TOPICAL: Digital transformation of scientific data management to increase investigation transparency, reproducibility, and efficiency and data re-use

Dan Berrios
Space Biosciences Research Branch
1 (650) 604-0470
KBR/NASA Ames Research Center
daniel.c.berrios@nasa.gov

Sylvain V. Costes
Space Biosciences Research Branch
1 (650) 604-5343
NASA Ames Research Center
sylvain.v.costes@nasa.gov

Alan Wood
Space Biosciences Research Branch
1 (510) 548-5333
KBR/NASA Ames Research Center
alan.e.wood@nasa.gov

Karen Stephens
Exploration Systems Development
1 (256) 975-9905
Marshall Space Flight Center
karen.l.stephens@nasa.gov

Introduction

Conducting scientific research requiring spaceflight presents special challenges not encountered in terrestrial investigations. Chief among these are the enormous and complex engineering and operations efforts required to realize scientific investigations in space environments. Scientists typically do not have deep expertise regarding these efforts, and conversely, engineers and operators typically do not have deep scientific expertise in the area being investigated. Consequently, research requiring spaceflight to date has traditionally been conducted by dividing responsibility for the investigation between two groups: the principal investigator's (PI) team and the flight experiment implementation team (FIT) provided by space agencies. The PI is responsible and has authority over the theoretical design of the experiment, receiving the experimental data from the FIT, data analysis and publication. The FIT has authority over and is responsible for the execution of the experiment, and acquisition and delivery to the PI of the experiment data. While this collaboration properly utilizes the expertise of each team, it has resulted in each having their own separate processes and systems for data management. This is not only inefficient, requiring additional effort to bridge these systems for data transfers, but results in a lack of insight between these two teams regarding data and metadata management, which is essential to the tight integration of investigatory processes.

This dichotomy in data management also has implications for implementation of Open Science and Open Data directives at the national level [1] [2] that promote investigators and institutions providing the scientific community with prompt and wide access to all investigational data. The motivations for the directives include supporting the ability of other investigators to reproduce experimental results and promoting data reuse by other investigators to derive additional results, as well as fostering multi- and interdisciplinary investigation. Additionally the initiatives aim to open up avenues for the convergence of ideas and thoughts through world-wide scientific collaborations. While this evolution in data management could in fact be revolutionary in its ability to accelerate the pace of scientific discovery, it is not likely to be widely embraced by the space biology research community without 1) additional investigator tools supporting data management; 2) advancement of data management policies by organizations operating science data repositories; and 3) additional funding for these repositories. Institutions that support scientific investigation need to address the increased burden on investigators, as they evolve their investigational data management processes to meet the goals of Open Science/Open Data.

We propose an evolution of the spaceflight scientific investigations paradigm with respect to data management, in part, to provide this support to PIs and FITs, and realize the goals of Open Science and Open data. This shift will require practices by investigators that are different from what has been expected of them in the past, including a) more consistent and detailed planning for data management at the proposal stage of science investigations; b) forming an agreement with Open Science Data Repositories (OSDR) early in the process of investigation to submit all data; and c) using the OSDR as a safe harbor for investigational data, as a source of mission data, and for general guidance on data management.

Early Life-Cycle Process Improvement and Automation

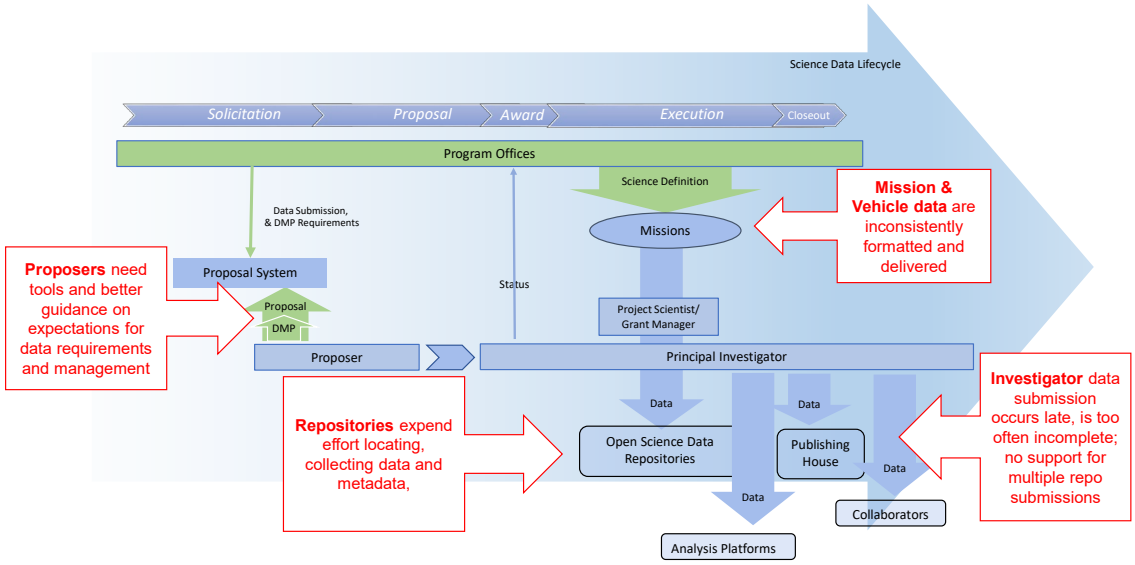

Figure 1.The current paradigm for management of data from space biological and physical science investigations.

Early in the life cycle of an investigation (Figure 1), sponsoring research programs frequently require proposers to provide information regarding assay parameters, and data acquisition, storage, back-up, transfer and dissemination (often called the "Data Management Plan," DMP). However, many programs lack processes and tools to assist proposers in drafting data management plans, or for evaluating them prior to submission with their proposal. This particularly impacts investigators new to the field, whose lack of experience with such planning may represent a daunting obstacle to proposal. The result is investigators expending substantial effort using manual methods that generate a wide variety in format, quality, and level of detail in DMPs, with some proposers providing little more than a few sentences of vaguely defined intentions. Furthermore, the repositories that ultimately will archive data from these investigations, many of which are managed by the sponsoring organization, often have no engagement with sponsoring programs or proposers during data management plan development [3]. This represents a missed, early opportunity for repositories to coordinate with and inform investigators on best approaches for data planning, organization, annotation, means of transfer, and transfer schedules, long before the data are acquired.

While there could be several different approaches to providing investigators with data management plan support, including referring proposers to recommended "general-purpose" data management plan tools (e.g., [4]), the development by research sponsoring institutions of specialized tools for proposers would offer several advantages. First, such tools could incorporate required domain-specific standards for data and metadata in the support offered; candidate plans could be validated against this set of standards and corrections recommended when necessary. Second, tools hosted by sponsoring organizations could retain generated data management plans from proposers and, for awardees, automatically match up planned products with data products eventually delivered to repositories hosted at the organization. This not only

reduces delivery checking efforts by investigators and repositories, but would provide automated reliable, available status information on the progress of investigations to programs and other interested parties.

Research awards for investigations in space biology and physical science have required that investigators provide copies of their investigation data to sponsoring organization repositories for decades. Nevertheless, the compliance with these requirements has been, at best, modest and for some programs and periods, abysmal. This potentially represents a great loss of research investments dollars if these data are never acquired by sponsoring organizations and never shared with others in the scientific community. One way to counter this trend would be to develop processes for and require investigator signature on data submission agreements (DSA) or similar compacts. Even without any legal force, such agreements would serve as additional motivation for researchers to comply with award requirements for submitting investigation data, and provide a formal basis for inquiries by sponsor organizations on status of data submissions. The development of tools to generate a DSA could be incorporated into sponsoring organizations' repositories, and such tools could leverage information from a previously generated data management plan (see above) as the foundation for the DSA.

Mid-lifecycle Automation

As shown in Figure 1, investigators in the biological and physical sciences commonly only share their data through public repositories at the end of the investigation life cycle. It has long been the case that research information was closely guarded by investigators fearful of colleagues being alerted to their experiment designs, methods, or results, and using any of these to advance their own research before an investigator can lay claim to them through report publication. This has led many investigators to procure their own systems for the storage of investigation data, and, particularly for high volume and "big" data, at significant cost. Some investigators choose data storage and transfer systems that lack robustness and scalability, sometimes resulting in loss of data integrity or additional effort switching to more capable solutions. Additionally, at or near the end of the investigation, the investigator and archive repositories expend much effort organizing and executing the deposition of the investigation data into the required archive(s). Much of this effort is in the form of time-consuming human-to-human communication between investigation and repository operations, since developing automation of these processes has not been practical due to the wide variety in investigation data storage solutions.

To address these inefficiencies in mid-lifecycle processes, sponsoring research organizations should develop and offer cloud-based, investigator-controlled "experiment workspaces" (EW) as standard components of, and services offered by, the open science data repositories they manage (Figure 2). These EW would be akin to collaborative systems that have been proposed for terrestrial research and clinical investigators [5]. Not only could such workspaces free investigators from the time and expense of designing or procuring reliable, accessible, and maintainable investigation-related data storage, but the time required to publish data could be substantially reduced. Investigators would be free to add data to EW at any time during the

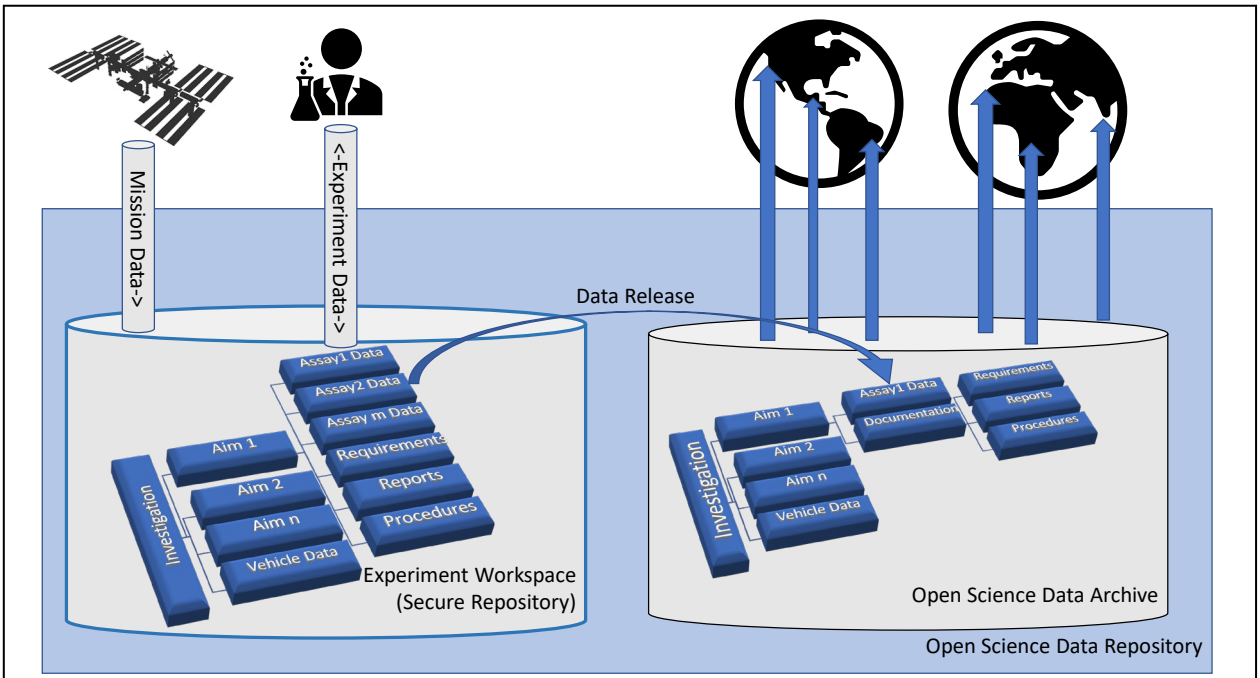

Figure 2. The Experiment Workspace consisting of a restricted access repository for experiment raw data (left), and a published repository for released experiment data, mission and environmental data, science requirements and reports, etc. (right).

investigation, and could share data with collaborators, journal editors, and article reviewers (i.e., EWs should have all the functions of investigator-owned data storage).

Spaceflight investigations almost always involve the collection of vehicle and experiment environment data and metadata ("Missions & Vehicle data," Figure 1), because these data provide critical context on space environment exposures and/or investigation treatments. These data are often deposited into research sponsoring organization repositories independently of investigator-provided data. Much effort is then spent by all involved parties associating these data with investigator-provided data to yield correctly defined scientific contexts. This effort would be much more straightforward and timely (and the results more reliable) if instead these data were deposited at the time an experiment is conducted into the same EW used by the PI team (Figure 2). Not only would repositories have greater insight on capture of these data, but investigators could have immediate access to them, perhaps influencing decisions regarding experiment operations or planning. Ensuring that metadata are centralized and searchable would provide for better experimental designs through improved understanding of how environmental factors impacted prior spaceflight experiments. Finally, another important advantage of this approach is that investigation "dashboard" systems that leverage EW usage metadata could provide sponsoring programs with more accurate and timely reports of investigation progress.

Late Life-cycle Automation

Once the PI team and FIT have acquired most or all investigation data, each team spends significant effort sharing these data with each other and submitting them to the required archives.

The development of EW, as described above, could accelerate the transfer of FIT-gathered data into long-term archives. However, the final publication of all investigations requires the detailed structuring and association of these data with those acquired by the PI team. NASA should invest in the development of more capable data curation tools that automate as much of this process as possible. These Advanced Curation Tools (ACT) would guide investigators in constructing metadata records for all the critical objects involved, including the investigation protocols, samples, assays, and generated data files. This construction would involve automated recognition of the temporospatial correspondence of vehicle, payload and experiment equipment data with sample, assay, and measurement data. ACT should also guide investigators in creating detailed, standards-based metadata as much as possible, pointing them to biomedical concepts that leverage community-based knowledge models [6].

Science investigators today face an ever-increasing demand in time and effort to clean, annotate, package and transfer investigation data to a burgeoning set of recipients and systems, with each having different requirements for all these processes. Many scientific publishing houses now have their own requirements for the public accessibility of data supporting article publications and some have specified means for data publications. Furthermore, researchers themselves often wish or need to share their data through community-managed data repositories, particularly in the field of life and physical sciences. Many investigators are taking advantage of the growing number of low-cost, publicly available data analysis platforms, and this requires them to transfer data into these destinations also.

To address these increasing demands on investigators time, ACT should support the delivery of investigation data to multiple recipients and partnering organizations easily (for example, by submitting an investigation's data at the request of an investigator both to a sponsoring organization's repository and a second, publishing house-specified repository). Such a service should only require that the investigator guide data annotation, packaging, and transfer to one of the desired destination repositories, and would then automatically provide any required data re-annotation and repackaging, as well as data transfer (copying) to the other.

Conclusions

Research sponsoring organizations that fund investigations in space life and physical sciences should invest in affordable, sustainable processes and tools to support the automated, early-, mid-, and late-lifecycle capture of investigation data into the open science repositories they manage. These organizations should also develop and offer to investigators cloud-based experiment workspaces as standard components of their data repositories. This could reduce the costs to investigators of designing or procuring reliable, accessible, and maintainable investigation-related data storage systems. As a result, sponsoring space research programs would have greater insight into investigation progress from dashboard systems that summarize use of these workspaces. Finally, the open science paradigm is likely to increase demands on investigators as they participate in more data sharing efforts, but this increase could be substantially lessened with the development by sponsoring research organizations of advanced data curation tools that assist investigators with standards-based data sharing.

REFERENCES

[1] Burwell, S.M., VanRoekel, S, Park, T, Mancini, D.J., Open Data Policy-Managing Information as an Asset, 2013, U.S. Office of Management and Budget. p. 1-12.
[2] National Academies of Sciences, E. and Medicine, *Open Science by Design: Realizing a Vision for 21st Century Research*. 2018, Washington, DC: The National Academies Press. 232.
[3] Williams, M., J. Bagwell, and M. Nahm Zozus, Data management plans: the missing perspective. *Journal of biomedical informatics*, 2017. 71: p. 130-142.
[4] Library, C.D. *DMPTool*. 2021; Available from: https://dmptool.org/.
[5] Kanbar, L.J., et al., Organizational principles of cloud storage to support collaborative biomedical research. *Annu Int Conf IEEE Eng Med Biol Soc*, 2015. 2015: p. 1231-4.
[6] Berrios, D., S.V. Costes, and D. Jameson, Semantic and linked rata retrieval. *Decadal Survey on Biological and Physical Sciences Research in Space 2023-2032*, 2021.