

Choice of cell size for HAP mosaic images

We are proposing to define a grid of cells (sometimes called "tiles" or "skycells") that will be used to construct the new HAP mosaics. This approach has several advantages discussed [here](#). The major advantage is that it makes the generation and augmentation of the mosaics more manageable (for the pipeline processing system and for users) as new observations get added. It also restricts the size of a region that must be reprocessed when new data gets added (e.g., making it unnecessary to reprocess the entire COSMOS mosaic when an overlapping observation is found.)

We plan to base the grid on the [Pan-STARRS sky tessellation pattern](#), which has a set of 2644 4°x4° projection cells arranged in rings of constant declination around the sky. Each projection cell is divided into a grid of small skycells. In PS1 there are 10x10 skycells per projection cell with each skycell being 0.4°x0.4°. The skycells from a projection cell share a tangent projection (they have identical CRVAL1/CRVAL2 and CD matrix values, with only CRPIX1 and CRPIX2 changing for different cells.) The same scheme can be used with a different choice of skycell size.

An important parameter in the definition of the cell grid for the Hubble Advanced Products is the size of an individual cell. A small cell size reduces the computing for each cell and leads to an efficient sky coverage without many blank pixels. However, if the cells are too small then individual exposures and objects often get divided up between different cells, making it necessary to combine multiple cells to get the data needed for science. On the other hand, if the cell size is large, the cells consist largely of blank pixels sprinkled with islands where there is some Hubble data. The resulting images can also be very large in pixels. For example, the 0.4°x0.4° PS1 skycells would have 36000x36000 pixels (at a scale of 0.04 arcsec/pixel), leading to a floating point image of ~ 5 GB (without compression).

- [The HLA Mosaics](#)
- [Effects of the cell size](#)
- [Conclusions](#)
- [Caveats](#)

The HLA Mosaics

I used the sizes of the existing HLA mosaic images to determine expected sizes for the HAP mosaics. The HLA mosaics do not use the tiling algorithm, but instead use the drizzle algorithm that determines the minimum size of a rectangular image that will enclose all the contributing exposures in the mosaic. Since the HLA mosaics use all the images that overlap in a region, they are probably representative of the size of a region that a user might like to use. By comparing these mosaic sizes to the HAP tile size, we can determine how often these "natural" mosaics will be split across the boundaries of cells (tiles).

There are 4306 filter images from ACS/WFC, WFC3/UVIS and WFC3/IR observations in [HLA DR10](#). Those images belong to 1348 mosaic fields (so there are an average of 3.2 detector/filter combinations per field). The first figure shows the distribution of the longest axis size in degrees for these fields. The field sizes range from 0.042 deg to 0.32 deg. The lower bound is determined by the size of an ACS or WFC3 image that is oriented north-south (rotated images require larger bounding boxes). The mean image size is 0.09 degrees, which is not too much larger than a single ACS/WFC image rotated at 45 degrees (0.078 deg).

[blocked URL](#)

The second plot shows the RA and Dec sizes for the mosaics. Most of the fields are fairly square, with the most oblong fields being about 2.5 times longer in the longest axis compared with the shortest axis.

[blocked URL](#)

Effects of the cell size

The field sizes combined with the cell sizes determine the likelihood that a given mosaic will be split across a skycell boundary. Since the skycell boundaries are randomly located on the sky compared with the mosaic edges, for any size mosaic we can compute the probability that the mosaic will be split by a boundary. For example, for a small mosaic with sizes that are less than the cell size, the split probability is

$$P(\text{split}) = 1 - \max(1-x,0) \cdot \max(1-y,0)$$

where **x** and **y** are the RA and Dec sizes of the mosaic divided by the cell size **s**:

$$x = \text{RA} / s$$
$$y = \text{Dec} / s$$

The **max** function returns the larger of the two parameters. If either **x** or **y** is greater than 1, the split probability is unity since the mosaic is bigger than a cell.

Obviously we could choose a large value for the cell size to reduce the split probability. However, then we generate a lot of mosaic tiles that have a small region with image data that is embedded in a large empty block of pixels. The **x** and **y** parameters can also be used to calculate the *filling factor* for the images. I define the filling factor as the fraction of pixels in a tile that fall within the boundaries of the original HLA mosaic. The mean filling factor for a particular mosaic is the average filling factor for all the tiles that wind up with a piece of the mosaic. The third figure shows an example for a 0.1° mosaic embedded in 0.2° cells. At this particular shift, the mosaic splits into 4 cells with various filling factors, and the mean filling factor is $f = 0.062$.

[blocked URL](#)

It is possible to derive an analytical expression for the mean filling factor (averaged over all shifts within the cell) as a function of the x,y parameters. For small mosaics ($x, y < 1$):

$$f = x \cdot y \cdot (1 - x/2) \cdot (1 - y/2)$$

A similar expression can be derived for any mosaic size (including mosaics that are larger than the cell size in one or both dimensions):

$$\begin{aligned} j &= \text{ceil}(x) \\ k &= \text{ceil}(y) \\ f &= 4 \cdot x \cdot y \cdot (1 - x/(2 \cdot j)) \cdot (1 - y/(2 \cdot k)) / ((j+1) \cdot (k+1)) \end{aligned}$$

where $\text{ceil}(x)$ is the smallest integer $> x$, meaning that $j-1 < x \leq j$ and $k-1 < y \leq k$.

The fourth figure shows the split probability and mean filling factor as a function of the cell size. The split probability is 1 for small cells (if the cell size is smaller than the mosaic, it is certain to be split). It decreases as the cell size increases and is around 50% for a cell size of 0.25° . On the other hand, the mean filling factor declines rapidly with cell size. So choosing a large cell size decreases the number of split fields but greatly increases the fraction of empty pixels in the tiles.

[blocked URL](#)

The final plot shows this same information in a different way (that I prefer). The x-axis is the split probability and the y-axis is the filling factor. For very small cells, the split probability is unity and the filling factor is large (little wasted space in the tiles). As the cell size increases, the filling factor drops and then (eventually) the split probability decreases. The dots mark particular values of the cell size.

[blocked URL](#)

A table of the marked values is given below. Along with the split probability and the filling factor, the table also gives the image size in pixels (assuming 0.04 arcsec pixels).

Split probability and filling factor vs. cell size			
Cell size	P(split)	Filling factor	Image size
0.01	1.000	0.792	900 x 900
0.02	1.000	0.649	1800 x 1800
0.05	1.000	0.408	4500 x 4500
0.10	0.938	0.236	9000 x 9000
0.15	0.786	0.158	13500 x 13500
0.20	0.655	0.110	18000 x 18000
0.25	0.556	0.080	22500 x 22500
0.30	0.482	0.061	27000 x 27000
0.40	0.378	0.038	36000 x 36000

Conclusions

My conclusion is that while a cell size of 0.2° might be a reasonable compromise, it still leads to a lot of split mosaics (about 65% of them are split), and it also leads to a lot of empty space in the mosaics (about 10% of the tiles are filled). That argues that:

(1) A cutout service that can combine tiles from neighboring skycells is going to be very important for usability, since targets and regions will be commonly split across cell boundaries.

(2) Storing the images in compressed format to "squeeze out" the empty pixels will also be essential, both for reducing the storage volume and also to improve speed of access (it is faster to read a compressed version of empty regions.)

Caveats

The HLA mosaic membership is determined from the Hubble Source Catalog source groups. Note that those groups include WFPC2 data (even though we have not generated any WFPC2 mosaics), which means that some of the HLA mosaics have WFPC2 "bridges" that join two smaller ACS+WFC3 image groups. So the HLA mosaic sizes will sometimes be larger than the HAP mosaics using the same data. That makes these tests slightly pessimistic (although that is not a big effect).

I have not included the effect of an overlapping strip around edges of the cells. That can reduce the split probability slightly: it makes the cells larger, and it also creates regions at the edge where a small mosaic may be split by one cell edge but not by the other. That will help with the splitting (but not the filling factor, I think).