
**Proceedings for the 26th Annual Conference
of the Society for Astronomical Sciences**



Symposium on Telescope Science

**Editors:
Brian D. Warner
Jerry Foote
David A. Kenyon
Dale Mais**

**May 22-24, 2007
Northwoods Resort, Big Bear Lake, CA**

Reprints of Papers

Distribution of reprints of papers by any author of a given paper, either before or after the publication of the proceedings is allowed under the following guidelines.

1. The copyright remains with the author(s).
2. Under no circumstances may anyone other than the author(s) of a paper distribute a reprint without the express written permission of all author(s) of the paper.
3. Limited excerpts may be used in a review of the reprint as long as the inclusion of the excerpts is NOT used to make or imply an endorsement by the Society for Astronomical Sciences of any product or service.

Notice

The preceding "Reprint of Papers" supersedes the one that appeared in the original print version

Disclaimer

The acceptance of a paper for the SAS proceedings can not be used to imply or infer an endorsement by the Society for Astronomical Sciences of any product, service, or method mentioned in the paper.

Published by the Society for Astronomical Sciences, Inc.

First printed: May 2007

ISBN: 0-9714693-6-9

Selective Availability of Astronomical Data

P. R. McCullough
Space Telescope Science Institute
3700 San Martin Dr., Baltimore MD 21218
pmcc@stsci.edu

Abstract

We discuss potential advantages and disadvantages of selective availability of astronomical data. Selective availability would enable prompt dissemination of data to the community at large by maintaining the proprietary nature of only selective characteristics of the data to only a select set of persons. For the purpose of illustration, we present a simplistic example of an algorithm that modifies a two-dimensional image such that it corrupts stellar photometry of specific (bright) stars while preserving other characteristics such as stellar astrometry. We advocate that the astronomical community already could benefit from selective availability, and we suggest that the need will increase in time as ever-larger volumes of data are collected but not disseminated in a timely fashion for lack of appropriate algorithms to create selective availability.

1. Introduction

The term “selective availability” (SA) refers to the intentional masking of certain characteristics of data to be made publicly available while retaining those characteristics for authorized users, typically by some form of encryption. For example, in the U.S. global positioning system (GPS), pseudo-random deviations were introduced into the satellite signals to reduce the accuracy of positions on Earth, i.e. longitude and latitude, available to those users that lack the knowledge of the pseudo-random deviates. The SA for the U.S. GPS system was de-activated on May 1, 2000, because the advantages of doing so outweighed the disadvantages.

(http://www.ngs.noaa.gov/FGCS/info/sans_SA/)

Civilian users of GPS would benefit by the ~10-times better precision of positions without SA compared to with SA, while the US could still deny GPS on a regional basis as necessary for national security using alternatives to SA.

In this paper, we introduce “Alice” as an astronomer with rights to data in its original full-fidelity form (hereafter, O-data), and “Bob” as another astronomer who wishes to have access to the O-data in some form (hereafter, SA-data) that is useful to him and agreeable to Alice. The SA concept we propose here enables Bob’s access; otherwise, Bob would have access only at Alice’s discretion or alternatively he could wait until Alice’s proprietary period expires.

Selective availability generally is not applied to astronomical data; typical practice is described in this

paragraph. Availability is accomplished with a single bit: either a person has access to data in all its intrinsic fidelity, or the person does not have access to the data at all. The bit is active for a so-called proprietary period, which begins when the data are obtained and extends for a period of time specified in the observing proposal. The nominal proprietary period is 18 months, which allows an investigator to analyze the data and potentially obtain additional data in a second observing season, if necessary, prior to publication.

The proprietary period can be longer or shorter than its nominal value, depending on circumstances. For example, data may be collected over an extended period of time to enable the intended science goals. For example, NASA’s Kepler mission requires at least three years of observing to confirm multiple transits of extrasolar Earths, so its investigators may justifiably request to retain their proprietary rights for at least that duration. As another example, spectra obtained at the Keck observatory for the purpose of radial-velocity detection of extrasolar planets have been archived in a form where the section of the spectra with iodine absorption lines (used as radial velocity fiducials) has a longer proprietary period than the rest of the spectra, in order to allow the principal investigators time to see oscillations in the radial velocities associated with planets that can have many-year periods (Beichman *p.c.*; see also <http://www2.keck.hawaii.edu/koa/public/koa.php>).

The XO Project, and others like it, aim to discover transiting planets via photometric monitoring of hundreds of thousands of stars and have accumulated multiple terapixel sets of images of significant fractions of the sky that could be useful for a number

of other fields of astronomy (McCullough et al. 2006 and references therein). However, only a small fraction of those images have been released to the public domain, because the teams that have labored to produce the data continue to analyze them to find additional transiting planets. At the other extreme, data have no proprietary period; for example, observing time with major observatories that is awarded via Director’s discretion traditionally is afforded zero proprietary period – the Hubble Deep Field and the Ultra Deep Field are examples.

If “Alice” could apply a SA algorithm that would corrupt all stellar photometry by 3% rms, she could be confident that releasing the data in that form would retain to herself the ability to detect planetary transits which typically are less than 2% in depth. Another astronomer, “Bob” could use the XO data in its SA-form for other purposes, such as studying larger-amplitude variable stars, detecting moving objects such as near-earth asteroids or comets, studying diffuse objects such as galaxies, zodiacal light, or artificial light pollution, or studying rare transients such as optical counter parts to gamma-ray bursts (e.g. Paczynski 2006).

For the purpose of illustration, we present a simplistic example of an algorithm that modifies a two-dimensional image such that it corrupts stellar photometry of specific (bright) stars while preserving other characteristics such as stellar astrometry. Trivial algorithms for SA of tabular data can be effective also.

2. Lossy Compression as a SA Algorithm

The desirable property of lossy compression (e.g. HCOMPRESS, White, Postman, & Lattanzi 1992) usually is the reduction in file size in bytes for a given image size in pixels. For a typical image from the XO survey, HCOMPRESS increases noise by 0 (i.e. lossless compression), 0.5, and 1.6 millimag rms for aperture photometry with a radius of 3 pixels of bright stars at corresponding file-size compression factors of 1.8, 3.8, and 7.5, respectively. Scintillation noise for the same images is 2 to 4 millimag rms, depending on atmospheric conditions, so potentially one could use lossy compression without significant loss of photometric accuracy. However, to be certain one wouldn’t later regret irreversibly lossy-compressing images (i.e. not saving the uncompressed data), very extensive testing would be required, so for critical scientific data, astronomers tend not to use lossy compression. The benefits of reduced disk storage space and transmission bandwidth are outweighed by the cost of that testing. Also, the cost, volume, and mass of disk storage all decrease ap-

proximately a factor of two every two years, so the smaller file size benefit of lossy compression is largely ephemeral.

Lossy compression also could be a practical SA algorithm to corrupt stellar photometry. However, lossy compression is a blunt tool for SA, because while it does corrupt stellar photometry by somewhat predictable amounts, it simultaneously corrupts astrometry and surface brightness of extended objects. Ideally, the SA could be turned on and off with a small encryption key, a.k.a. a password, that would transform the SA-data to its original form, i.e. O-data. Lossy compression algorithms, by definition, are irreversible, so they would not satisfy the desire for such a key. However, if such a key-based SA algorithm could be implemented, it would provide an incentive for astronomers to release to the public domain keyed-SA data, because the public domain could provide a robust backup for and access to the O-data. We imagine Alice might use the keyed-SA data herself and use the key whenever she wished to obtain the data with full fidelity.

The PHOTZIP algorithm was designed to losslessly retain the pixel values near stars while lossy-compressing regions of sky without stars, all in a user-selectable and predictable manner (Shamir & Nemiro 2005). The PHOTZIP algorithm thus may be an algorithm to achieve SA for the photometry of galaxies or other diffuse objects while allowing public access to stellar (or more exactly, point-like object) photometry (i.e. PHOTZIP as a SA algorithm would have the “opposite” goal of the algorithm described in Section 3.1).

3. Examples of SA algorithms

3.1. SA Algorithm for Photometry

The following is a simplistic algorithm for SA of precision photometry of bright stars in a 2-D image. The objectives of this SA algorithm are

1. SA-data will differ from O-data only at specifically identified pixels.
2. Photometry of SA-affected stars will differ from that of O-data by a predictable random fraction.
3. Astrometry of all stars, including the SA-affected stars, will not be affected.
4. SA-data could be converted to O-data by anyone that knew the encryption key.

The proposed algorithm would first identify which stars are desired to be subject to SA. They could be identified by coordinates or a parameteriza-

tion of their brightness or other characteristics. An associated table would store the (x,y) coordinates of a fiducial “center” pixel associated with each star subject to SA. Pixels within a specified vicinity of the tabulated pixels would be adjusted as part of the SA procedure, and all other pixels would be unaffected. The algorithm would solve for the “sky” value within each star’s vicinity and store that with the coordinates of the center pixel. The SA algorithm would generate a unique semi-infinite series of pseudo-random numbers from the encryption key. For each set of pixels associated with each star, the SA algorithm would subtract the sky value (a scalar), multiply the residuals (a vector of a few pixel values) by the next pseudo-random number (a scalar), add back the sky value and replace the appropriate pixels with the SA-adjusted values. Because the pixel coordinates and sky values are stored, and because the random numbers are pseudo-random (i.e. reproducible), the SA-data can be converted to an identical copy of the O-data by anyone who knows the encryption key. For the special circumstance of overlapping SA-affected regions around nearby stars, the order of the stars will be important and the inversion algorithm will need to operate on the stars in the correct (reverse) order.

The above algorithm can be described in equations as follows. Let I be the O-data value of a pixel in common between two stars (The algorithm can be generalized to pixels that are associated with one, two, or more than two stars), identified by their sequence indices m and n . The SA-algorithm converts I to I' and then to I'' by the following linear equations:

$$I' = R_m(I - S_m) + S_m \quad (1)$$

$$I'' = R_n(I' - S_n) + S_n \quad (2)$$

where the S are the sky values and the R are the pseudo-random numbers, equal to $1 + \epsilon$ where ϵ is a normally distributed pseudo-random number of zero mean and specified standard deviation (e.g. 0.03). In order to recover the original pixel value, I from the SA-value I'' , S_m , S_n , R_m , and R_n , the intermediate value I' should be recovered first from I'' , S_n , and R_n , then the I can be recovered from I' , S_m , and R_m .

The simple algorithm described above should permit SA of photometry while retaining astrometry to a high degree of precision. Astrometry should not be affected significantly because all the pixels within a region of interest associated with a specific star are multiplied by the same pseudo-random number (R) after the sky has been subtracted (and later restored). Differences between a specific stellar astrometry algorithm and the algorithm used to determine the sky value (S) could perturb the astrometry to some de-

gree. That possibility and the circumstance of pixels being shared by multiple stars both suggest that some empirical testing would be required to measure empirically the effect the proposed algorithm, designed for SA of photometry, would have on astrometry. Presumably the proposed algorithm’s effects, if any, on galaxies or extended structures would be insignificant, because as we have imagined its implementation, only pixels within approximately one or two FWHM of bright stars would be SA-adjusted, and only by multiplying them by scalars of order $1 + \epsilon$.

A significant improvement in the proposed algorithm would be elimination of the table of affected pixels and associated sky values. If the elements of the table could be algorithmically generated uniquely from both the SA-data and the O-data, then no separate table would be required, which would eliminate the need for a separate table. We encourage the reader to consider how to accomplish this improvement.

3.2. SA Algorithm for Astrometry

An algorithm that would provide SA of astrometry while not affecting aperture photometry could move counts from some pixels to other pixels, all within the PSF and the photometric aperture. We leave it as an exercise to the reader to design such an algorithm with similar objectives as those enumerated for photometry in Section 3.1 (The author has not completed that exercise himself).

3.3. SA Algorithm for Tabular Data

The SA algorithms for tabular data can trivially be simple, because the various parameters are separate already. For example, a table of photometric and astrometric measurements can easily have one or the other (or both) SA-adjusted by applying (i.e. adding or multiplying as appropriate) a sequence of pseudo-random deviates to each value in the table.

4. Discussion

Potential advantages of SA could be:

1. Bob will have access to SA-data sooner than he’d have access to O-data. Bob’s interest may be purely complementary science enabled directly by the SA-data, or his interest might be planning similar observations.
2. Alice enables additional science earlier than if she held all her data for their proprietary period.
3. SA may foster collaborations, by providing a means for Bob to realize and to demonstrate to

Alice that his own analysis or complementary data can add value to Alice’s SA-data and by inference also her O-data.

4. Alice could benefit from an archive of her SA-data because she can use it as an alternative user interface to her O-data, because she knows the decryption key. Alice can use the SA-data as a backup, either the sole copy or a copy redundant to her own.

Potential disadvantages of SA could be

1. The traditional proprietary period is simple to implement and to understand, whereas SA-data are more complicated to create and to understand.
2. Some may have the opinion that the intentional corruption of data (i.e. from an SA algorithm) is antithetical to science.
3. The community, journals, or referees may not trust SA-data or conclusions based upon them.
4. SA-data could be mis-interpreted in a manner that O-data would not be.
5. Scientific data products based upon SA-data could be considered ephemeral.
6. The SA-data products may require updating when the associated O-data became available.
7. Two “strains” of subsidiary data products potentially may co-exist in the literature.

The latter disadvantages (items 5-7) of SA would apply equally to the current practice at STScI of “on-the-fly reprocessing” in which the archive provides data that has been calibrated with the latest, and presumably best available, reference files (e.g. flat fields) and algorithms (e.g. cosmic ray identification). The latter practice shares some of the same motivation as SA, specifically to disseminate data known to be of lower fidelity sooner than data of higher fidelity to be made available at some time in the future.

5. Conclusions

The main purpose of this draft manuscript is to foster discussions of the potential advantages and disadvantages of selective availability (SA) of astronomical data prior to revising and publishing it in a journal. A auxiliary purpose is to encourage others to create and test superior SA algorithms. Comments, opinions, suggestions, and questions are welcome.

6. References

- McCullough, P. R., et al. (2006). *ApJ* **648**, 1228.
- Paczynski, B. (2006) *PASP* **118**, 1621.
- Shamir, L., Nemiroff, R. J. (2005) *AJ* **129**, 539.
- White, R.L., Postman, M., Lattanzi, M.G. (1992). *ASSL* **174**: Digitized Optical Sky Surveys, 167.