VO Data Model - FITS, XML plus NDF: the Whole is More than the Sum of the Parts

David Giaretta, Patrick Wallace, Martin Bly, Brian McIlwrath

Starlink, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, UK

Clive Page

Department of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK

Norman Gray

Astronomy Group, Department of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ, UK

Mark Taylor

Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 0HA, UK

Abstract. We believe that the coming Virtual Observatory implementations will need to access astronomical data in a more structured, hierarchical way. This paper proposes a way to combine the advantages of FITS and XML to provide this functionality. In addition we believe that the lessons learned by Starlink in developing the HDS and NDF systems will prove valuable.

1. Introduction

In the Virtual Observatory era, when astronomers will have easy access to large volumes of many kinds of distributed data, it seems likely to us that they will need to be able to use such data in a more structured, hierarchical way.

Interrelationships will need to be specified between data files which may be distributed between archives, without having to copy large volumes of data into a single local (possibly extremely large) file. In addition the same piece of astronomical data may be part of several different such groupings of files, for example once as a member of a time sequence of observations and again as a part of a sky mosaic.

At other times a user may collect large numbers of files locally, either from archives or from observing runs. It will be convenient not to have to combine these together into a single file.

The hierarchical distributable data structure which seems to be demanded must also be portable, flexible and extensible. It must be accessible through an

interface which facilitates efficient, distributable and parallel processing. Finally we would hope to take the maximum advantage of industry standard tools to provide sophisticated ways of displaying and interrogating this type of structured data.

The use of XML is now the obvious way to provide hierarchically structured data. A number of proposals have been made in the past to replace FITS with something else based on XML for example XDF. The problems with such proposals are that (1) FITS is widely used and works well and (2) XML still requires something outside itself to hold binary data.

2. FITS

FITS is the pre-eminent astronomical data format. It is however handicapped by not being inherently hierarchical. Other data formats, such as Starlink's NDF/HDS, are hierarchical and provide powerful ways to express images and spectra; but none has achieved significant penetration simply because they are not FITS.

Meanwhile proposals to extend FITS abound, making what began as a simple data transport format increasingly baroque.

On the other hand FITS clearly allows efficient access to binary data, and most astronomical data is carried away from observatories as FITS files. Its use is also deeply embedded in the bulk of astronomical tools. Thus it seems unnecessary and indeed unwise to try to replace FITS as the bulk data container of astronomical data.

3. Combining XML and FITS

What is proposed in this paper is to combine XML and FITS in order to take advantage of the strength of each. XML provides the structural information, while the FITS files contain the bulk data.

In the simplest case one has a bare FITS file. In the next simplest case one has an XML structure in which one leaf points to a (possibly non-local) FITS file. Increasing in complexity one can have for example a single FITS image referred to by multiple XML files, as illustrated in Figure 1.

A certain amount of metadata, such as image size and World Coordinate System information, could reasonably be copied or relocated from the FITS file into the XML file, in order to use XML tools to query and display the structural information of the combined dataset. It would further be sensible to keep the FITS keywords unchanged in the XML.

At all times the individual FITS files would be accessible using existing tools.

4. Complex Relationships - possible problems

Having structured data has, as should now be clear, major advantages. On the other hand it should be recognised that there may be problems associated with flexibility introduced.

These problems are exactly the sort encountered by Starlink when it introduced its present structured data format – Hierarchical Data System (HDS)– in 1985. The problem was that application writers could generate arbitrarily complex data structures. Their own applications could interpret and understand the interrelationships between the components, but in general other applications could not. This could easily lead, in the worst case, to a complete lack of the ability to exchange data between applications without some human intervention by, for example, reading comments and acting appropriately.

In principle there are techniques such as RDF or XTM which allow one to specify relationships. However these are at the moment insufficiently mature to allow automated processing of arbitrarily complex data structures.

To solve the problem arising from the use of HDS, in 1990 the Starlink project imposed a strong astronomical data model to be used by all applications. This model, known as NDF, was built on top of HDS. Building this model – in particular balancing the tensions between abstraction, usability and completeness – proved unexpectedly difficult, and the final design represents a hard-won consensus amongst a broad spectrum of applications authors and data users. That this format is still in constant use by Starlink's broad range of applications and users testifies to the validity of the compromises it represents. We expect to build on and develop this expertise in this proposed combination of XML and FITS.

5. The Starlink NDF

Starlink's NDF is at the simplest just a data array together with a data type and a name, within an HDS file. The power lies in the optional additional components which include:

- Variance - allowing one to track data errors
- Data Quality - allowing one to store complex quality information for each pixel.
- Units
- Axis information, including errors, and normalisation information
- World Coordinate System information
- History
- Extensions - allowing a way, supported by application conventions, which ensure that additional information can be carried forward even when a particular application does not know how to otherwise deal with the information contained therein.

Figure 2 shows an XML implementation of NDF; the data could, if necessary, be contained within the XML file, but the more normal case would have the data contained in an external FITS file.

6. API

An API is being developed to provide uniform access to this type of structured data, whether it be a simple FITS file, an XML/FITS combination or even an HDS file.

In order to traverse the structure, an API based on the DOM (or JDOM) interface is envisaged. This will allow access to industry standard tools. Furthermore the structure will be displayable by such common tools also. To accompany this will be a thread-safe, distributable, API with an associated efficient implementation to access image and table data.

References

¡b¿NDF¡/b¿ http://www.starlink.rl.ac.uk/star/docs/sun33.htx/sun33.html#xref_
¡b¿HDS¡/b¿ http://www.starlink.rl.ac.uk/star/docs/sun92.htx/sun92.html#xref_
¡b¿XDF¡/b¿ http://xml.gsfc.nasa.gov/XDF/XDF_home.html
¡b¿FITS¡/b¿ http://fits.gsfc.nasa.gov/fits_home.html