Advanced Data Products for the Next Decade

S. Casertano, J. Anderson, T. Brown, M. Stiavelli, R. L. White (STScI);
G. Fabbiano, I. N. Evans, J. McDowell, A. Rots (CXC);
D. Schade (CADC); K. D. Borne (George Mason University);
B. F. Madore (IPAC); P. Rosati (ST-ECF)

March 31, 2009

The impact of archival research

The amount and quality of research carried out using archival data has been steadily increasing over the last decade (see, e.g., the statistics presented in White et al. 2009, submitted to Astro2010), thanks in no small measure to the improved quality of the data that most projects and missions make available to archival researchers. Within the vision of the Virtual Observatory, advanced data management techniques will allow the integration of globally distributed data collections and services, putting a vast array of data and products within easy reach of researchers everywhere. The development of high-quality advanced data products is a central component of that vision. Over the next decade, Archive Centers—in cooperation with the Virtual Observatory—should expand their involvement in the development, validation, and dissemination of advanced data products for their respective missions, identifying the tools and data that offer significant promise of broad applicability, and developing them to the extent needed to make the results accessible to the whole astronomical community.

Survey projects, such as IRAS, HIPPARCOS, SDSS, and WMAP, have traditionally provided archival users with high-quality processed data and software tools that greatly facilitate their research. The quality of such products generally is as good as it can be achieved, since the processing is carried out by the most experienced people and the data are homogeneous. Pointed missions, such as today's Great Observatories, face a more difficult challenge because of the heterogeneity of the data they produce; PI teams can often achieve higher quality for specific data sets, thanks to fine-tuned processing and

quality control. Traditionally, these missions have focused primarily on producing the highest quality single-exposure calibrated data, placing a lower emphasis (at least for HST) on combined images and ancillary products. In some special cases, high-quality combined images for very deep surveys have been produced and made available to the community; examples include the seminal Hubble Deep Field (Williams et al. 1996, AJ 112, 1335) and the following Hubble Deep Field South and Ultra-Deep Field for HST, and the Chandra Deep Fields (e.g., Zezas et al 2006, ApJS 166, 211). For other special data sets, combined images have been produced by the PI teams (e.g., HST Treasury and Spitzer Legacy programs) and distributed through the respective Archives. Similarly, ESO now asks PIs of Large Programmes to return their Advanced Data Products to be ingested in the public ESO Archive in the form of science-ready data (both imaging and spectra) with a homogeneous set of VO-compliant metadata.

Evolving Archive collections

In the last few years, mission archives have substantially expanded the type and quality of data products that go beyond single-image calibration, such as combined images, extracted spectra, and source catalogs.

Recently, CXC has released the first version of the Chandra Source Catalog, which includes information on point-like (or quasi point-like) sources detected in the public Chandra data. Besides providing access to databases of source information, the Catalog provides to users a set of file-based data products to facilitate science-specific analyses of the observations and sources included in the catalog. The Chandra catalog is expected to be used by a broad group of scientists, not just X-ray astronomers, including those who may lack a detailed understanding of Chandra data or who may be unfamiliar with astronomical data analysis in the X-ray regime. A primary goal of the catalog is to provide file-based data products that are generally useful for a wide range of science studies, are familiar and readily understandable by users from any wave band, and that rigorously meet the needs of those users who routinely analyze Chandra data. A secondary goal is to provide data products that may be difficult for the end-user to prepare at their home institution, or that may be significantly computationally expensive to reproduce. CXC is also developing an automated capability for coadding observations, and anticipates making more high-quality multi-observation datasets available to users. Ancillary information, such as noise images and/or sensitivity maps, is a necessary complement to such datasets to enable their use in multi-wavelength and temporal analysis.

For HST, the Hubble Legacy Archive (HLA), a collaboration of STScI, CADC, and ST-ECF, now offers much of the HST data in a form that combines multiple exposures obtained within the same visit, together with multi-band source catalogs suitable for

point and extended sources. Similar information is also available from the Spitzer Science Center as part of their automated pipeline. The HLA also offers extracted source spectra from NICMOS GRISM data, and will soon extend this offering to ACS GRISM data.

Such products are still new, and—at least for HST—some researchers choose to carry out specialized combinations with slightly different options and independent quality control; it is a goal of the HLA to understand how the combination and quality control can be carried out so as to make their products useful to the majority of the users.

Further enhancements will become commonplace with the cross-linking of information across missions and the development of Virtual Observatory capabilities. For example, the Chandra, HST, and Spitzer archives are in the process of sharing their footprint information, thus enabling one-stop federated archive searches; the increasingly common linking of data to papers that present them in the literature offers additional semantic search capabilities; and cross-correlation of catalogs from different missions, including ground-based products, will eventually expand to encompass all missions that make such products available.

The need for more advanced data products and tools

Combined data, source catalogs, and extracted spectra, together with the necessary ancillary data and metadata needed to search, access, and assess the information seamlessly, are essential resources for archival researchers. These products represent a clear improvement over what was available only a few years ago, yet community-based efforts—both PI teams and independent researchers—have proven that there is much more potential hidden in the enormous quantity of high-quality data available in these archives.

Better information and characterization of the data can open up new avenues of research and analysis. For example, Anderson and King (2000, PASP 112, 1360) have shown the improvements that can be obtained in both photometry and astrometry of point sources by using more complex algorithms that fully characterize both the effective point-spread function and the geometric distortion of HST's cameras as part of the analysis process. Laidler et al (2007, PASP 119, 1325) have shown how the spectral energy distribution of distant galaxies can be reconstructed from panchromatic data with wildly varying angular resolution across the spectrum by using image fitting techniques. Rhodes et al. (2006, HST Calibration Workshop) illustrate how the characterization of the time-dependent HST point-spread function is necessary in order to enable weak lensing analysis of wide-field imaging data (Massey et al. 2007, Nature 445, 286). Brown et al. (2009, AJ 137, 3172) identify the need for special calibrations and procedures in order to characterize multiband stellar photometry and extract metallicity and age information to the full extent the inherent accuracy of the measurements allows. Several

of these examples are described in more detail in the Appendix; more can be found in Ferguson et al (2009, submitted to Astro2010). Research groups have developed techniques and distributed tools that enable users to extract the necessary information; however, their use typically requires time and expert knowledge, and in most cases they have not been widely used except for the original data for which they were developed.

Advanced data products and tools will become even more important with the increasing size and complexity of future projects such as JWST, JDEM, Pan-STARRs, and LSST. For example, JWST will largely build upon the experience of HST, but will present special challenges because of its active primary, which can potentially produce slightly different PSFs each time it is realigned. In addition, the JWST instruments, particularly the spectrographs, will produce more complex data than the previous generation of space instruments. The integral field units in MIRI and NIRSpec and the NIRSpec microshutter-based multi-object spectrograph will require particular care in their reduction and analysis. Handling and visualization of the data cubes, spatial and wavelength resampling and optimal extraction for point sources or compact objects are some of the issues that need to be addressed. It is essential for the success of JWST that high quality data sets are distributed to the users and to the archive and this might require input from the community. One idea to gather this input is that of carrying out data challenges where realistic simulated data are distributed to the community so as to enable them to begin developing tools even before JWST is launched.

The role of Archive and Science Centers

These examples illustrate how advanced data products, characterization, and tools can put within easy reach of the whole astronomical community a broad range of research projects that take advantage of the wealth of data currently available. Formation, enrichment history, and dynamics of tens of globular clusters in the Galaxy; population synthesis and star formation history of tens of thousands of galaxies to $z = 2$ and beyond; gravitational lensing mass mapping of individual halos, groups, and clusters in a wide range of parameters; systematic study of active galaxies and their nuclear properties in the optical, IR, and X-ray; and many other projects would become accessible to individual researcher and small groups. The range of questions that can effectively be asked will be limited by the intrinsic quality of the data, not by the work needed for each group, independently, to extract and collate the information, working through a very large amount of inhomogeneous data, requiring a variety of tools and expertise to understand and characterize. Critical elements of these enhanced data collections include: complete, reliable source catalogs covering all the data, including time-resolved catalogs to identify possible variability; exquisite characterization of point-spread function and

instrument response function, including variability due to observational and instrumental parameters; and the ability to correlate and search data from different missions, including all of the ancillary information and metadata needed for their interpretation.

How can we move towards such a scenario? There is no shortage of brilliant ideas and novel techniques to develop better tools that can exploit the precision and stability of our Observatories. NASA and other funding agencies have long supported, through archival research funding and other dedicated programs (e.g., ADP), the development of advanced techniques and data analysis procedures *aimed at answering specific scientific questions*. In many cases, these techniques can be generalized and applied to a much broader range of data than they were originally developed for; in practice, as shown by the above examples, the original researchers focus primarily on the specific target of their program, and this potential often remains untapped.

If given the right mandate (and the necessary resources), archives could become the focal point of these efforts. Archive centers already encompass much expert knowledge on their respective missions, and contain the type of scientific and technical knowledge to serve as a bridge between the researchers who develop advanced tools for their own use, and the broader community that can benefit from an expanded application of such tools. Identifying and selecting the data and tools that have the most promise for discovery and that benefit the community the most; testing their viability, accuracy, and generality; and distributing them in readily usable form are all tasks for which science and archive centers are ideally suited, and that they carry out successfully today. As an added benefit, many of these techniques have elements of self-calibration, and their generalized application to a large quantity of data can greatly improve our understanding of the regularity of the data themselves; for example, the use of the effective PSF approach on several different projects has shown that only a handful of stars are needed to fully characterize the point-spread function for a specific data set, taking into account time-variable properties of the observation (Anderson, private communication).

We envision a push towards a new generation of advanced data products and characterization tools that will make practical for researchers to realize the full potential of the wealth of panchromatic, high-resolution, high-accuracy data our observatories are collecting and will continue to collect in even greater amounts. Archive Centers can play two key roles in realizing this vision. First, they will identify community-provided products and tools developed for specific science goals and build on them, refining, documenting and supporting their wider use, and—where appropriate—applying them in a wholesale way so as to bring the entire archive collection up to the same standard of product as the original PI-based effort. Second, they will ensure that the products thus obtained contain all the necessary ancillary information and metadata to ensure their easy accessibility and interoperability through the Virtual Observatory.

Appendix: Some examples of advanced data products and tools

In the following we provide a few signal examples of data products and tools developed over the last few years, and whose widespread application can greatly enhance the ability of archival researchers to extract qualitatively better information from the vast amount of data now available in the archives of major missions. Often, such tools are developed for a specific application, to enable a focused research project, or as part of the work of PI teams on major proposals or of Instrument Scientists attempting to better understand their instruments. Many such tools are especially applicable to space-based platforms, as their intrinsic stability allows a deeper level of analysis to improve the understanding and characterization of instrumental effects, and to remove their signatures; but similar techniques have been developed for ground-based observations as well, and will become even more commonplace with the advent of a new generation of ground-based survey instruments (Pan-STARRs, LSST; see for example Ivezić et al. 2008, astro-ph/0805.2366). In addition, JWST, with its well-measured but most likely time-dependent PSF, and the array of on-board diagnostic tools (pupil imaging, in-focus and out-of-focus wavefront sensing), will provide an a more complex set of challenges with potentially bigger gains if the time-dependent PSF effects can indeed be fully calibrated.

Example #1: Panchromatic photometry with multi-resolution data

Several major surveys in the last decade have been dedicated to obtaining imaging and spectroscopy data over a broad wavelength range, in order to characterize the formation and evolution of galaxies at high redshift (e.g., SWIRE, Lonsdale et al. 2003, PASP 115, 897; GOODS, Giavalisco et al. 2004, ApJ 600, L93; GEMS, Rix et al. 2004, ApJS 152, 163; COSMOS, Scoville et al. 2007, ApJS 172, 1; DEEP2 and associated data, Davis et al 2003, SPIE 4834, 161) and the stellar population and structure of nearby galaxies (e.g., SINGS, Kennicutt et al, 2003, PASP 115, 928; ANGST, Dalcanton et al. 2007, AAS 211, 79.05). As a result, there is now a wealth of data covering relatively large areas of the sky—from tens of square arcmin to square degrees, depending on depth—over wavelengths ranging from from the radio to the X-rays. The availability of such data has changed how some of the research is carried out in areas that require large data sets; a new project no longer needs to collect all the necessary data, but often can leverage existing data and only supplement them with a small amount of data needed to cover the gaps.

Major observatories provide detailed information about the basic calibration and processing of the data they distribute, and high-level products are routinely produced

either by the original research team (delivery arrangements are a common requirement for funding) or by the archives and science centers themselves. However, the challenge of combining multi-wavelength, multi-resolution data in an optimal way remains daunting, especially when dealing with sources (galaxies at high redshift, stellar systems and star-forming regions in nearby galaxies) that may be fully resolved at some wavelengths, unresolved and possibly blended at other wavelengths.

One approach to taking best advantage of such data is to use the high-resolution images in the visible and near-infrared, together with a spatially resolved stellar population model, to match the low-resolution images in the mid-IR (or in any other passband that may be available). This approach, called TFIT (Laidler et al. 2007), requires a thorough understanding of the peculiarities of the data, the point-spread function appropriate to each dataset (which itself depends not only on the instrumental PSF, but also on the method used to combine the data), the relative alignment of the images at different wavelengths, and the throughput function in each image. As a consequence, its application is very labor-intensive, setting a high barrier to entry.

With the appropriate resources and priorities, archives and science centers could make TFIT (or equivalent methods) far more immediately accessible to the astronomical community, by 1) further developing and generalizing the procedures to make them more of a push-button operation, and 2) especially by characterizing the properties and maintaining the quality of the data they produce to the more exacting standards required by these programs. Archives will also ensure that the products and tools conform to community-wide standards and contain standard metadata. This is especially important in a multiwavelength, multipurpose context, for which data have to be thoroughly annotated, making explicit parameters that are implicitly assumed within a particular wavelength or analysis domain. The framework provided by the Virtual Observatory Characterization Schema allows the description of a dataset's spatial, spectral and temporal calibration and reliability, and its Spectrum Schema which defines the metadata needed for interoperable usability of spectra and SEDs.These schemas allow for the fact that different instruments and archives may represent calibrations at varying levels of fidelity—e.g., for one instrument a single number may represent the spatial resolution, while for another a whole library of position and wavelength dependent point-spread-functions may be provided. When an astronomer accesses the archives, the VO standards allow them to determine what calibrations are available for a dataset and at what level of fidelity, in the same format for every archive.

Example #2: Ultra-precise photometry and astrometry for dense stellar regions

Accurate photometric and astrometric measurements for stellar populations in the nearby Universe are at the basis of understanding their formation history, age, and dynamics. Typically, standard procedures—such as aperture and PSF-fitting photometry and moment centroiding—can achieve an accuracy of about 0.01 mag and 0.05 pixels on well-exposed, moderately crowded HST images. (Their performance degrades rapidly with crowding and, for centroiding, for very undersampled data, such as WFPC2 images.) More accurate procedures, such as those developed by Anderson and King (2000), have demonstrated the capability of achieving photometry with relative precision of a few millimag, and relative astrometry with errors well below 0.01 pixels. Such precision opens the field to new discoveries, such as the multiple main sequences—indicative of a non-monolithic formation history—in the most massive globular clusters (Bedin et al 2004, ApJ 605, L125; Anderson et al 2009, ApJ, submitted; see review in Piotto et al 2009, IAU Symp 258), and the possibility to study systematically the internal and external dynamics of individual globular clusters through proper motion studies (McLaughlin et al. 2006, ApJS 166, 249; Anderson and van der Marel 2009, in prep.). In the most favorable cases, a direct detection of the parallactic motion of globular clusters is not out of the question (Anderson and Bedin, in prep.), although this analysis may be at the limit of what is achievable through automated methods.

The methodology developed by Anderson and King is based on a careful determination of the effective point-spread function (ePSF) and of the geometric distortion for a field observed multiple times over several years. In principle, both the PSF and the geometric distortion can vary with observation; under optimal circumstances (a single field observed multiple times at different epochs and orientations), the method is applied iteratively until a satisfactory multi-parameter solution is found. The technique is in principle straightforward and the necessary procedures are freely available, but its application depends on the type of data—rich stellar fields are optimal and permit the most accurate results—and at present it requires expert judgment by the user. As a result, this technique has not been widely used to date except by the inventors and their collaborators. The Anderson and King approach has been developed for HST data, but a similar approach could well be devised for other missions with similar types of imaging data.

As this procedure is applied to more data, we are achieving a better understanding of both the technique and the properties of the data themselves. It now appears that most of the advantages of the ePSF technique can be realized without an iterative solution, and thus are applicable to a broader set of data than originally thought. Now that a library

of point-spread functions has been established for several instruments, a few bright stars are generally sufficient to determine the ePSF applicable to a specific data set, and the time and wavelength dependence of the instrumental geometric distortion has been well characterized. An accurate ePSF improves not only the photometry and astrometry for point sources, but also the ability to distinguish reliably between point and extended sources, and the characterization and fitting of the properties of the latter. Thus, while optimal data sets allow a more precise solution, as well as providing an intrinsic self-calibration of the procedure through quantification of the residuals, the procedure itself is applicable to a large degree to the vast majority of the HST data.

In this case, we can envision the relevant Archive or Science center obtaining the necessary expertise to apply the Anderson and King technique (or a functional equivalent) to a large fraction of suitable data in their collection, far exceeding those for which it has been applied to date. The immediate benefit for archival researchers will be direct access to very high precision catalogs based on this method. In addition, routinely applying this type of analysis to a much larger set of data will uncover systematics that will introduce additional constraints (e.g., PSF properties relating to known parameters of the observation) and thus further enhance its applicability.

Example #3: Extracting stellar parameters from high-accuracy multi-band photometry

At the present time, it is difficult to conduct investigations of resolved stellar populations using archival data that spans multiple observing programs. For example, while the Hubble Legacy Archive is automated to the point of producing photometric catalogs, these catalogs are nowhere near the fidelity of those produced by the PI teams of the individual programs. Even if one could collect those individually-produced catalogs in a common location, an analysis of the assembled data would be hampered by systematics inherent to the reduction process employed by each team and distinctions in the particular instruments and bandpasses used. The first challenge, improving the quality and reliability of photometric measurements for individual sources, can be met by advanced photometric methods that take into account the instrumental properties of HST, such as those presented in Example # 2.

In order to meet the second challenge, that of providing a high precision, empirical calibration of the photometric system to the extent required to fully quantify the properties of well-observed stellar population, Brown et al. (2009) have developed a two-part strategy for WFC3, soon to be installed aboard HST. First, Brown and collaborators introduced a new photometric system employing five WFC3 bands spanning the UV, optical, and near-infrared: F390W, F555W, F814W, F110W, and F160W (analogous

but not identical to the ground-based filters Washington C, V, I, J, and H). With these bands, one can construct reddening-free indices of temperature and metallicity. Second, they will obtain deep photometry in six fields in well-studied star clusters, spanning a wide range of metallicity, and four fields in low-extinction windows of the Galactic bulge. The cluster data serve to calibrate the reddening-free indices, provide empirical population templates, and correct the transformation of theoretical isochrone libraries into the WFC3 photometric system. The bulge data will serve as empirical population templates for other studies. These data, part of an HST Treasury program, will provide an ideal calibration set for this photometric system, and enable observers targeting other stellar systems to place their results in a well-understood observational and theoretical context.

Improved catalogs and photometric calibrations will enable a broad range of scientific investigations of stellar populations in nearby galaxies. We outline here two such examples.

The color of the red giant branch (RGB) is a sensitive tracer of metallicity in a stellar population. In a reasonable HST observing program (100 orbits or less), one can obtain color-magnitude diagrams (CMDs) of the brightest two to three magnitudes of the RGB at distances of up to $\sim$ 10 Mpc (e.g., Mouhcine et al. 2005, ApJ, 633, 828). Many of the nearby galaxies have already been imaged to sufficient depth. An ideal catalog of such observations for nearby galaxies could explore the metallicity of the stellar population as a function of environment—both as a function of galaxy size/morphology but also as a function of location within galaxies. The result would provide critical observational constraints for theoretical models of chemical evolution in galaxies.

A CMD offers several standard candles, such as the tip of the RGB and the red clump on the horizontal branch. If the archive provides catalogs of both the coadded deep images in a field but also the time series of images in a field, then variable stars provide additional standard candles (e.g., Cepheids, RR Lyrae stars). Because many nearby galaxies are blanketed with observations from multiple imaging programs, a comprehensive catalog of these images can provide the three-dimensional structure of these galaxies. When combined with radial velocities in these populations (e.g., Gilbert et al. 2007, ApJ, 668, 245), we get a snapshot at one point in the structural history of a galaxy, which will provide robust anchors for simulations of this history. Ambitious simulations are currently attempting to demonstrate how recent satellite mergers can reproduce much of the substructure in nearby galaxies (Fardal et al. 2007, MNRAS, 380, 15).