Snowmass2021 - Letter of Interest

The Need for a Cosmology Data Repository

Thematic Areas: (check all that apply □/■)

□ (CompF1) Experimental Algorithm Parallelization
□ (CompF2) Theoretical Calculations and Simulation
□ (CompF3) Machine Learning
■ (CompF4) Storage and processing resource access (Facility and Infrastructure R&D)
■ (CompF5) End user analysis
□ (CompF6) Quantum computing
■ (CompF7) Reinterpretation and long-term preservation of data and code
■ (CF) Cosmic Frontier (general)

Contact Information:
Submitter Name/Institution: Stephen Bailey / Lawrence Berkeley National Lab
Collaboration (optional): Dark Energy Spectroscopic Instrument
Contact Email: StephenBailey@lbl.gov

Authors:
Stephen Bailey, Lawrence Berkeley National Lab
David Schlegel, Lawrence Berkeley National Lab
Benjamin A. Weaver, NSF's National Optical-Infrared Astronomy Research Laboratory

Abstract:
There is a need within the cosmology community for a Cosmology Data Repository, which curates cosmologically useful datasets and simulations from across experiments and funding agencies, and co-locates these data at a computing center capable of jointly processing them. This would provide a method for archiving these datasets beyond the funding lifetime of individual projects, and, equally importantly, facilitate joint analyses of these data. A prototype repository used by the DESI Legacy Imaging Surveys exists at NERSC; broader coordination with the cosmology community and other data centers would maximize its usefulness for future projects.

1 Introduction

It is common within astronomy to release data publicly for the benefit of the entire community. NASA supports this through the Mikulski Archive for Space Telescopes (MAST)[1] and IPAC[2]; NSF's National Optical-Infrared Astronomy Research Laboratory (NOIRLab) provides data through its Astro Data Archive[3] and Astro Data Lab[4], in addition to NSF partnering with IPAC. These centers provide public data to the world, beyond the lifetime of the missions and experiments that generated the data. This meets the minimal federally mandated requirements for *archiving* data, but are insufficient for the computational *analysis* of these data — these data centers are primarily oriented toward downloading catalogs or small subsets of data, not for large-scale user-analysis of entire datasets at the data archive center, nor for providing bulk downloads to other computing centers for large analyses.

The Department of Energy (DOE), however, lacks a similar structure for curating data releases for DOE-funded cosmology experiments, as well as combining those data with external datasets for joint analysis. The result is an ad-hoc system where each project makes different arrangements for the long term preservation of its data, making it difficult to access and jointly analyse. At the user level, this can also result in the same dataset being downloaded multiple times to the same computing center, because users on one project were unaware that users from another project already had a copy of a 3rd party dataset needed by both.

DOE should invest in a cosmology-focused data archive hosted at a computing center sufficient for analyzing the data, not just downloading subsets of the data to elsewhere. This archive should include external datasets from NASA, NSF, the European Space Agency (ESA), the Sloan Digital Sky Survey (SDSS), and elsewhere that are useful for combining with DOE-sponsored data, and in the future should include copies of data releases from DESI, LSST-DESC, CMB-S4, and Euclid. Additionally, this data archive should host simulation datasets needed for the analysis of these data.

2 Prototype Cosmology Data Repository

A prototype of this vision is implemented as the "Cosmology Data Repository" hosted at the National Energy Sciences Research Computing Center (NERSC). It hosts 1.5 PB of images, spectra, and catalogs from SDSS, DES, Gaia, WISE, Galex, 2mass, and the DESI Legacy Imaging Surveys.

This repository was used by the DESI Legacy Imaging Surveys[5] to combine new and pre-existing public images from 3 ground-based telescopes plus the WISE satellite, as well as catalogs from at least 3 additional telescopes and satellites. These data were originally spread over multiple data centers, some of which only offered individual file http as a download method to access the 10s of millions of files needed. The Legacy Surveys performed a joint multi-band fit of O(10M) images to identify 1.6 billion objects as input to target selection for the Dark Energy Spectroscopic Instrument (DESI)[6]. This analysis for DESI would have been impossible without first curating the datasets co-located with computing resources capable of analyzing them.

[1] https://archive.stsci.edu
[2] https://www.ipac.caltech.edu
[3] https://astroarchive.noao.edu
[4] https://datalab.noao.edu
[5] https://www.legacysurvey.org
[6] https://desi.lbl.gov

3 Sharing data between computing centers

At the same time, we recognize that many people prefer their local computing center or university cluster and that it is unrealistic to expect a single data center to meet all needs of all users at all analysis scales. Data archive and computing centers should invest in technologies to easily and automatically replicate public datasets (and subsets) from one computing center to another ("bring the data to the computing"), as well as cross-site grid-like "bring the computing to the data" options. Data centers should be able to subscribe to datasets from each other and have them automatically replicated as needed using efficient transfer methods such as Globus[7], without individual end users having to initiate custom transfers in an ad-hoc manner.

4 Comparisons to astronomy and particle physics

Compared to astronomy, a cosmology-focused data repository is in some sense easier since fewer disparate datasets are needed. At the same time, to maximize the cross-project potential of future cosmology analyses, the multi-petabyte datasets need to be hosted at computing center(s) capable of providing many millions of CPU hours. Simply making the data available for download on a different website per experiment is no longer sufficient for many cosmology analyses using these data.

Compared to particle physics, more is to be gained in cosmology from combining datasets from different experiments, e.g. in different wavelengths, imaging plus spectroscopy, and multiple time epochs, as well as combining data taken by one project with simulations generated by another. Within particle physics, the CERN Open Data Portal[8] provides public access to public data from CERN experiments, though the authors of this LOI are unaware of an equivalent service for non-CERN data.

5 Summary

A Cosmology Data Repository hosted at a large computing center such as NERSC would maximize the usefulness of cosmology datasets by facilitating their joint analysis, while also providing a means for meeting the federally mandated requirements of archiving datasets beyond the funding lifetime of individual projects.

[7] https://globus.org
[8] http://opendata.cern.ch