfaces and outlined a path forward that included clear goals and milestones (Scopatz, 2015).

Initial steps along this path were taken during a workshop during 2016 (Python And HDF5 Hack-fest), but substantial work remains. We propose to further that work and increase sustainability of the HDF platform in three ways:

1) two focused design and development workshops with principle developers of h5py, pyTables, Pandas and THG will be organized during 2017 to identify obstacles and outline solutions to the merger of the code bases.
2) hiring an experienced python developer that can attend these meetings to gain the specific experience needed to make the code changes required to implement the outlined solutions. This developer will focus 100% on this project until the items identified by Scopatz are accomplished.
3) hiring an experienced community manager to facilitate communication, connections and training across the scientific python/HDF community.

Engaging and Supporting the HDF Scientific Python Community

The second element of this work will involve on-going engagement of HDF users in the scientific python community through moderation of the new streamlined HDF interface and on-going interactions with the community. This will begin during the development workshops described above and continue into the future. Our goal is to foster and strengthen connections among existing HDF users and to leverage the expertise and resources already invested in HDF technologies by diverse user communities.

When funded, this project will provide the seed resources required to help us build the foundation for a broader HDF community of practice and allow us to support community-driven collaboration and sharing. It will provide a foundation for sustaining HDF technologies, accelerating scientific and technological breakthroughs, encouraging innovation, and promoting current and future use, understanding and preservation of data.

Project Timeline

The goal of this proposal is to provide start-up resources required for THG to set out towards a goal of active support for and engagement with the community of scientific python users. We expect this initial effort to last two years. Tasks include workshops, code streamlining, and community outreach. We expect the ongoing work to become self-sustaining after two years.

Two workshops with principle developers and THG staff. Deliverable: Output of these workshops will be detailed plans addressing the development goals listed above.

A. Code Streamlining – Deliverable: single streamlined code base for access to data in HDF using Python.
B. Addressing existing issues, pull requests, and other community actions – Deliverable: Up-to-date and actively moderated git repository for unified code base.
C. Initial community outreach – Deliverable: a connected source of information about how HDF is being used in multiple scientific communities.
D. On-line and face-to-face trainings - Deliverable: A guide to existing materials developed in the community and a integrated video/example set that covers the basic skills required for effective use of HDF in any discipline.
E. Attendance and presentations/tutorials at scientific python conferences.

Risks

Whenever code is reused and refactored there are risks of unforeseen technical obstacles that can delay or even stop progress. The principals in PyTables and h5py have discussed this project at length and held a workshop exploring potential problems and ways forward, so we do not expect

technical issues that cannot be overcome. Our goal will be to identify important issues early and work with current experts to solve them in the design phase – before they become showstoppers.

Building functional and effective communities is also challenging. In this case, we are working with well-established open-source communities with long-term track records of success. Facilitating connections between these communities will make them stronger and more effective.

Participants

We have positive feedback from a diverse group of global HDF community members, keenly interested in participating in the community building process. We have gathered experts from the biomedical, earth, environmental, plasma and particle physics, and planetary sciences together with private sector representatives such as MathWorks, Esri, and Tech-X. These community members represent an open source community serving a diverse, multi-disciplinary community of scientists and decision makers. The diversity of expertise and applications of HDF technologies among these groups is significant and will result in exciting new discoveries and applications.

The expertise of The HDF Group combined with that of the various user communities is a powerful collaboration to advance data usability, understanding and preservation. Together we can leverage resources to create a whole that is greater than the sum of its parts. In this process we will demonstrate the value of cross-disciplinary technology transfer that can facilitate innovation and new scientific discovery.

References

Collette, A., (2015), HDF5 is Eating the World, https://www.youtube.com/watch?v=nddj5OA8LJo, Retrieved 15:11, Nov. 26, 2016 (GMT).

Ford Foundation, (2016), Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure, https://www.fordfoundation.org/library/reports-and-studies/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/, Retrieved 15:11, Nov. 26, 2016 (GMT).

GitHub Search (HDF5), https://github.com/search?utf8=✓&q=HDF5, Retrieved 15:11, Nov. 26, 2016 (GMT).

GitHub Search (PyDAP), https://github.com/search?utf8=✓&q=pydap, Retrieved 15:11, Nov. 26, 2016 (GMT).

GitHub Search (PyTables), https://github.com/search?utf8=✓&q=pytables, , Retrieved 15:11, Nov. 26, 2016 (GMT).

H5Py, http://www.h5py.org, Retrieved 22:00, Nov. 25, 2016 (GMT).

Habermann, Ted, et al. (2014): The Hierarchical Data Format (HDF): A Foundation for Sustainable Data and Software. http://dx.doi.org/10.6084/m9.figshare.1112485. Retrieved 15:04, Aug 10, 2015 (GMT)

Pandas, http://pandas.pydata.org, Retrieved 22:00, Nov. 25, 2016 (GMT).

PyDAP, http://www.pydap.org/en/latest/, Retrieved 22:00, Nov. 25, 2016 (GMT).

PyTables, http://www.pytables.org, Retrieved 22:00, Nov. 25, 2016 (GMT).

Python and HDF5 HackFest, https://curtinic.github.io/python-and-hdf5-hackfest/, Retrieved 22:00, Dec. 10, 2016 (GMT).

Scopatz, A., (2015), Python & HDF5 – A Vision, https://hdfgroup.org/wp/2015/09/python-hdf5-a-vision/, , Retrieved 22:00, Nov. 25, 2016 (GMT).