Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

Formatting Journal Tables in XML at the ADC

Edward Shaya[1], Jim Blackwell, Jim Gass, Vincent Kargatis, Gail Schneider, Kirk Borne

Raytheon ITSS, 4400 Forbes Boulevard, Lanham, MD, 20706

Cynthia Cheung, Richard White

Code 631, GSFC, Greenbelt, MD, 20771

Abstract. The eXtensible Markup Language (XML) is a powerful, internet-ready storage format for data. It provides a precise and clear data description language that is both machine- and human-readable. Documents marked up with XML can be easily parsed for efficient and focused searching. XML's expanded hyperlinking power can be used to enhance connectivity among archive documents. It is fast becoming the standard format for business and information transactions on the Internet and it is an ideal common metadata exchange format.

The Astronomical Data Center (ADC) plans to develop scripts to explore and display table data and metadata. The ADC is also developing XML authoring tools for scientists to convert their data and metadata into XML and to create conversion software for the ingest of journal tables into the ADC XML archive.

1. Introduction

The Astronomical Data Center (ADC)[2] at NASA's Goddard Space Flight Center has a 20-year history of preparing, documenting, validating, and distributing computer-readable versions of published astronomical data tables and catalogs. It maintains an extensive archive of such data and its services are heavily used by the community. There are two major problems facing the Astronomical Data Center (ADC).

- The ADC needs to prepare to manage the impending delivery of about 1,000 published scientific tables per year from the major astronomical journals. Documentation needs to be extracted from the articles to describe the tables and the tables must be converted to machine-readable format.
- At present, it is difficult for computer programs to parse the ADC's plain text documentation of catalogs and journal tables by information content

[1] U. of Maryland, Department of Physics, College Park, MD 20742

[2] `http://adc.gsfc.nasa.gov/`

(for example, the author or wavelength band). Precision searching and smooth navigation require intelligent information tagging.

The eXtensible Markup Language (XML) is an ideal solution to the major challenges above. XML is a powerful Internet-ready structured format for documentation (metadata) and data. It allows high-precision searching, enhanced web navigation, and seamless file exchange between users.

2. The problem of archiving and navigating thousands of tables

The ADC is very concerned with the need to improve the process of making published journal data tables computer-readable. However, the current journal submission and publication process is strongly oriented to the production of human-readable tables. The formats and organization of journal article submission do not fully support the automatic creation of data tables that can be easily read by data analysis software.

Although journal tables are available on the Web right now, their display formats are incompatible for loading the data into analysis or database environments. Some of these common problems are: 1) the inclusion of footnotes into the data 2) words, like 'glitch', embedded in columns of numbers 3) columns with multiple usage and 4) flags embedded with the numbers, like '10:'.

Probably the most difficult task is to extract sufficient information from the journal articles so that tables in the archive can be used correctly and intelligently. A single paragraph description of each table needs to be written (hopefully by the author) specifically for this use and complete information on the units, formats, blanks handling, etc. needs to be included in the metadata.

Given the expected growth rate, in just a few years, finding a particular datum from a table or even finding a particular table will require an unacceptable amount of effort. A way must be found to smoothly navigate between related tables, where the relationship can be defined by the user. A new way is needed to search very large archives, and it needs to be implemented in the broadest possible sense.

3. The wonders of XML

The essential concept of XML is to label, or 'mark up', the information content of a document such that computers can easily parse the document structure and retrieve the data and metadata it contains. A simple example of this is the following: `<author>` Ed Shaya `</author>`. Such tagging of content allows for focused and efficient automated scanning for information. This exceedingly simple concept, having been standardized in the metalanguage XML and accepted by the World Wide Web Consortium (W3C)[3] as a Recommendation in February 1998[4], is causing a revolution on the Internet, especially at inventory and data centers.

---

[3] `http://www.w3c.org/`

[4] `http://www.w3.org/TR/1998/REC-xml-19980210`

• **Defining the structure** – The rules for a set of XML documents are set down through the specification of a Document Type Definition (DTD). Since the structure can be fully prescribed, automated procedures can be used to validate the structure of the documents.
• **Hyperlinks** – XML has a companion schema called eXtensible Linking Language, XLink, (now a W3C working draft) for advanced hyperlinking between documents that allows for a much richer hypertext linkage. For example, XLink links can be type-cast by area of interest and users can choose to activate only specific link types. Another example is having a menu appear with multiple links to chose from when the mouse cursor goes over a certain set of words.
• **Navigation aids and vocabularies** – The Resource Description Framework (RDF) is an XML compatible schema that provides a robust and flexible architecture for supporting website descriptions and vocabulary definitions. Its primary purpose is to aid in locating data on the Web in a way that is understandable to the browsers. It could allow data centers to query each other about the nature, structure, and format of their holdings and services. The ADC needs tools with these features to properly manage the large volume of diverse tables and catalogs that will come with the new pipeline from the AAS Journals.

3.1. Some Details

XML is a version of SGML that will make the ingestion of the publisher's SGML marked up tables a straightforward process.
• **Defining the structure** – The rules governing how information is organized between the start and end tags are defined in a file called a Document Type Definition (DTD). The DTD rule for `<field>` might be:

```
<!Element field (name, description?, unit, format, blank?, chars?, min?, max?,
validMin?, validMax?)>
```

This declares that nested between the tags `<field>` and `</field>` there will necessarily be a name, unit and a format declaration. The other metadata attributes (description, unit, etc.) are optionally contained within field. This is signified by the question marks. The significance of the DTD is that it defines the order and organization of an acceptable archive document and XML provides checks that all documents conform to the rules.
• **Parse and display** – To translate an XML document into something that can be meaningfully displayed on an output device requires an intermediary. The, eXtensible Style Sheet (XSL)[5] is a full programming language for the display of XML documents. An XSL processor executes a series of rules which match tags or patterns and assigns actions to take place when a match is made. The assigned actions can be different depending on the particular device type or resolution that is used for display. One can choose various ways to display XML documents. Some examples are tree diagrams, directed acyclic graph (arrows), ASCII stripped of tags. One can direct the output to RTF (for Microsoft Word, WordPerfect, etc.) or direct it into database format (Excel, Access, Borland 123, etc.). In addition XSL can be used to translate XML into HTML which

[5] `http://www.w3.org/TR/WD-XSLReq`

can then be viewed with a web browser and be augmented with a Cascade Style Sheet (CSS).

• **Authoring tools** – We are examining automated methods for scientists to convert their data and metadata into ADC XML format for archive delivery. And, we are creating an SGML-to-XML application that will enable an automatic translation of tables from the major astronomical journal presses into the ADC XML archive. Since XML is derived from SGML, this should be straightforward, but certainly not trivial.

• **Image formats** – We need to investigate the pros and cons of translating standard image data to XML. It may be the case that image and observational raw data are best described by previously defined standards such as FITS, CDF, netCDF, and HDF. We may need to define XML-MIME types for each of these to make suitable use of these older standards as well.

• **Lists and arrays** – We must find how to best represent lists or tables of objects. Information like keywords, authors, and in fact the data tables themselves, can be tagged in several different ways: by individual item, list of items, or multi-dimensional array. We will find which way is optimal for each element, and will determine how to unambiguously delimit items within lists and arrays. Tables are a special issue because at this time there is not an XML standard table element. In our example DTD[6], we have accepted a simplified version of the `<array>` element from TecML (an application independent technical Markup Language written by the Open Molecular Foundation[7] to describe 2-D data.

4. Search tools

We can make use of off-the-shelf search tools as back ends to our website browsing tools. Search tools could be modified to be aware of our tag names and standard field names, and searches for these could be optimized by keeping up-to-date indices of occurrences. In fact, with all of the information well organized and clearly tagged, one can make great use of simple C language programs (like sgrep) that search for tags nested within tags and output their contents. It will become possible for users to bypass the ADC web tools and do searches based on their own software or intelligent agents. This is a major advantage because it allows users to analyze the data in new and innovative ways.

5. Conclusion

XML will probably be used in many areas within astronomy beyond catalogs and tables. It will take a united effort to make the most use of XML's power to organize and interoperate. We will be working with the AAS, the U. of Chicago Press, the AstroBrowse community, the Astrophysics Data Centers Coordinating Council (ADCCC), NED (NASA/IPAC Extragalactic Database), and our sister organization CDS to create a community standard for XML documents.

---

[6] `http://messier.gsfc.nasa.gov/xml/acatalog_dtd.html`

[7] `http://www.ch.ic.ac.uk/omf/`