Open Source White Paper

Primary author:

Brian R. Dennis, NASA GSFC, Solar Physics Laboratory, (Code 671)
Research Astrophysicist, RHESSI Mission Scientist and Co-Investigator
Phone: 301-286-7983
Email: brian.r.dennis@nasa.gov
X-ray and gamma-ray instrumentation and data analysis since 1967.

Contribution authors:

Joel Allred, NASA GSFC, Solar Physics Laboratory, (Code 671)
Research Astrophysicist, IRIS Deputy Project Scientist
Developer and Co-developer of many computational modeling tools, including the flare modeling package, RADYN, and the particle transport code, FP.

Charles N. Arge, NASA GSFC, Solar Physics Laboratory, (Code 671)
Head, Solar Physics Lab.
Primary developer of the Wang-Sheeley-Arge (WSA) coronal and solar wind model, now operational at NOAA's National Centers of Environmental Prediction.

Gordon D. Holman, NASA GSFC, Solar Physics Laboratory, (Code 671)
Research Astrophysicist, RHESSI Co-Investigator
Depends on open-source data and code to carry out scientific research since solar physics research generally depends on data and code from multiple resources. Open source data and code are critical to carrying out quality solar physics research and maintaining a vibrant, successful research community.

Andrew Inglis, NASA GSFC and The Catholic University of America
Research Scientist
Developer and former Board member of the SunPy project with 10 years of data analysis experience in solar physics.

Richard Schwartz, NASA GSFC and American University
RHESSI Co-Investigator
Written and used scientific software for almost all solar X- and gamma-ray instruments. One of the lead developers for SolarSoftWare (SSW)

Albert Shih, NASA GSFC, Solar Physics Laboratory, (Code 671)
Research Astrophysicist, RHESSI Assistant Mission Scientist and Co-Investigator
Scientific researcher and code developer/maintainer using NASA funds.

Anne K. Tolbert, NASA GSFC and American University
Software Analyst, RHESSI and SMM/HXRBS team member.
Active in writing software and making it freely available, primarily on SSW, for the past 35 years.

Dominic Zarro, NASA GSFC and Adnet Systems Inc.
Group Manager/Principal Support Scientist
SolarSoftWare (SSW) Co-Developer

Open Source White Paper

Given NASA's goal of maximizing science return, we believe that the case for open source, as with open data, is self-evident. Most solar physicists would agree that the open data policy has been a great success. Few would argue that we should return to the policies of the early space program that extended through the 1990s, when the PI of each instrument was able to limit access to the data from that instrument to the PI team and favored collaborators. While officially the PI had these restrictive data rights for only a year or so after the observations were made, in practice, many PIs maintained control by making it difficult to access the data and refusing to release software for accessing and analyzing the data to obtain useful scientific results. The great successes of missions like the Ramaty High Energy Solar Spectroscopic Imager (RHESSI) and the Solar Dynamics Observatory (SDO), both of which have a completely open data and data analysis policy, demonstrate the value of this open approach. Furthering our understanding of most solar phenomena now requires observations from different instruments sensitive to radiation over a wide range of the electromagnetic spectrum from radio waves to gamma-rays. In situ particle measurements from different locations in the heliosphere are also valuable in many cases. Thus, it is seldom possible to make any advances in solar physics with observations from a single instrument. The ability to easily access and use multi-wavelength observations has become crucial and this is only optimized with a fully open data policy.

The great success of the current open data policy is the motivation for developing an open source policy with all of its similar attendant benefits. A brief history of the development of NASA's open data policy shows how the current system has greatly increased the scientific return from most space missions. In the early days, before about 1980, the PI of a given instrument had complete control of the data obtained from the observations with that instrument. As a result, a person not on the immediate PI team could not get access to the data without the explicit approval of the PI. The principal author of this white paper is most familiar with the difficulties this caused with data from the different instruments on the Solar Maximum Mission (SMM) in the 1980s. This meant that the scientific results from each instrument were limited by the capabilities of the instrument team members. The argument was made that only the PI team members were sufficiently familiar with the instrument to both fully exploit its scientific capabilities and at the same time avoid publishing incorrectly analyzed results. The argument was even made that it would be difficult to assemble a team to build a new instrument if it was known that all data from the instrument would be openly available following launch. Only one of the instruments on SMM had an open data policy, Ken Frost's Hard X-Ray Burst Spectrometer (HXRBS). To our knowledge, this was the first NASA instrument to have a completely open data policy. It led directly to the HXRBS solar flare hard X-ray light curves and spectra becoming a standard that everybody used, similar to the standard Geostationary Operational Environmental Satellite (GOES) soft X-ray light curves that are available today. Even during the 1990s, data from all the instruments on the Solar and Heliospheric Observatory (*SOHO*) were not made freely available, greatly limiting the ability to jointly analyze data from different instruments observing the same events. Now, with the open data policy of RHESSI, SDO, and most other solar missions, everybody has the ability to compare observations at will, leading to an explosion of new multi-wavelength analysis and rapid advances in our understanding of many solar phenomena.

Along with access to data, access to the analysis software is equally important. Most modern scientific instruments are complicated and require specialized software to retrieve the data and to produce instrument-free results. Thus, it is important that this software and all calibration data be also publicly available so that the observations can be analyzed with minimum effort on the part of non-PI-team members. This has been made possible with the SolarSoftWare (SSW) facility that contains all IDL code for analyzing observations from most of NASA's solar instruments. Even HXRBS light curves and spectra from observations made in the 1980s can still be obtained and analyzed using software in SSW.

Now the question is, should this open policy be extended to various data bases and computer codes written to carry out theoretical modelling and simulations of different solar phenomena under investigation? Many such resources are already open. For example, the CHIANTI data base and codes to predict the X-ray spectrum from plasma at different temperatures, densities, and compositions. CHIANTI has been publicly available since its inception and, as a result, is widely used as a standard for solar physics. A similar combination of a data base and software to calculate the expected gamma-ray line and continuum emission from ions accelerated in solar flares that was developed by Reuven Ramaty is also now available on line. These publicly available facilities developed over many years have been a great benefit to solar physics thanks to their open and ready availability.

But what about more specific software developed for more limited applicability such as the codes for predicting the emissions from the energy release and particle acceleration in solar flares? Do the authors of these codes have the exclusive rights to using them, especially if they were funded by the US tax payer? We hear the same arguments used by instrument PIs in the last century that they would not have been motivated to build their instruments if they knew that they would not have control of the data that it produced. They argued that only members of their team were sufficiently knowledgeable about the instruments to carry out the data analysis correctly. It is true that some refereed papers have appeared in print with erroneous results because the authors did not fully understand how the instrument worked, but we would argue that these mistakes have been easily corrected and are greatly outweighed by the vastly increased numbers of papers that are made possible by the free access of the data to a much larger and more diverse number of scientists.

What about the argument that a person who develops a certain powerful code should retain exclusive rights to its use so that he or she can use it in a proposal for continued funding? But if the original work was funded by the US tax payer, then the author does not own the code. Furthermore, the reputation of the author of the code would be enhanced if the power and value of the code became more widely known through its general availability and use by other scientists. And, of course, the scientific return would be increased if more people could use the code.

Once it is generally agreed that open source is the best policy to optimize science return, the question arises as to how best to achieve that goal. We believe that that answer is with carrots rather than with sticks. By that we mean that individual researchers should be rewarded for doing the additional work involved in making their code publicly available and not punished for failing to do so. Until the advantages of an open source policy are recognized and accepted voluntarily by the whole community, an open source policy will fail, or at least it will not provide its full potential benefits. Each scientist must see the rewards of the policy, such as free access to other people's code, extra funds to make their code more user friendly and comply with community standards, and ultimately greater scientific return from their own work. They must see for themselves that the rewards of an open source policy significantly outweigh the added burdens.

One lesson about the open data policy that would also apply to open code is that significant funding must be made available to maintain the individual data sets and to ensure that they continue to be accessible on line given the continual changes in computing and software protocols. With open data, this is achieved by continuing to fund the PI team while the instrument is operating at a level sufficient to achieve this goal and by transferring the data sets to an active repository such as the Solar Data Analysis Center at Goddard once the particular mission is terminated. A similar policy should be adopted with an open code policy. In this case the writer of an original, highly useful, model code should receive sufficient funds to make it useable by other interested scientists and to maintain it in a useful form. That is, developers of computer code deemed by the community to be of high value (through peer review) could, in exchange for open access, receive sustained financial support much like those that receive funding to provide open access to data from operating instruments. The code itself could then be used and modified by other users to meet their requirements with some procedure to preserve a working version at all times. The Python code in SunPy, the community developed solar data analysis environment, is being developed along these lines and seems to be very successful with its open source policy. Currently, the main problem with this approach is the difficulty that model developers have in getting continued funding to maintain their code and to further improve it in the face of competition from other people with access to the same code. This issue must be addressed before the community will accept a fully open source policy.

We would welcome the development of an open-source policy for future NASA funded scientific research codes, as we believe it would be complementary to the very successful open-data policy that has been in place for a number of years now. Our impression is that the younger scientific community, in particular, is already familiar with and supportive of open-source software, and the typical tools that are used to maintain software archives, e.g. version controlled repositories such as GitHub and GitLab. This support seems to be growing throughout the community - see for example the very successful Astropy and SunPy projects.

Regarding the specific issue of NASA requiring that existing or previously developed scientific codes be made open source and freely accessible (Bullet point #2 in the Call for White Papers), this would likely incur significant costs and require the provision of additional financial support to researchers. NASA should be aware of this need and be prepared to support it. A reasonable expectation for an open-source code is that, although it may be provided as-is, it should be provided in a useable and readable state to other users. However, typical scientific code development practices are such that many existing codes would be all but impenetrable to only a very select group. At a minimum, to release an existing private code as open source, the code authors would be expected to ensure that the code is well-commented and documented. They should perform some testing of their code to ensure that it operates on different systems, or at a minimum determine what the operating requirements are.

It is important that no onerous requirements be placed on the authors of scientifically useful code to make it publicly available. Thus, for example, it may be unnecessary and counter-productive to impose any legal requirement on published papers for the release of all code used in the generation of the results presented. That is probably best left up to the refereeing process to work itself out rather than attempting to make it a requirement by edict.

Considering all of these factors, it seems to us that the case for open source is as strong as that for open data. As with open data, problems will arise in the case of work not paid for by the US tax payer, code subject to ITAR, etc. However, we believe that the benefits of an open source policy far outweigh any disadvantages, and the science return will be greatly increased.