

Reference ID: 11220559147\_Fabbiano

---

**Reference ID:** 11220559147\_Fabbiano

**Submission Date and Time:** 12/13/2019 11:51:42 AM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

*Consent Statement:* "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

#### **Author Names & Affiliations**

Submitting author: Giuseppina Fabbiano - Center for Astrophysics | Harvard & Smithsonian

**Additional authors:** Bruce Berriman (Caltech/IPAC-NExSci); Raffaele D'Abrusco (Center for Astrophysics | Harvard & Smithsonian); Mark Lacy (NRAO); Francesca Civano (Center for Astrophysics | Harvard & Smithsonian)

**Contact Email Address** (for NSF use only): (hidden)

**Research domain(s), discipline(s)/sub-discipline(s)**

astrophysics

**Title of Response**

Data Infrastructure Challenges

#### **Abstract**

This input is submitted on behalf of the US Virtual Observatory Alliance (USVOA), a member of the International Virtual Observatory Alliance (IVOA). With the blossoming of time variability surveys, and

the detection of gravitational waves, Astronomy is facing a future where multi-messenger and multi-wavelength large-volume data will be the research norm for breakthrough discoveries. This landscape is representative of interdisciplinary research at large, as most of the challenges will be the same across disciplines: the comparison of multi wavelength imaging data; large date set comparison and ‘mining’; techniques and procedures for multi-wavelength follow-up of time alerts; and flexible, friendly data discovery and access. Related CI challenges include the production of calibrated interdisciplinary ‘science-ready’ data hyper-cubes, the establishment of data banks, ready access to data and speedy follow-up with new observations, and the enhancement of data, metadata, and registries to allow flexible, speedy and user-friendly searches across archives. Besides allowing for these developments, it is important to realize that interdisciplinary collaborations need to be supported (e.g. the USVOA - IVOA for astronomy), and long-term data hardware and software infrastructure curation funding will be needed. We suggest that NSF may also consider new interdisciplinary programs along the lines of the European ESCAPE.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

Astronomy is facing increasingly growing data volume and complexity. With the blossoming of time variability surveys and LSST soon to come, and the opening up of gravitational wave astronomy, we are looking at a future where multi-messenger and multi-wavelength large-volume data will be the research norm for breakthrough discoveries in astronomy. This landscape is also representative of interdisciplinary research at large, as some if not most of the technical challenges will be the same across disciplines. Examples are: a) Comparison of multi wavelength imaging data. This is certainly becoming increasingly more important in astronomy (e.g., multi-wavelength surveys such as COSMOS; comparison of ALMA/Chandra/HST/ground-based IFU datasets), and it offers similar challenges as comparison for example of earth science data, weather maps, crops, diseases etc. Interdisciplinary comparisons in astronomy, and beyond, require the creation of science-ready calibrated data sets, and of flexible and efficient data structures that will allow positional cross-matching, resolution matching, pixel-based image algebra, and a suite of image enhancing, yet scientifically valid statistical tools and simulations. b) Large date set comparison and ‘mining’. Given the data volumes, there will be a demand in the near future for copies of all the large astronomy datasets (LSST, Gaia, DESI... etc) to be co-located to allow rapid cross-matches, searches, and data mining across projects for objects with specific properties (e.g. PM from Gaia and colors from LSST). Something like that would definitely require a lot of cyber-infrastructure to support. c) The implementation across observatories of techniques and procedures for multi-wavelength follow-up of time alerts (transient electromagnetic events; gravitational wave events; neutrino surges; ...). Given the complexity inherent to scheduling both ground and space observatories, some intelligent conflict resolution method may be needed. Similar methods may apply to Earth-based events, or solar-flare-Earth follow-up. d) Flexible, friendly data discovery

and access. In astronomy there has been a movement in this direction with the development of IVOA (International Virtual Observatory Alliance) standards and data registries. These structures will need to evolve to be more dynamic and allow for wider and complex data searches, based for example on object type or other general characteristic. This suggests that the metadata associated with each data object may need to be enriched to allow for more transparent multi-disciplinary searches.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).** Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

The challenges discussed above will require several types of CI investigations and development. In particular: a) Interdisciplinary data comparisons require the production of calibrated interdisciplinary ‘science-ready’ data hyper-cubes (from real observations and simulated observations), to support the richness of data being produced by different instruments. This translates into: a. The implementation of calibration workflows at the data source or data center or to be run by users in some cases. b. While presently the data objects are instantiated into ‘files’ with accompanying metadata (FITS in astronomy), more flexible data structures will need to be created for multi-dimensional, inter-disciplinary data access and operations. New data access standards may be needed to access this information that go beyond the IVOA data-cube modeling presently being developed. c.

Spatial and temporal cross-matches in astronomy may be achieved with the IVOA Space-Time-Coordinates (STC) suite of data models, but application to Earth data will require extension to the IVOA STC models. d. Development of tool sets for interdisciplinary data comparisons and data mining will need to be considered, and may require either new approaches or extension of the approaches presently in use. b) If the large data volumes require data to be collocated for efficient analysis, data banks need to be established. The use of commercial cloud technology may be the more desirable approach, although non-commercial solution may also be considered. In any case, engineering studies and prototyping will need to be performed, to satisfy the present requirements and more important to allow for future scalability in both volume and complexity. These developments have both hardware and software implications. It is important to realize that this will not just be a one time effort, but it will require sustained long-term funding for infrastructure (hardware, software, data set) curation. c)

The follow-up of time alerts will require both ready access to the relevant archival data (e.g. multi-wavelength data at the possible position of the event in the case of astronomy), and possibly speedy follow-up with new observations to study and inter-compare the time evolution of the event in different spectral windows. A new challenge is therefore the scheduling of follow-up observations.

These may conflict with observatory scheduled operations, weather, technical constraints, and may require some intelligent operational software. d) How do we find the data? This is a challenge especially for interdisciplinary research. As astronomical data evolve into more complex and heterogeneous data structures, the infrastructure that makes these data searchable and supports the tools used to search and retrieve the data will have to change. New requirements will need to be addressed. a. Robust and flexible inter-disciplinary searches may need an enhancement of the data's informational content. This may be achieved with a continuous integration of heterogeneous data into a single data record (for example, a single astronomical observation fused with new data from other instruments), or continuous aggregation of similar data (i.e., stacking or co-adding of single observations into value-added entities). b. Metadata will also require enhancement. The high-level description of data resources that constitute the core of the interdisciplinary registries will require the inclusion both of metadata describing instrumental, spatial and spectral properties of each of the different contributors to the final heterogeneous data structure, and new metadata, describing the integration or "fusion" process. c. The description of data resources will need to change dynamically with the data. These are the disciplinary registries that provide an accurate representation of the data resources i.e. the virtual "rolodex" of data, curated and maintained by data providers.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

a) Collaborations need to be fostered and maintained. Interoperability protocols and standards will be needed for these future developments. This means that targeted collaboration/interoperability funding will have to be established and maintained. For example, in astronomy interoperability work is done under the umbrella of the IVOA collaboration. In this collaboration, there is a marked difference between members that received sustained funding (for example the European Union nations), and those that do not. In the US, the interoperability collaboration is the USVOA. Here again, members need to find their source of funding, which is given different emphasis in different data centers. The NASA archives, for example, are adequately supported, but the radio or optical ground-base funding varies depending on the needs of different activities. It is important that funds for these activities (including travel, meetings, and some prototyping/development efforts) be recognize worthwhile across the board. b) As discussed above, for these transformative activities to be successful there will also be need for long-term data hardware and software infrastructure curation funding, to maintain and evolve the hardware and software systems, tool production, maintenance & migration. In particular, there will be a demand in the near future for copies of all the large astronomy datasets (LSST, Gaia, DESI, SDSS V, Euclid, ...) to be co-located to allow rapid cross-matches and searches across projects for objects with specific properties (e.g. PM from Gaia and colors from LSST). Something like that would definitely require a lot of CI support. c) NSF may consider interdisciplinary programs along the lines of the European ESCAPE project (European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure; <https://projectescape.eu/>). Possibilities in the US may include a similar partnering or a

*Response to NSF 20-015, Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: 11220559147\_Fabbiano

---

partnering of Astronomy with Earth observation/Meteorology, disciplines that use large volumes of imaging data.

-- End Submission --