SIA Upgrade Special Topics

Complex Data and Data Cube Data Access

D. Tody, April 2006

Image Data. The updated SIA interface will be based on the generic Dataset query interface and metadata, development of which is already proceeding as part of the SSA effort. Recent work on SIA has focused primarily on several special topics which are being driven primarily by image access requirements. These topics include access to multi-dimensional (cube) data, description and access to "complex" data and the related issue of general data discovery, and more programmatic issues such as asynchronous data staging.

A use-case analysis of selected projects or instruments producing spectral data cube data, begun in the last quarter, was completed in February and published as a draft IVOA Note (D. Tody, F. Bonnarel, et. al.). This looked at data from several radio data cube surveys (GALFA, CGPS, SGPS), plus an O/IR IFU instrument from ESO (SINFONI), which produces true data cubes based on image slicing techniques combined with a grating spectrograph. Scientific representatives from all the projects mentioned contributed to the analysis. This was a very interesting study which improved our understanding of data access requirements in several areas, including how to go about accessing potentially very large cube datasets, and approaches for handling complex data.

Complex data. By complex data we mean a data collection which consists of more than one type of data or which can be viewed in multiple ways. The current DAL interfaces are object-specific and are optimized for access to a single type of data (catalog, image, spectrum, etc.). A typical complex data collection might include one or more data cubes in different wave bands, standard 2-D projections or continuum images corresponding to these cubes, and one or more catalogs produced from this data. Access to such a data collection might support multiple views of the same data, e.g., a 2-D service for viewing continuum images, a 3-D cutout service for accessing cube data, plus a service to dynamically extract 1D spectra from a spectral data cube. A client application may need all these capabilities to visualize or analyze the data.

The problem of complex data breaks down into two areas: data and service discovery, and data access. We need to be able to describe such data collections, associating and describing the different types of data making up the collection (at the level of individual datasets matching some discovery query), and tell the client application what services are available to access the data.

This goes beyond what DAL can currently do as there is no way at the level of an individual object-specific service (SIA, SSA, etc.) to associate different types of data. It would be inappropriate for an image service for example, to know anything about spectra or catalogs. Something higher level is required to be able to make this association.

The proposed solution to this problem currently being studied is the **generic dataset query**. As noted earlier, Image, Spectrum, SED, etc., are all subclasses of Dataset, to the extent of sharing a common basic query interface and generic dataset metadata. The generic dataset query would query for datasets of any type, providing a mechanism to discover all elements of a complex data collection (the individual datasets) in a single discovery query. The single query response would make it straightforward to directly associate the elements of a complex data association, as well as point to the services available to access the data, including cases where multiple services (views) of the same data are available, such as image projection or spectral extraction.

Under the current proposal, the generic dataset query would be used primarily for high level discovery and association. The actual data access would be done using the current unmodified SIA, SSA, etc. data access services. Only the data type-specific services would have capabilities for precision data access to a given type of data, i.e., data model mediation and virtual data generation, including capabilities to subset, filter, or transform the data. Existing services such as SIA would be unaffected by the generic dataset query. Data access which does not require access to complex data, e.g., retrieval of single image cutouts from multiple distinct data collections, would work as at present, with a single SIA query (posed to multiple services) providing all that is required.

Data cubes. Radio surveys commonly produce cube data (spectral data cubes). Within O/IR, integral field unit (IFU) instruments producing cube data are becoming increasingly common. Time cubes, i.e., synoptic images, can also be considered a type of cube data.

The most notable aspect of cube data is that such cubes can be very large, making analysis by simply downloading a precomputed data cube for local processing often impractical. Current data cubes are typically in the range of several hundred megabytes to several gigabytes in size. In the near future, instruments with a higher spectral bandwidth will be capable of producing much larger cubes. For example, a 2K by 2K cube with 8K spectral channels would be 128 GB in size at 4 bytes per pixel. A single observation with multiple polarizations or bands could potentially be up to half a terabyte in size!

For datasets this large, some approach which moves much of the computation to the data becomes essential. Our focus at this point is on advanced data access services capable of dynamically subsetting and filtering the data. In this scenario the client application might access a single large data cube multiple times as analysis proceeds, each time accessing a different portion of the cube or computing a different view.

The access modes thus far identified include at least the following:

- Whole dataset (for smaller cubes or high bandwidth scenarios)
- Spectrum extraction
- Cutout 2-D plane along any pair of axes
- Cutout 3-D sub-cube
- 2-D projection, collapsing the 3-D cube along one axis
- 3-D projection, resampling the data
- 2-D slice through a 3-D cube at an arbitrary position and orientation

Implementing such capabilities for cube data is a generalization of the 2-D cutout and projection capabilities already present in SIA. The basic approach for subsetting the data is to specify the WCS and image geometry of the desired data product and let the service determine how to generate the specified data. Filtering can be provided by the use of a range-list to specify intervals of the spectral or time coordinate axis for which data should be returned or used, for example in computing a 2-D projection of a cube.

Whether cube access should be provided by generalizing the current 2-D SIA interface, or by defining separate 2-D and N-D image access interfaces, is not yet clear. Either approach would work. The 2-D case may be important enough to be worth separate treatment in order to simplify the interface.

Catalog access. In addition to providing a mechanism for dealing with complex data associations, the proposed generic dataset query would provide a powerful data discovery mechanism capable of finding all ("atlas" or static) data from a given site matching some query, be the query positional or otherwise.

In effect this would be a query against a global data catalog, where the generic dataset metadata defines a data model used to uniformly describe all available data regardless of the object type (catalog, image, spectrum, etc.). The availability of such well-defined, uniform dataset metadata is a prerequisite to being able to pose a general data discovery query. Since a generic dataset query interface is based on the assumption of such uniform generic dataset metadata, this would be a logical place to introduce ADQL into the DAL interfaces. Since SIA, SSA, etc., inherit from the generic Dataset, this may provide a strategy for eventually introducing ADQL capabilities into the DAL interfaces.

.