HiPS ingestion strategy

Caroline Bot, Thomas Boch and the Aladin team

June 13, 2023

1 Introduction

In an ideal world, we aim for having a collection of Hierarchical surveys that would represent the whole landscape of images in astronomy: the largest available sky coverage (whether observations are scattered pointed fields or global sky surveys), images at all wavelengths, images at all resolutions up to the finest ones, images of all depth up to the deepest ones, ... In practice, while we keep that aim, we have to make choices according to the availability of the data as well as resources (human, storage, computing time,...) and this document aims at defining our strategy, laying down our guidelines and criteria when ingesting images into HiPS.

One fundamental criteria we have chosen refers to the quality: we create HiPS out of data sets that have been published, i.e. there is a paper associated to each dataset we transform. In practice, this means for example that images taken by amateur astronomers that do not have an associated publication are not ingested. Example: The Mellinger survey is available as a HiPS, while images from Ciel Austral group were not ingested.

Another fundamental guideline is that we always favor situations where experts of the datasets are making the HiPS. We will ingest images for which the data providers are not making a HiPS themselves, but we always prefer to encourage that the people who know their data best make the image selection, documentation and HiPS creation. We therefore are happy to help teams that would need advice or feedback and will favor this kind of interaction. Our goal and missions are clearly not to create at CDS all the possible HiPS from all existing datasets, even if we had the resources to do so. This is also a way to improve the quality of the HiPS (ideally made by the people who know best) as well as a criteria to select datasets to ingest, and a way to foster the usage of HiPS by a wider community.

In practice, there is however a large need that CDS creates a given amount of HiPS directly. Even if we presently manage to do most of the HiPS we would like to ingest, even if sometimes with a long delay and different priorities, we are already not able to create all the HiPS we would like to. The pressure is likely to increase even more in the (near) future and this implies to put priorities. In all cases, our criteria to decide whether we want to ingest a HiPS or not will be the same as they are right now and this is what we detail in this document. This decision procedure is made of two parts:

- to ingest a dataset, we need to be aware that it exists and could make a potential HiPS. We describe our strategy for dataset watch in section 2.
- we put priorities on a potential dataset to ingest by balancing different aspects that we describe in section 3

This summarizes the strategy we have to watch for, prioritize and eventually select the image data sets that are converted into Hierarchical Progressive Surveys.

2 Keeping up to date with existing and new surveys

In order to be informed about new surveys and datasets of interest, we make use of several channels :

- all members of the Aladin team (as well as CDS) go to *conferences* and gather informations on current missions and surveys, different available image data sets becoming public, ... This includes scientific conferences, community conferences (e.g. AAS, EAS, SF2A) or data and

service-centered conferences (e.g. ADASS, IVOA Interoperability). This is a good way to keep track of the projects and interests of the astronomical community at large

- we also receive a given number of *newsletters*, including from large astronomical data centers (e.g. ESO, IPAC, MAST, ESA). This is a good way to be aware of large datasets or surveys becoming public.

- some of us follow different *social media*, which is a good way to keep track of press releases, highlights, or interesting/trendy images and news

- part of the data watch is also done *as part of other activities at CDS*. For example, scientists at CDS can be aware of interesting data sets simply by checking the literature, either for their own research, or through the validation of tables for VizieR (this includes associated data in VizieR but not only), or through the weekly bibliography check that happens for Simbad (=g meetings). This information of potentially interesting image surveys is then passed on to the Aladin team.

- we can also be informed of image data sets of interest *from external users*, either by PIs who contact us directly or through users looking for a specific image survey and requesting it through cds-question hotline or direct contact.

Despite these various ways to check for image data sets of interests, we are aware that it is difficult to be complete. We hope that the variety of information sources and people at CDS that perform this watch (either naturally on their daily routine or purposefully), we are not missing datasets that would be of high priority to ingest.

3 Selection and prioritization of datasets to ingest

For any new identified image dataset that can potentially become a HiPS, an evaluation is done of whether a HiPS should be made and how high a priority it is. In defining this selection and priorities, several aspects are taken into account and weighted before making a decision:

- *How does it completement existing HiPS data sets?* This includes arguments like the sky coverage, the availability of other HiPS in the same wavelength range, whether it brings a higher resolution or a deeper intensity view, ...

- *What is the added value of doing this HiPS given the already existing image data sets?* The added value can include cases like gaining the effect of a global view that is not available with the collection of small individual images. It can also be the fact to add a new data release of a dataset for which we already have a HiPS, or to redo a HiPS to be able to have access to FITS images and hence pixel values and dynamics with respect to an existing HiPS with only PNG files.

- *Is there a specific opportunity?* This includes cases like someone willing to invest a certain amount of effort for a specific survey (e.g. D. Durand collaboration for HST and JWST HiPS), a special moment to highlight our services (e.g. JWST early release data sets) or an internal interest (e.g. redoing the GALEX surveys as a testbed for filtering tools).

- *How easy/difficult is it to do?* E.g. a survey distributed as a HEALPix file is a straightforward product to add and is usually an easy addition to the collection for a small amount of computation and storage resources. On the other hand, very large optical surveys may require months of computation and some image data sets may require specific knowledge about the data (e.g. which files, pixel units, filters, keywords for blank pixels or instrumental effects, ...). Again, we reinforce the idea that we always favor the case where experts of the datasets or providers of the datasets make the HiPS. Hence the last question we ask ourselves:

- *Can we encourage and help others to do this HiPS?*

With all these elements to weight on, we assign a priority degree to the considered data set and compare it with other surveys that are currently in our waiting list. In practice, this means that different data sets will take different times to get ingested and become a HiPS. As the data amounts

continues to rise, we will eventually set a limit at which some datasets will not be considered. At the moment, this has mainly happened when no publication was associated or if the HiPS was already available elsewhere (even in a different format, e.g. PNG only).