

# The NVO Hyperatlas Standard

for building multi-wavelength images

Roy Williams

## 1 What is it

Several groups are interested in reprocessing astronomical image surveys through resampling and mosaicking. These include Caltech CACR, SDSC, NASA JPL, and IPAC. We are impelled by (a) the NASA Montage and Polysamp projects provide the trusted software for switching between projections, (b) the large data computing facilities provided by projects such as Teragrid and NASA IPG, and (c) the data handling software emerging from the Grid community, such as SRB and Globus.

I would like to suggest a collaboration between these groups and the NVO that directs and focusses our work to building an "Image Hyperatlas" of the sky. The idea is to agree on a set of fiducial map projections -- like pages of an atlas -- and all images of each survey are rendered to that set of projections.

In this paper, we define an *atlas* to be a collection of *pages* (or charts in the language of differential geometry). Each page is a mathematical mapping between part of the celestial sphere and a rectangular pixel space. The hyperatlas standard defines some standard atlases, and a way to express the atlas information in a string, for example TM-5-TAN-20 is an atlas that complies to the standard. Each standard atlas specification includes:

- A set of directions in the sky that act as "plate centers" for the pages of the atlas.
- A projection type chosen from the WCS enumeration, eg SIN, TAN, CYL.
- A scale (arc-seconds per pixel) for the resolution of the pages, that is chosen from a discrete set of possibilities.

### 1.1 All-Sky Mosaicking

The initial Teragrid implementation of hyperatlas will reprocess the 2MASS and Sloan imagery into a single set of atlas pages with eight channels – three from 2MASS and five from Sloan. This will be extended by a further three channels from the DPOSS survey, and a channel from the FIRST radio survey. We expect to use the TM-5-TAN-20 atlas for the reprocessing (see below for exact definition).

### 1.2 Multiwavelength Astronomy

There is a lot of science that can be done with federated pixels -- see the papers [1], [2] on Multiwavelength Image Space. Some salient points are:

- Finding fainter sources with spectra: we can go fainter in image space because we have more photons from the combined images and because the multiple detections can be used to enhance the reliability of sources at a given threshold. Because the object is detected in multiple channels, we also have significant spectral information.

- Image differencing can pick out changes over time in an automated fashion, as well as provide spectral specificity in source extraction. Before reaping the gains, we will need to research effects of different point-spread functions, as well as effects of positional error.
- Robust detection and flux measurement of complex, extended sources over a range of size scales. We will be able to combine multiple instrument imagery to build a multi-scale, multi-wavelength picture of such extended objects. It is also interesting to make statistical studies of less spectacular, but extended, complex sources that vary in shape with wavelength.
- A trend in astronomy is the synoptic survey, where the sky is imaged repeatedly to look for time-varying objects. Hyperatlas will be well-placed for mining the massive data from such surveys.
- We will use the new multi-wavelength images to specifically look for objects that are not obvious in one wavelength alone. Quasars were discovered in this way by federating optical and radio data. We hope for discovery of new classes of essentially multi-wavelength objects. We will make sophisticated, self-training, pattern recognition sweeps through the entire image data set. An example is a distant quasar so well aligned with a foreground galaxy to be perfectly gravitationally lensed, but where the galaxy and the lens are only detectable in images at different wavelengths.

### 1.3 Education

The Hyperatlas can be used for other purposes besides multiwavelength science. Further processing of the atlas images can provide an educational dataset showing the night sky in many scales and many wavelengths. Such a resource can be used to build interactive educational materials of great depth, materials where data can be drilled down all the way to the original image data.

### 1.4 A Deep Registry of Astronomical Images

We can also think of the Hyperatlas as a directory of astronomical imagery. Given a point on the sky, the directory services of the Hyperatlas can point to all the sky surveys that cover that point. The sky can be available in these multiple wavelengths with the pixels exactly co-registered, meaning that data-mining software can be run as well as sophisticated visualization.

## 2 Data Model

### 2.1 Pages

First we define a **virtual page** to be a reference point on the sphere together with a WCS projection mapping type (eg TAN, SIN, AIT), a scale number, and a rotation angle. Combining these provides a mapping from the sphere to a plane, where the integer coordinates on the plane correspond to the scale of the virtual page. It is also assumed that the distortion of the virtual page projection approaches zero at the reference point. The scale number (in degrees per pixel), is assumed to be the magnitude of the Jacobian matrix of the virtual page at the reference point.

The main method associated with a virtual page is a bidirectional map from a subset of the plane (pixelated at the given scale) to a subset of the sphere. Note that the domain of the map on the plane is determined only by the WCS type: a TAN projection extends to infinity, and a SIN projection has a disk domain corresponding to a hemispherical range on the sphere. The reason it

is called a virtual page is because of this infinite extent -- it is a way of talking of a map projection, rather than a frame with which real data is associated. In the terminology of FITS WCS, a virtual page carries the CTYPE keywords (projection type), the CRVAL keywords (reference point), and the CDELTA/CROTA/CD keywords (scale, rotation, etc)

We now define a **rectangle** to be a pair of integer side lengths, and the position of the rectangle on the plane of a virtual page, relative to the reference point of the virtual page.

We then define a **page** to be a virtual page combined with a rectangle: this is the vehicle for geometrical metadata of astronomical images. One of the most useful methods of this class evaluates whether two pages may overlap: this provides a definition of computing jobs when an image collection is to be reprojected. A page is equivalent to the FITS WCS header, since the rectangle on the pixel plane is defined by the NAXIS keywords (image size), and CRPIX keywords (placement of the rectangle relative to the reference point)

The important data objects that would be used in application of the proposed standard are **pages**: rectangular data arrays. However, the standard itself is written in terms of **virtual pages**. This concept is an infinite plane on which pixels may or may not be placed, rather than a particular rectangular part of that plane (page).

The standard we propose is about naming virtual pages, and about ways to build standard sets of virtual pages (atlases) from astronomical image archives.

## 2.2 Atlas of Pages

The NVO Hypermap is a protocol for defining collections of virtual pages so that images may be rendered on them. In the words of differential geometry, an *Atlas of Charts* defines a manifold through a collection of mappings. For our purposes, we will define an **Atlas** as a unified collection of virtual pages, in the sense of the following objects:

- An AtlasType, which corresponds to an algorithm for generating the collection of virtual pages and their reference points. In IT terms, the AtlasType is the name of a class, or shared object.
- A parameter string that is interpreted in the context of the chosen AtlasType, for example "4" to mean "level 4" to one AtlasType, or "4 degree grid" to another.
- The three-letter projection specification, as defined by the FITS-WCS standard. For the purposes of defining these standard atlases, we do not consider any WCS projections that require extra data in addition to the three-letter code -- no polynomials for example.

The atlas is defined by a string which is these strings concatenated with a hyphen ("-"). Atlases that have been suggested are:

- The "TM" scheme, where page reference points count around RA from zero, with regularly spaced lines of constant declination. The parameter represents the spacing between the lines of declination. Thus "TM-5" is similar to the DPOSS plate layout of five degree centers, and covers the sphere in 1732 pages, with no point of the sphere further than 3.5 degrees from a page center.
- The "HV" scheme, which is the vertices of the Hierarchical Triangular Mesh (HTM) at some level. The parameter represents HTM level, and "HV-4" covers the sphere in 1026 reference points

Methods associated with an Atlas include:

- How many virtual pages are there in the atlas.
- Return a virtual page from its number (0 to N-1).
- Return a virtual page whose reference pixel is closest to a given point in the sky.

We would like to further propose that the virtual pages of an atlas be numbered from 0 to N-1, where N is the number of virtual pages in the atlas.

### 2.3 Discrete Scale: $2^n$ seconds per pixel

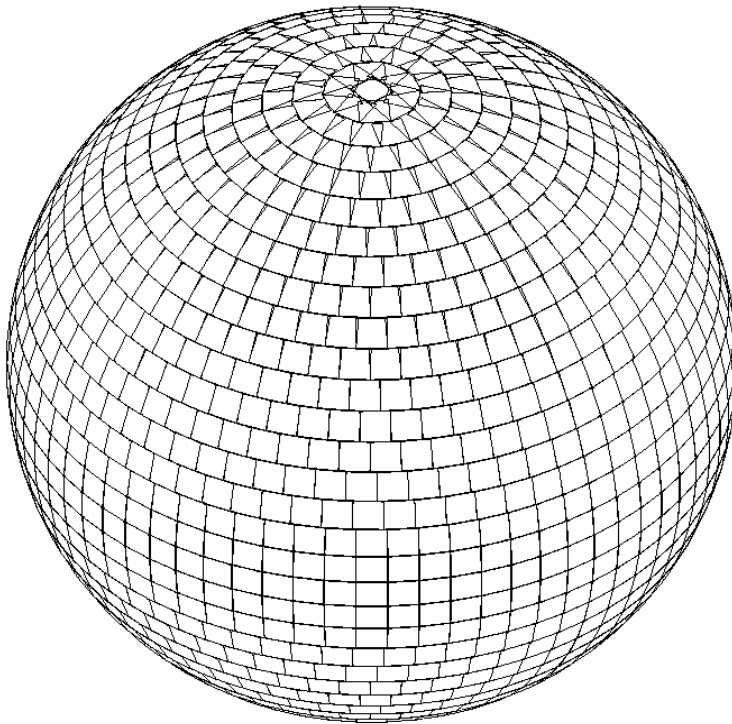
A suggestion for discretizing the scale of images uses a geometric sequence differing by a factor of two between scales. Suppose we pick scale  $S=20$  to be exactly one arc second per pixel,  $S=19$  to be 2 sec/pixel, and so on, including  $S=10$ , 1024 seconds = 17.07 minutes per pixel. At scale 10, the SIN projection renders an entire hemisphere with an image 403 pixels on a side. On the other hand, Scale 25 (0.03 seconds per pixel) would work well for the Hubble Deep Field.

### 2.4 Names for Atlases and Pages

We have defined a quite general scheme for atlases, yet they can be named in a very compact fashion, for example:

**TM-5-SIN-20**

would be an atlas with the TM-5 layout of pointing centers, with SIN projection at scale 20 (one second per pixel). The objective here is to build a "menu of atlases", with enough choice to allow



TM-5 hyperatlas

4.869147607046481 degree chart width

for special needs, while also allowing the community to settle on a default atlas in the light of experience.

A virtual page of the atlas can be defined with a further field specifying the page number:

### **TM-5-SIN-20-1733**

In the case of the TM-5 atlas, virtual page number 1733 covers the south pole.

The Figure shows the layout of pages for the TM-5 scheme.

## **2.5 Atlas Pages as Virtual Data**

The Pages in the Atlas are abstract objects -- the plane of a virtual page is potentially infinite in extent. In contrast, the real page has a bounding rectangle in the plane, so it is semantically equivalent to the WCS part of the FITS header, the part that places the pixels of a rectangular image on the sphere.

In the language of the Griphyn project, the virtual page forms a virtual data space, where the rendering of an image is possible, but may or may not have been done yet. When requested, any part of the potentially infinite plane can have data rendered on it, and that computed product is stored away. This means that next time that data is requested (or a subset of it), it can be retrieved from the cache and produced quickly.

Virtual data is akin to the idea that a data object is somehow equivalent to the program that produces it. We can start an atlas project without extended computing, by setting up a server which computes on demand. Results of computation can be stored in the cache so that they are available quickly the second time around. A data cache can also be built up by low-level compute cycles ("SETI-at-home"). Users of a data federation system can use a remote service to render their own data on to the atlas. The virtual data model provides a flexible way to structure the computing work.

In deciding to render image data to an atlas, there are practical decisions to be made about the virtual pages in the atlas. Considerations include the number of pixels that might be wanted, which is connected to the overlap of the rendered pages. For the TM-5 atlas, for example, if every page is 6 degrees wide, then there is a one-degree overlap between neighboring pages. Astronomers would not want to move further than this on a given page, for fear of pixel shape distortions. But other users, for example a planetarium show, might want to render very wide, continuous imagery on a single virtual page.

Another consideration is the number of pages that are to be rendered in an application. If we render a survey of the northern hemisphere, we would not be interested in a lot of southern hemisphere pages that have no data attached. If we render the Hubble Deep Field, we might choose a single virtual page, but a very deep scale.

## **3 Summary**

The hyperatlas protocol is a way to assign concise names to the virtual pages of a standard atlas: in the example above TM-5-SIN-20-1731 is such. Mosaicking an image collection to this virtual page involves also the choice of images, and the choice of algorithm for doing the mosaicking.

Different data providers will make different collections available, but we would hope that each is rendered to some NVO-approved hyperatlas page at some suitable scale. This will be the product that astronomers use for data mining. If the data were further transformed to other virtual pages at coarser scale, then a visual index is established, allowing zoom out. The zoomed out view could connect to the NVO registry, and thereby contain directory information, such as a blue box (cf HTML clickable image) to indicate that other data has been rendered on this virtual page.

We can easily establish an XML specification for a virtual page, based on the model above. This could be used by digital library tools that allow the creation of a community around a metadata standard. In this case the relevant standard is the definition of a virtual page, and its parent atlas. In this way, we provide a “publish and discover” service for people who have rendered data to an NVO-compliant atlas.

## 4 The proposed standard

An atlas is defined as a combination:

```
<atlasType>-<atlasParam>-<ProjectionID>-<Scale>
```

In regard to this:

- The NVO provides services and software to support some atlas types (eg TM, HV), and
- If the atlas is at fine scale ( $S \gg 17$ ), projectionID is chosen to be TAN; at coarse scales (all-sky), pages should be in AIT or SIN. (is this needed ???)
- That scales be discretized to a power of two times one second per pixel.
- A particular virtual page of an atlas can be named with an integer page number, that runs from zero to N-1, where N is the number of virtual pages in the atlas.
- A virtual page may be named more fully with the construction

For example:

The atlas HV-4-SIN-13 is built from the HV (HTM vertices) at level 4, with 1026 virtual pages. The projection and scale makes the celestial sphere into a ball, seen from infinity. Each virtual page has a diameter of 403,000 pixels, and each pixel is about 2 arcminutes.

## 5 Services

The hyperatlas standard is instantiated as a set of services, as described below, to connect page numbers with sky positions and with WCS fragments of FITS images headers. The service can be called with the name of an atlas which defaults to TM-5-TAN-20. In this case it responds with a table of the essential parameters defining the virtual pages. The columns of this table are:

- Virtual page number
- Scale in degrees per pixel
- WCS Ctype strings
- Right ascension of plate center
- Declination of plate center

### 5.1.1 All Pages

Calling the hyperatlas service with just an atlas name provides the whole table:

```
<baseUrl>/getChart?atlas=TM-5-SIN-20
0      2.77777778E-4 'RA---SIN' 'DEC--SIN' 0.0 -90.0
1      2.77777778E-4 'RA---SIN' 'DEC--SIN' 0.0 -85.0
2      2.77777778E-4 'RA---SIN' 'DEC--SIN' 36.0 -85.0
...
1731   2.77777778E-4 'RA---SIN' 'DEC--SIN' 288.0 85.0
1732   2.77777778E-4 'RA---SIN' 'DEC--SIN' 324.0 85.0
1733   2.77777778E-4 'RA---SIN' 'DEC--SIN' 0.0 90.0
```

### 5.1.2 Best Page

Calling the service with RA and Dec parameters provides the best page for that position in the sky, for example:

```
<baseUrl>/getChart?atlas=TM-5-SIN-20&RA=182&Dec=62
1604 2.77777778E-4 'RA---SIN' 'DEC--SIN' 184.61538461538458 60.0
```

### 5.1.3 Numbered Page

Calling the service with a specific page number gives the relevant parameters for that page:

```
<baseUrl>/getChart?atlas=TM-5-SIN-20&page=1604
1604 2.77777778E-4 'RA---SIN' 'DEC--SIN' 184.61538461538458 60.0
```

### 5.1.4 Implementations

Currently there are two implementations of the hyperatlas standard services:

baseUrl = <http://mercury.cacr.caltech.edu:8080/hyperatlas>

baseUrl = <http://virtualsky.org/servlet>

## 6 Grid Computing

The Atlasmaker project will use Teragrid and NASA IPG technology, in combination with NVO interoperability, to create new knowledge resources in astronomy. The product is a multi-wavelength, scientifically trusted image atlas of the sky, made by federating many different surveys at different wavelengths. The atlas is implemented as virtual data – the data is intimately linked with the software that makes it – which is implemented on the NPACI SRB system. Computations take place on the Teragrid, either as on-demand instantiations or backfill, and computed data stored in a distributed virtual file system based on SRB.

The proposed work will

- create new knowledge in astronomy,
- serve as a magnificent canvas for education and outreach,
- prove the combination of SRB and Teragrid, and
- enhance penetration of the NSF-NVO data protocols.

The computational power for Atlasmaker comes from Teragrid/IPG, while the scientific motivation is strong connection to the National Virtual Observatory (NVO) project. Atlasmaker uses Montage, a new and rigorous code for mosaicking images, that is one of the Teragrid flagship applications. Under NPACI funds, the code has been parallelized and scripted into the

Atlasmaker package for high-performance, wide-area computation on Teragrid, using SRB for (some) input images, but essentially for the distributed storage of the resulting atlases. There is new code for connecting to arbitrary image archives that are using the NVO publishing protocol (SIAP). The protocol allows for multiple retrieval mechanisms: if the input data is on an SRB system, it can be retrieved that way, or else through HTTP.

Services have also been built for the creation of atlases -- coherent collections of mosaicked images that lead directly to multi-wavelength imagery. We expect these atlases to be a new and powerful paradigm for knowledge extraction in astronomy, as well as a magnificent way to build educational resources. As the Teragrid matures, we expect to be computing large numbers of mosaics, each a reprocessing of a particular image survey to a particular page from an atlas. The results will be stored back in a single virtual file system managed by SRB, but physically located at SDSC, JPL, and CACR.

## 7 References

[1] Williams RD; Grids and the Virtual Observatory, in *Grid Computing: Making The Global Infrastructure a Reality* by Fran Berman, Anthony J.G. Hey, and Geoffrey Fox, Wiley, 2003, pp 837-858.

[2] Williams RD, Berriman GB, Deelman E, Good J, Jacob J, Kesselman C, Lonsdale C, Oliver S, Prince T, Multi-Wavelength Image Space: Another Grid-Enabled Science, *Concurrency & Computation*, vol 15 (2003), pp 539-549.

[3] Montage, <http://montage.ipac.caltech.edu/>