Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921317002058

The WFIRST Science Archive and Analysis Center

Sara R. Heap[1,2], Alexander S. Szalay[3]
and the WFIRST Science Archive Team[†]

[1]Visiting scientist, Johns Hopkins University (JHU), Baltimore, MD, USA
email: sara.heap@Gmail.com

[2] Emerita Scientist, Goddard Space Flight Center (GSFC), Greenbelt MD 20771, USA
email: sally.heap@NASA.gov

[3] Dep't of Physics & Astronomy, JHU, 3701 San Martin Drive, Baltimore, MD 21218, USA
email: Szalay@JHU.edu

Abstract. The Wide Field Infrared Survey Telescope (WFIRST) is a 2.4 m telescope with a large field of view (∼0.3 deg^2) and fine angular resolution (0.11"). WFIRST's Wide Field Instrument (WFI) will obtain images in the Z, Y, J, H, F184, W149 (wide) filter bands, and grism spectra of the same large field of view. The data volume of the WFIRST Science Archive is expected to reach a few Petabytes. We describe plans to enable users to find the data of interest and, if needed, to analyze the data *in situ* using sophisticated software tools provided by the archive. As preparation, we are building a mini-archive that will help us to define realistic science requirements and to design the full WFIRST Science Archive.

Keywords. astronomical data bases, catalogs, surveys, WFIRST, methods: data analysis, galaxies: statistical, galaxies: evolution, infrared: galaxies

1. WFIRST

The Wide-Field Infrared Survey Telescope (WFIRST) was the top-ranked large space mission recommended by the 2010 Decadal Survey in its report, *New Worlds, New Horizons*. It has four scientific objectives:

- Answer basic questions about dark energy by surveying a ∼2200 sq. deg. area using a Wide-Field Imager (WFI) and an integral-field spectrograph (IFS);
- Complete a census of exoplanets by using the WFI to monitor a region in the galactic bulge in search of microlensing events;
- Detect and characterize exoplanets orbiting nearby stars using a high-contrast coronagraph;
- Support a guest-observer program making WFIRST observations of specific astronomical sources.

WFIRST is unique in having the high sensitivity of a 2.4m telescope combined with a large field of view (∼0.3 deg^2) and fine angular resolution (0.11"). As shown in Figure 1, the field of view of WFIRST's Wide Field Instrument (WFI) is *much* larger than those of the Hubble and James Webb Space Telescopes. A single WFI observation will contain

[†]Thomas Budavari, Gerald Lemson, Brice Menard, Alex Szalay (PI), Ani Thakar (JHU); Sara Heap, Thomas McGlynn (NASA's GSFC); Andrew Connolly (U. Washington); Jay Anderson, Michael Fall, Joshua Peek, Marc Postman, Swara Ravindranath, Gregory Snyder, Richard White (STScI)

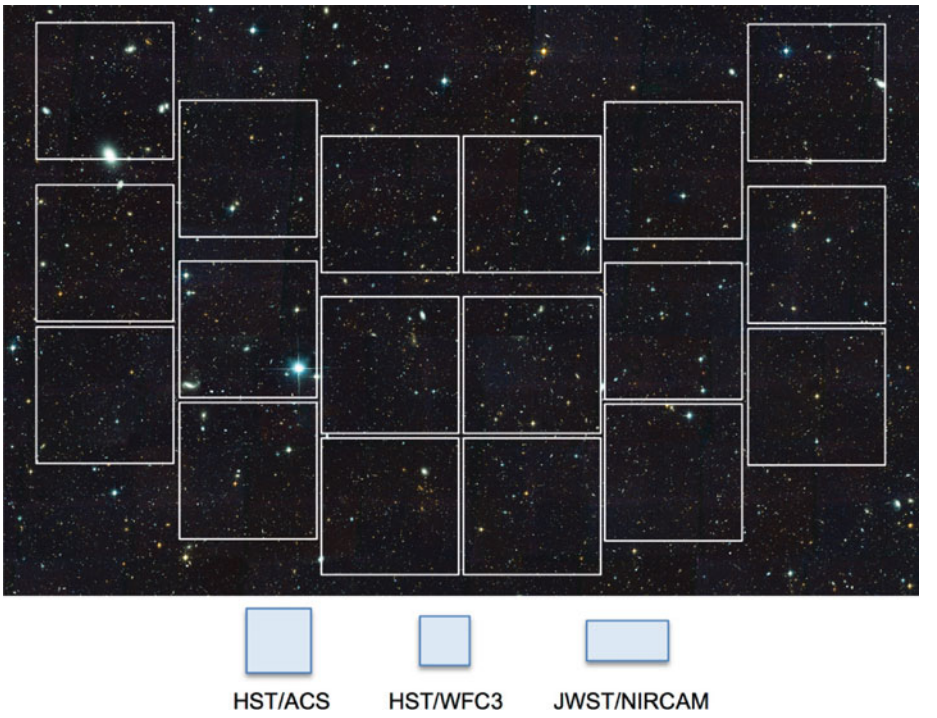

Figure 1. The Wide Field Instrument has eighteen 4K X 4K HgCdTe sensor arrays (288 million pixels, mapped to 0.11 arcseconds on the sky). (Figure reproduced from the 2015 WFIRST-AFTA SDT Report)

a million galaxies. As described in the 2015 *WFIRST-AFTA SDT Report* (Spergel *et al.* 2015), WFIRST will survey an area of ∼2200 sq. degrees to investigate properties of dark energy by:

- Measuring the large-scale structure of the universe out to nearly redshift 3
- Making improved measurements of redshift space distortions
- Conducting a deep-wide supernovae survey
- Measuring the shapes of 380 million galaxies

WFIRST will also impact a broad range of astrophysics. A sample includes:

- Making an accurate 3-D map of the distribution of dark matter
- Observing "Degree-Deep-Fieds" almost 100 times larger than Hubble deep fields
- Determining the precise positions and motions of $> 200M$ stars in our galaxy
- Finding counterparts to gravitational-wave merger events detected by LIGO
- Enabling studies of the properties of stars in the Milk Way and neighboring galaxies
- Generally enabling "cosmic origins" science (Heap *et al.* 2015)

Information on the Wide-Field Instrument (WFI) channels is given below.

- Wide Field Channel with very large FOV: 0.8° x 0.4°
  - Filter Bands: Z, Y, J, H, F184, W149 (wide band)
  - Grism: $\lambda\lambda 1.0 - 1.89\mu$ with 2-pixel resolving power, R=461 λ
- Integral Field Channel (IFC) Spectrograph
  - Supernova FOV: 3"x3" with 0.075 "/pix resolution
  - Galaxy Photometric Redshift Calibration FOV: 6"x6" with 0.15 "/pix resolution

https://doi.org/10.1017/S1743921317002058 Published online by Cambridge University Press

- Very high sensitivity over $\lambda\lambda 0.6 - 2.0\mu$
- Low spectral resolving power, R=70-140 $\lambda/\Delta\lambda$

In late 2015, NASA asked for proposals for WFIRST science investigation teams (SIT's), which together would compose a Formulation Science Working Group for the next 5 years. Each SIT would be responsible for developing scientific performance requirements and data analysis techniques, and for developing simulations and data challenges to validate the requirements and techniques. Our team, led by Alex Szalay, was selected to study the scientific capabilities of the WFIRST science archive (WSA) with a focus on galaxy evolution as a function of cosmic time, environment and intrinsic properties. As of October 2016, we have completed the first half year of our 5-year commitment.

2. The WFIRST Science Archive (WSA)

The usual way to access WFIRST science data and perform basic operations will be to query a catalog of measured properties of detected sources or selected regions, rather than work on the images or spectra themselves. This mode of operation calls for searchable catalogs containing a wide range of measurements. For galaxies, the catalog might contain such properties as source location, elliptical-aperture magnitudes in concentric, log-spaced apertures; Petrosian and Kron magnitudes and sizes; PSF-convolved, elliptical-model fits of exponential and de Vaucouleurs surface brightness profiles; two- and one-dimensional spectra of grism sources with line fluxes and equivalent widths of hydrogen and oxygen lines when present; photometric redshifts and spectroscopic redshifts (B. Robertson, priv. comm.).

We have adopted the "20 Questions" methodology of Gray *et al.* (2002), which has proven to be a powerful key to unlocking data in astronomical data archives. As recounted by Szalay (2008), "Gray asked about our 20 queries, his incisive way of learning about an application, as a deceptively simple way to jump-start a dialogue between him (a database expert) and me (an astronomer or any scientist). Jim said, 'Give me your 20 most important questions you would like to ask of your data system, and I will design the system for you.' It was amazing to watch how well this simple heuristic approach, combined with Jim's imagination, worked to produce quick results." We expect the "20 queries" methodology help us to discern the most important science analysis patterns and use these to establish the performance required of the WFIRST archive.

The 20 queries (Szalay *et al.* 2000) are broad in scope; they include not only questions involving source attributes but other questions such as: find all objects with unclassified spectra; find all star-like objects that are 1% rare; and find all objects within 1' of one another that have very similar colors - this last query being a search for gravitational lenses.

The original 20 queries were formulated for the Sloan Digital Sky Survey, but they are general enough that they are easily applied to WFIRST data. However, we are extending the original set of queries into take into account new capabilities not available in the SDSS era. One new capability is precise astrometry made possible by GAIA reference data. It should be possible, for example, to learn not only whether an x-ray point source is associated with some galaxy, but where in that galaxy. Another is the superb angular resolution of WFIRST (0.11"). This resolution should make it possible to examine the properties of the central kiloparsec of galaxies (usually where the action is) over a wide range in stellar mass and redshift. A third is the growing number of detailed cosmological simulations begging to be tested through comparison with observations.

https://doi.org/10.1017/S1743921317002058 Published online by Cambridge University Press

The projected data volume from WFIRST will be a few Petabytes, and simulation data may be equally large. A single storage system capable of holding this data volume will not be expensive by the time WFIRST is in orbit (∼2024). However, transmitting and replicating these data locally at a multitude of user sites will be quite expensive. Networks are the choke point. Hence, we will bring the algorithms to the data. We expect to provide enough computing power right on top of the archive to enable users to perform at least the first stages of data analysis in close proximity to the data. This set up will help for query execution over large catalogs, but will be even more important when users start running their own analysis scripts over the data. We are working to ensure that the WFIRST archive will have such capabilities. Thus, the WFIRST archive will be more than a simple passive file or database server. Instead, users will interact with it algorithmically though well-defined and well-designed Applications Programming Interfaces (APIs).

3. AstroInformatics

A major part of our study is to prototype a system capable of executing complex user-defined scripts including access to a shared computational facility with tools for object classification, large-scale cross-correlations, parallel Bayesian cross-matching of large object catalogs, scripting capabilities and machine learning, joining WFIRST to other surveys, and comparing WFIRST observations to physical models.

3.1. *Advanced Object Classification and "Feature Vector" Libraries*

Classification of astronomical sources is at the core of the WFIRST mission with applications ranging from the characterization of populations of transient, variable and moving sources, to the identification of unusual or anomalous sources, to the evolution of galaxies and their spectra. Numerous algorithms and techniques have been developed by the astrophysical community to address classification issues such as nearest-neighbor techniques, non-parametric Bayesian classifiers, kernel density estimation, neural networks, etc. An emerging approach is probabilistic classification rather than discrete assignment of a source to one type or another. Classification and machine-learning (ML) algorithms can be run within the archive, so we will investigate how advanced ML algorithms can be optimized for galaxy morphology and SED classification, and how best to represent uncertainty in classification within a database.

Machine learning (ML) methods typically work on features extracted from images ("feature vectors") rather than on the images of objects themselves. These methods reduce the feature vectors to a much smaller series of numbers representing aspects of the image, then reduce them further by some form of dimensionality reduction (e.g. PCA). The remaining information-rich vectors are then used in a learning algorithm. Astronomical feature vectors are also important for spectra. We plan to build and share two feature vector libraries for use with WFIRST data: one for kpc-scale galaxy morphology classification, and one for grism emission-line classification.

3.2. *Large-Scale Cross-Correlations*

Working directly with the source catalog enables data analyses based on one-point statistics, but the vast majority of extragalactic studies are based on two-point correlation functions, i.e. operations requiring the use of *pairs* of objects or attributes. An example of such operations is, measure the mean reddening of quasar sample Q, through angular cross-correlations with galaxy sample G. The ability to obtain such measurements without having to develop the necessary tools from scratch will enable scientists to better

https://doi.org/10.1017/S1743921317002058 Published online by Cambridge University Press

understand their data, test ideas, and interact with the data in new ways. Such early explorations of the data can be followed by more advanced statistical analyses requiring massive computations. Efficient cross-correlation methods have been developed, and members of our team have written various implementations running on different platforms. We will adapt and share these cross-correlation tools for use with WFIRST data.

3.3. *Parallel Bayesian Cross-Matching of Large Object Catalogs*

The most basic galaxy parameter, photometric redshift, cannot be estimated with high accuracy ($<0.05(1+z)$) from WFIRST data alone; it can only be accurately estimated in combination with LSST data, which provides photometry at bluer wavelengths (u,g,r,i,z,y). Analysis of combined LSST and WFIRST photometry is not simply a matter of cross-matching; non-detections in one or more LSST filters are just as informative in indicating a Lyman Break galaxy. Construction of a full spectral energy distribution (SED) spanning $0.3-2.0\mu$ and its use in estimating photometric redshift (with errors) will need to be carried out for hundreds of millions galaxies detected by WFIRST. Because of the importance of this problem, we will formulate a strategy for treating it in the context of the WFIRST mission. A key part of this strategy is SkyQuery, a system developed and hosted at the Johns Hopkins University. This system can respond within minutes even for searches across large archives with 500 million sources such as SDSS, GALEX, 2MASS or WISE. We will validate the performance of parallel Bayesian cross-matching for WFIRST by incorporating mock WFIRST catalogs into the existing system.

3.4. *Server-side Scripting and Machine Learning*

Most users will want to refine their queries and generate new hypotheses without the tiresome and time-consuming cycle of query, download, investigate, and repeat. Any software that allows for this has to be general enough to work on the majority of science questions, powerful enough to efficiently manipulate large WFIRST data sets, well enough designed to be learned quickly by new users, and advanced enough to work in browsers without any client side installation. STScI has successfully implemented this kind of interactive visualization through the MAST portal, where users explore data before downloading by generating scatter plots, spectra, and time series. Goodman and colleagues (Baumont, Goodman & Greenfield, 2015) are developing a Python-driven code called `glue` which allows for fast, expandable visualization of tabular and image data simultaneously (c.f. `glueviz.org`). We plan to adopt the `glue` interface for the WFIRST Science Archive, so that users can easily explore large data sets on multiple dimensions simultaneously.

4. A Mini-Archive as Pathfinder

As a first step toward prototyping the WFIRST archive and analysis system, we are building a mini-archive based on the COSMOS field. The COSMOS field (∼2 sq. deg.) covers about $1/1000^{\rm th}$ of the WFIRST survey area. The field is located at RA$\sim 150^\circ$ close to the equator so that telescopes in both northern and southern hemispheres can (and have) made observations of it. These ground-based telescopes include Keck, Subaru, Very Large Array (VLA), the ESO-VLT, UKIRT, NOAO, and CFHT. The COSMOS field has also been observed by space telescopes including Hubble, Spitzer, GALEX, XMM, Chandra, Herschel, and NuStar. More information on the COSMOS survey can be found at `cosmos.astro.caltech.edu`, and a mosaic of the COSMOS field can be viewed with COSMOS SkyWalker (`www.mpia.de/COSMOS/skywalker/`).

We chose the COSMOS field for the mini-archive because it is so well studied, the field is big enough that statistical analyses can be made, and Hubble/WFC3 IR images and

https://doi.org/10.1017/S1743921317002058 Published online by Cambridge University Press

slitless spectrograms of the field are similar to the observations that WFIRST will obtain, since the two telescopes have the same aperture (2.4m) and use the same detectors (HgCdTe). The main differences are in the IR spectra: WFIRST spectra extend to $\sim 1.9\mu$ in order to include Hα out to redshift, $z = 1.9$, whereas Hubble/WFC3 spectra extend only to the H-band; and WFIRST spectra have a higher spectral resolution (R=461λ) than does Hubble in order to obtain the precision needed for Baryonic Acoustic Oscillation (BAO) studies of dark energy.

The core holdings of the mini-archive comprise Hubble/WFC3 JH images (F140W) and G141 slitless spectrograms covering a small portion (0.05 sq. deg.) of the COSMOS field. The spectra cover $1.1 - 1.65\mu$ at a resolving power, R$\sim$ 130. These data were obtained by the 3D-HST team (van Dokkum *et al.* 2011), which derived and made public high-level data products from their observations (Momcheva *et al.* 2015) as well as deriving useful scientific results (`3dhst.research.yale.edu/Publications.html`). All these data from the 3D-HST program – observed & calibrated data, catalogs, output products like image cut-outs of the 2-D spectra – will be ingested by the mini-archive. These data will be supplemented by Hubble/WFC3 images of the same field in the J and H bands.

There is some discussion about including a bluer filter than is presently in the WFIRST filter set, so the mini-archive will ingest Hubble/ACS F814W image data obtained by the CANDELS consortium. But even if a "blue" filter is included, the spectral range of WFIRST will still be inadequate for obtaining the broad spectral energy distributions (SEDs) needed to estimate photometric redshifts of detected objects. WFIRST is thus highly dependent on the Large Synoptic Survey Telescope (LSST) to supply u,g,r,i,z,y photometry. Fortunately for the mini-archive, useable optical photometry of the COSMOS field has already been derived for most of the LSST passbands and included in the COSMOS2015 catalog (Laigle *et al.* 2016). This catalog contains measurements of over 500,000 objects in the COSMOS field obtained by a myriad of telescopes observing in hard X-rays to radio wavelengths. The mini-archive will ingest this catalog as well. The catalogs of the 3D-HST team and the COSMOS2015 catalog will be combined and converted into a searchable, relational database.

We expect that the mini-archive capable of basic functions (search, display, download) will be ready for the WFIRST Science Investigation Teams to try out and critique in the spring of 2017. Work on the algorithmic and analysis sides of the mini-archive will proceed thereafter.

References

Beaumont, C., Goodman, A., & Greenfield, P., (2015), *Proc. ADASS XXIV*, ASPC 495, 101
Gray, J., *et al.* (2002), http://arxiv.org/ftp/cs/papers/0202/0202014.pdf
Heap, S., *et al.* (2015), http://cor.gsfc.nasa.gov/sags/`SAG8_Final_Report.pdf`
Laigle, C., Capak, P., & Scoville, N. (2016), *ApJS* 224, 24
Momcheva, I., *et al.* (2015), *ApJS* 225, 27
Spergel, D., Gehrels, N., *et al.* (2015) http://arxiv.org/abs/1503.03757v2
Szalay, A. (2008), http://cacm.acm.org/magazines/2008/11/549-Jim-Gray-astronomer/fulltext
Szalay, A., *et al.* (2000), Proc. ACM SIGMOD 2000, p. 451
Van Dokkum, P., Brammer, G., Fumagalli, M., *et al.* (2011) *ApJ* 743, L15