Title: **A Software Archival Node for the Planetary Data System**

PI: Adam Brazier, Cornell University (abrazier@astro.cornell.edu, 607-255-1733)

Co-I: Chase Million, Cornell University (chase.million@gmail.com, 765-914-5336)

Co-I: Alexander Hayes, Cornell University (hayes@astro.cornell.edu, 607-255-1712)

Co-I: Mahadev Satyanarayanan, Carnegie Mellon University (satya@cs.cmu.edu, 412-268-3743)

Proposal Description

We propose to create a Software Node of the Planetary Data System (PDS) to archive and curate high-value software of specific scientific relevance to NASA's planetary missions. High-valued software includes that used to generate and analyze data archived elsewhere in the PDS, mission and instrument calibration and processing pipelines, and many analysis or processing tools written by other research groups or community members. In modern scientific workflows, the software is often nigh indistinguishable from the methodology itself, and therefore necessary for future researchers to understand, validate, reproduce, or build upon the earlier work. In this sense, such software *is* research data as stated in OMB Circular A110, and echoed in *NASA Plan: Increasing Access to the Results of Scientific Research*, where "*Research data* is [sic] defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings [...]" [1][2]. A Software Node fills a gap in the current capabilities and services of the PDS, which has no means for scientific software archiving, and serves the interest of NASA in maximizing the usefulness of mission data.

Simply sharing, publishing or releasing source code or a software description (e.g. via Github [3], a personal or institutional website, or a Software Interface Specification) is insufficient. Even assuming that the source code remains available outside of a formal archive like the PDS, all software becomes un-runnable over time through changes to the software and hardware operating environment in a process called "bit rot." This eventually results in complete obsolescence of the software—sometimes within just a few months—and any functionality is lost. To maintain the usability of software for decades or more requires a focused and intentional effort, as we propose, to create a record of the software, its operating environment, and the information necessary to recreate running instances of the software.

Scientific and mission software is especially susceptible to bit rot because it is often developed as a bespoke solution by small groups of people with relatively little support provided for ongoing maintenance. The field of Planetary Science is rare among scientific disciplines in that we largely share a single, core set of observational data—much of it archived by the PDS—from which almost all results are derived. These data were difficult and expensive to collect and represent unique observations, generated by unique instruments and missions. They have tremendous scientific and cultural value, which makes preservation of Planetary Science software particularly important; by preserving software, we give future researchers more capacity to understand or modify earlier methodologies to extract new knowledge from the data.

The practice of software archiving—as distinct from archiving of other digital data objects or the release of source code—is *new*. Only a small number of researchers are actively exploring this area, and the PDS Software Node will be the first archive of its type outside of extremely narrow contexts, test cases, or technology demonstrations. All of the required tools exist right now, though, and a delay in archiving valuable planetary science software will result in more of it becoming unrecoverable.

Examples:

- It is not uncommon to have a gap of decades between missions to specific planetary bodies, but deriving full meaning from new observations requires synthesis with prior. Without access to the software used during earlier investigations, future researchers working with ten or twenty year old archived data will be locked into the assumptions and contextual limitations of past. Under those conditions, even minor modifications to the prior work comes with considerable difficulty.
- Despite being on the same spacecraft, Compact Reconnaissance Imaging SpectroMeter (CRISM) [4] imaging data are usually map-projected using the 128-pixel-per-degree Mars Orbiter Laser Altimeter (MOLA) [5] shape model of Mars, while HiRISE [6] image projections typically use a Mars sphere model (which does not account for variations in local shape). A researcher may therefore need to exert a great deal of additional effort and accept large amounts of distortion in order to get images from these two data sets to align in the region of interest. Access to the actual pipelines used to generate the data in the first place—which could, in principle, be modified to use the same model—would dramatically simplify such activities.
- Mastrogiuseppe et al. 2014 [7] used the Cassini RADAR [8] as a sounder to probe the depth and composition of Titan's seas. The Cassini RADAR was not intended to operate as a sounder and the analysis required the application of custom tapering functions using the CPADS processing software currently only available within a subset of the Cassini RADAR Science Team. If Mastrogiuseppe were not a Cassini Co-I with access to the CPADS software, this important result would not have been possible.

Methodology

The Software Node will be a curated *archive* of software that has been used in the creation or analysis of other planetary data sets. It will be neither an active project management tool like Github [3] nor a directory like the IPDA Tools Registry [9]. Archived software will be those specific versions and builds of code which have supported PDS data creation or scientific publications. The software will undergo peer and compliance review prior to archiving. We will provide support to data preparers at all stages of work in service to the goal of a smooth and compliant archiving of data with maximum value to the community and public. Services will also be provided to enable and encourage use of archived software, including for search, retrieval, access, compilation, and modification.

The core units of data to be archived by the proposed node are: (1) source code, (2) documentation for installation, operation, and maintenance of the software, and (3) regression tests. These units can and will be stored as plain text and are therefore compliant with PDS4 standards (as "parsable byte streams" per 12.1-2 of the "PDS4 Concepts" document

[10]). All accompanying metadata will conform to PDS4 requirements. Regression testing is a software engineering practice that confirms that inputs to software continue to produce expected outputs; they require the input and output data, configurations, and running software. If "non-core" data is required by software or regression tests (e.g. calibration reference files), then it will also be archived. In many cases, such data may be appropriate for archiving in another Discipline Node, but otherwise it will be archived in the Software Node as supplementary data.

Core archival products will be peer reviewed by qualified members of the community for completeness and clarity of the source code, documentation, and regression tests with the objective that a future researcher would be able to reproduce the primary functionality by rewriting, porting or directly compiling the source. The reviewers will assess whether the source code and attendant documentation are comprehensible and reasonable based on their technical and domain expertise, as well as addressing the question of whether the software is of sufficient scientific value, or offers a meaningful contribution to the field of Planetary Science, that it *should* be preserved. Our emphasis is on the completeness, value, and utility of archived source code, not its *"quality,"* so reviewers will specifically not assess for that.

Core products will also be reviewed by node, network, or institutional staff for PDS4 standards, ITAR, and licensing compliance and adherence to the specific requirements of the Software Node. Node-specific requirements will be (1) that the software source code and attendant documentation for installation, operation, and maintenance must be structured and released into the public domain according to PDS policy and standards, and (2) node staff can reproduce the functioning software in a virtual machine (VM) using information in the source code and documentation, and (3) the software as reproduced in the VM must pass all regression tests. A VM is an emulation by one computer of another, and can be used to instantiate a precisely defined operating environment that matches the one under which the archived software was actually developed; in other words, not only does installation on a VM provide us with a means to verify that the software documentation, installation instructions, and regression tests are sufficiently detailed and accurate to permit reuse, but a VM also *freezes* the software's precise environment so that it can be executed even decades after creation.

For ongoing and completed missions or projects, project funds (and possibly the researchers themselves) may not be available to support the archiving process, and the software products may not have been designed with any intention of release or reuse. We will seek out, recover, accept and archive these "legacy" products, and will work with the appropriate teams to do so to the extent possible given limitations in funding and time. For such legacy software, case-by-case exceptions to the Node's baseline archival standards will be possible—related to completeness, clarity, regression tests, and the ability to build within a virtual machine—with the goal of archiving the products in as complete a manner as possible under any constraints and commensurate with their scientific importance.

Extant nodes already host some software as supplements to specific data; such codes will be candidates for migration to the Software Node. All past and current mission and instrument teams are likely to possess unreleased codes of high value, and we will seek out, identify, and acquire those. We will also request NASA's assistance in identification, location, acquisition,

and licensing for archiving of *other* planetary data processing and analysis software products, including encouraging and enabling current and future NASA planetary researchers to submit their software data products to the PDS or provide the Node team access to software-related assets from past projects which reside on NASA systems.

Encouraging and Enabling Use

The Software Node will be an active resource to the community. We will make the archived data products available and searchable through a public facing web portal and provide a range of means to access and use archived software (subject to the licensing and use restrictions of the archived software's dependencies, e.g. the Windows OS). Based upon archived information, we will develop and provide scripts for the creation of running instances of the software either natively or in VMs on users' own systems [11], and host virtual machine images of software running in its native environment for end users to run in a VM manager on their own systems or via "one-click," in-browser access using Olive Executable Archive technology (see below) [12]. For non-graphical software (e.g. many pipelines), we will provide remote API access to archived software running on our own servers. We will host and moderate a "Software Node Forum," through which to provide assistance on use of archived software, general advice and feedback on preparing software for archiving, and a space for community members to discuss specific pieces of archived software. We will also work with other nodes and the network to improve the use and utility of all PDS holdings.

Olive: Olive is a technology which provides access to archived executable content running within VMs on remote servers. The entire VM is streamed over the internet, much as video is streamed today. Users can browse an online library of available software and quickly assess, by *using it*, whether any specific resource meets their needs without having to download, configure or install it locally. Because changes to individual VM images do not modify the master image, of which they are clones, users can quickly make changes to source code and recompile within the Olive VM without the need to locally reproduce hardware and software dependencies, port the software to modern environments, or recreate (i.e. rewrite) the functionality entirely. Getting old, legacy software to work on modern hardware and operating systems can be a difficult and time consuming process. Olive dramatically decreases the "time to value" for users of archived software.

Team and Capabilities

Key Personnel: The proposed work is innovative and on the cutting edge of archiving practice. In addition to expertise in archiving, data management, and Planetary Science, it will require talent in software recovery, engineering, project management, and maintenance. Our team composition reflects this and contains all of the skills needed to meet the Node objectives.

- Dr. Adam Brazier (PI) works for the Center for Advanced Computing at Cornell University and has a decade of experience in scientific software engineering, scientific workflow design, database design and use, data management and archiving, software requirements elicitation and research in the physical sciences. His work includes data management on the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) [13] and the PALFA pulsar survey [14][15]. He also served as the CCAT

Science Software Architect, for which his responsibilities included leading the observatory software requirements elicitation and design of the observatory data flow and archiving for the entire CCAT Telescope project [16].

- Chase Million (Co-I) is Founder of Million Concepts LLC, a software contracting and consulting company that develops scientific research software—including calibration and analysis tools—in the fields of spacecraft-based Planetary Science and Astronomy. He has broad experience with PDS-held data, and worked on the MER Pancam [17] and Odyssey THEMIS-VIS [18] calibration pipelines. He led the effort to recreate core functionality from the completely obselesced Galaxy Evolution Explorer (GALEX) calibration pipeline (on which he was a developer) to calibrate and archive the approximately 1.1 trillion detected ultraviolet photons events at the Mikulski Archive for Space Telescopes (MAST) [19], making this data available for the first time. For the purpose of this proposal, his affiliation is with Cornell University and, if the Software Node is funded, he will become an employee of Cornell under this award.
- Dr. Alexander Hayes (Co-I) is an Assistant Professor of Astronomy at Cornell University. He has broad domain expertise in Planetary Science that includes planetary surface processes, spectroscopy, instrument development, and extensive mission experience at all phases. He is a Co-I of MastcamZ on Mars2020, a participating scientist of Mars Science Laboratory [20], a participating scientist of the Cassini-Huygens mission (as a team member of RADAR and associate of VIMS) [8][21], and served many roles as a collaborator on the Mars Exploration Rovers [22][23]. He is also the Director of the Spacecraft Planetary Imaging Facility (SPIF) at Cornell University [24].
- Dr. Mahadev Satyanarayanan (Co-I) is the Carnegie Group Professor of Computer Science at Carnegie Mellon University. He is a pioneer in distributed and cloud computing and was the principal architect of the Andrew File System (AFS) [25]. He is also the PI of the Olive Executable Archive project, which he will continue to develop and improve access for its applications in the Software Node [26]. He will also provide support and guidance on problems related to the recovery and archiving of executable content.

Facilities and Infrastructure: Node infrastructure will draw upon Cornell's existing IT security and data management capabilities and services, which will fully meet our needs in terms of security, disaster recovery, data integrity, and data sharing. These include high reliability server hosting, high performance computing, a scalable computing cloud (RedCloud), a full range of security and access controls, adaptable to need, and technical support staff with a wide range of expertise and experience. We will request that NASA provide access to its high performance and cloud computing infrastructures as needed to enhance our institutional capabilities to provide the best possible service for archive end users.

The Software Node Consortium: The Software Node will contain two consortium institutions. Cornell University will be the lead institution, and responsible for all scientific, technical, and archival domain needs. A "sub-node" led by Satyanarayanan at Carnegie Mellon will develop and provide technologies to maximize use and reuse of archived software products by the community and public by building upon expertise and capabilities already acquired through the Olive Executable Archive project.

References

[1] Office of Managment and Budget. Circular a-110 revised 11/19/93 as further amended 9/30/99. `https://www.whitehouse.gov/omb/circulars_a110/`. Section 36.d.2.i.

[2] NASA. Nasa plan: Increasing access to the results of scientific research. `http://science.nasa.gov/media/medialibrary/2014/12/05/NASA_Plan_for_increasing_access_to_results_of_federally_funded_research.pdf`. Online; accessed 12-April-2015.

[3] Github Inc. Github. `https://www.github.com`, 2015. Online; accessed 12-April-2015.

[4] S Murchie, R Arvidson, Peter Bedini, K Beisser, J-P Bibring, J Bishop, J Boldt, P Cavender, T Choo, RT Clancy, et al. Compact reconnaissance imaging spectrometer for mars (crism) on mars reconnaissance orbiter (mro). *Journal of Geophysical Research: Planets (1991–2012)*, 112(E5), 2007.

[5] M_T Zuber, D_E Smith, SC Solomon, DO Muhleman, JW Head, JB Garvin, JB Abshire, and JL Bufton. The mars observer laser altimeter investigation. *Journal of Geophysical Research: Planets (1991–2012)*, 97(E5):7781–7797, 1992.

[6] Alfred S McEwen, Eric M Eliason, James W Bergstrom, Nathan T Bridges, Candice J Hansen, W Alan Delamere, John A Grant, Virginia C Gulick, Kenneth E Herkenhoff, Laszlo Keszthelyi, et al. Mars reconnaissance orbiter's high resolution imaging science experiment (hirise). *Journal of Geophysical Research: Planets (1991–2012)*, 112(E5), 2007.

[7] Marco Mastrogiuseppe, Valerio Poggiali, Alexander Hayes, Ralph Lorenz, Jonathan Lunine, Giovanni Picardi, Roberto Seu, Enrico Flamini, Giuseppe Mitri, Claudia Notarnicola, et al. The bathymetry of a titan sea. *Geophysical Research Letters*, 41(5):1432–1437, 2014.

[8] Ch Elachi, MD Allison, L Borgarelli, P Encrenaz, E Im, MA Janssen, WTK Johnson, RL Kirk, RD Lorenz, JI Lunine, et al. Radar: the cassini titan radar mapper. *Space Science Reviews*, 115(1-4):71–110, 2004.

[9] International Planetary Data Alliance (IPDA). Ipda tools registry, 2015. Online; accessed 12-April-2015.

[10] Data Design Working Group. Pds4 concepts. Technical Report 1.0.0, May 2013.

[11] Kam Woods and Geoffrey Brown. Assisted emulation for legacy executables. *International Journal of Digital Curation*, 5(1):160–171, 2010.

[12] Mahadev Satyanarayanan, Gloriana St Clair, Benjamin Gilbert, Jan Harkes, Dan Ryan, Erika Linke, and Keith Webster. Olive: Sustaining executable content over decades.

[13] MA McLaughlin. The north american nanohertz observatory for gravitational waves. *Classical and Quantum Gravity*, 30(22):224008, 2013.

[14] JM Cordes, PCC Freire, DR Lorimer, F Camilo, DJ Champion, DJ Nice, R Ramachandran, JWT Hessels, W Vlemmings, J van Leeuwen, et al. Arecibo pulsar survey using alfa. i. survey strategy and first discoveries. *The Astrophysical Journal*, 637(1):446, 2006.

[15] Benjamin Knispel, P Lazarus, B Allen, D Anderson, C Aulbert, NDR Bhat, O Bock, S Bogdanov, A Brazier, F Camilo, et al. Arecibo palfa survey and einstein@ home: binary pulsar discovery by volunteer computing. *The Astrophysical Journal Letters*, 732(1):L1, 2011.

[16] Thomas A Sebring, Riccardo Giovanelli, Simon Radford, and Jonas Zmuidzinas. Cornell caltech atacama telescope (ccat): a 25 m aperture telescope above 5000 m altitude. In *Astronomical Telescopes and Instrumentation*, pages 62672C–62672C. International Society for Optics and Photonics, 2006.

[17] JF Bell, SW Squyres, KE Herkenhoff, JN Maki, HM Arneson, D Brown, SA Collins, A Dingizian, ST Elliot, EC Hagerott, et al. Mars exploration rover athena panoramic camera (pancam) investigation. *Journal of Geophysical Research: Planets (1991–2012)*, 108(E12), 2003.

[18] TH McConnochie, JF Bell, D Savransky, MJ Wolff, AD Toigo, H Wang, MI Richardson, and PR Christensen. Themis-vis observations of clouds in the martian mesosphere: Altitudes, wind speeds, and decameter-scale morphology. *Icarus*, 210(2):545–565, 2010.

[19] C Million, SW Fleming, and B Shiao. gphoton. `https://www.github.com/cmillion/gPhoton`, 2014. Online; accessed 12-April-2015.

[20] John P Grotzinger, Joy Crisp, Ashwin R Vasavada, Robert C Anderson, Charles J Baker, Robert Barry, David F Blake, Pamela Conrad, Kenneth S Edgett, Bobak Ferdowski, et al. Mars science laboratory mission and science investigation. *Space science reviews*, 170(1-4):5–56, 2012.

[21] Kevin H Baines, RH Brown, Dennis L Matson, RM Nelson, BJ Buratti, JP Bibring, Y Langevin, C Sotin, A Carusi, and Angioletta Coradini. Vims/cassini mission at titan: Scientific objectives and observational scenarios. In *Symposium on Titan*, volume 338, pages 215–219, 1992.

[22] Raymond E Arvidson, Steven W Squyres, Robert C Anderson, JF Bell, Diana Blaney, Johannes Brueckner, Nathalie A Cabrol, Wendy M Calvin, Michael H Carr, Philip R Christensen, et al. Overview of the spirit mars exploration rover mission to gusev crater: Landing site to backstay rock in the columbia hills. *Journal of Geophysical Research: Planets (1991–2012)*, 111(E2), 2006.

[23] Steven W Squyres, Raymond E Arvidson, D Bollen, JF Bell, Johannes Brueckner, Nathalie A Cabrol, Wendy M Calvin, Michael H Carr, Philip R Christensen, Benton C Clark, et al. Overview of the opportunity mars exploration rover mission to meridiani planum: Eagle crater to purgatory ripple. *Journal of Geophysical Research: Planets (1991–2012)*, 111(E12), 2006.

[24] JJ Hagerty et al. The regional planetary image facility network. In *Lunar and Planetary Institute Science Conference Abstracts*, volume 43, page 1548, 2012.

[25] James H Morris, Mahadev Satyanarayanan, Michael H Conner, John H Howard, David S Rosenthal, and F Donelson Smith. Andrew: A distributed personal computing environment. *Communications of the ACM*, 29(3):184–201, 1986.

[26] www.olivearchive.org. Online; accessed 13-April-2015.