Astronomical Data Analysis Software and Systems IX
ASP Conference Series, Vol. 216, 2000
N. Manset, C. Veillet, and D. Crabtree, eds.

Using XML for Accessing Resources in Astronomy

F. Ochsenbein[1], M. Albrecht[2], A. Brighton[2], P. Fernique[1], D. Guillaume[3], R. Hanisch[4], E. Shaya[5], A. Wicenec[2]

[1] *CDS, Observatoire Astronomique de Strasbourg, France*

[2] *ESO, Munich, Germany*

[3] *Univ. Illinois, USA*

[4] *STScI, Baltimore, MD, USA*

[5] *GSFC, Greenbelt MD, USA*

Abstract. XML—the Extensible Markup Language—is a developing standard in which the description of the data (the metadata) is included with the actual data in a single electronic document. This presentation focuses on the use of XML for accessing and understanding tabular data, particularly for handling the responses from queries to on-line catalog services. If such responses are encoded in XML using agreed upon tags and attributes, it is possible to both display the data in clearly formatted tables and use the data in other applications (such as generating graphical overlays of object positions on survey images). XML-encoded tables can also provide the basis for the next generation of data discovery and integration tools (Astrobrowse, ISAIA). The detailed definitions of XML tags and attributes, including examples and a working DTD, are described at http://vizier.u-strasbg.fr/doc/astrores.htx

1. The Challenge

The amount of astronomical data which can be accessed over the network is increasing at unprecedented rates, offering to the astronomer a lot of data, accessible from a lot of different servers all around the world, with a wide (should we say wild?) diversity of different *formats* and presentations. To answer to one of the most frequently asked question—*what does exist in the region of the sky I'm studying ?*—three steps are implied: (1) locate the potentially interesting servers in a changing world; (2) find out how to code the query parameters; and (3) interpret the results. Steps 1 and 2—what Hanisch (2000) names *good* and *better* in his presentation—could be automatized by tools like AstroBrowse[1] (McGlynn & White 1998), where the diversity of servers and ways to query them are coded in a distributed dictionary of URLs, the so-called *GLU*[2] (Fer-

[1] `http://heasarc.gsfc.nasa.gov/ab/`

[2] `http://simbad.u-strasbg.fr/glu/glu.htx`

nique et al. 1998). Our purpose here is now to tackle the third step: *interpret* the heterogeneous results coming from the numerous databases and services.

2. Improve the Data Understanding

Our perspective is not just to nicely display the heterogeneous results on a web browser, but to let a program interpret the results for further processing, e.g. for cross-correlation or visualization. A typical example is the visualization of sky images superimposed with the information coming from catalogue or archive services, like Aladin[3] or Skycat[4]: in order to overlay symbols showing the exact location of sources found in catalogues on top of images of the sky, the application has to query many resources for catalogued data existing in the target region *(steps 1 and 2)*, and then to find out, in the documents returned by the various servers, how the positions of the retrieved sources are coded *(step 3)*. This means that not only the actual numbers representing the coordinates have to be found, but also that minimal details about the coordinate representation (the *metadata*) have to be known (units, coordinate systems, accuracy, etc.).

For the Aladin application, the solution which was adopted up to now was to express the output of the main CDS data sources, VizieR[5] and Simbad[6], in a data format dedicated for this usage. The interpretation of the parameters provided by the NED database[7] was just based on a simple analysis of the HTML text coming out of the NED server—a solution which can easily break down if the NED layout happens to change. These two interfacing methods—develop a dedicated data format or try to interpret HTML documents—are not adapted for the long run, or impose constraints on the server side for no obvious benefit.

3. XML

XML *(e**X**tensible **M**arkup **L**anguage)* is basically a "superset" of HTML in which the various markup *"tags"* like `<TITLE>`*my title*`</TITLE>` in a *DTD (**D**ocument **T**ype **D**efinition)*. The basic advantage of XML is that the *same document* can either be parsed by simple-minded programs (XML uses hierarchical structuring), or can be displayed in the new generation of browsers (via an XSL style sheet which maps the DTD tags into typographical specifications) with the same capabilities as the HTML documents. There are other potential interests, especially regarding interoperability issues: many documents, each with its own DTD, can be analyzed with generic tools. It is a developing standard, proposed in many different contexts for astronomy—see http://pioneer.gsfc.nasa.gov/public/xml/

[3] http://aladin.u-strasbg.fr

[4] http://archive.eso.org/skycat/

[5] http://vizier.u-strasbg.fr/

[6] http://simbad.u-strasbg.fr/Simbad

[7] http://nedwww.ipac.caltech.edu/

HTML, the markup language in which the vast majority of the data servers are currently providing their results, has a frozen markup designed for a visual presentation of the documents (typically typographical specifications), forbidding therefore to have, e.g., a tag dedicated to mark the sky coordinates of the astronomical objects quoted in the document. A significant overhead is also generated by HTML documents, especially with the `<TABLE>` tag which is the natural way of presenting catalogued results.

On the other hand, FITS[8], also widely used in astronomy, includes a full support of tabular data. FITS tables however are used only in astronomy, and require specific tools to even read their contents; there are moreover two different formats for tables. However, if FITS cannot be included in XML documents, we took care of preserving a *compatibility with FITS*, in the sense that the definitions required for FITS can be included in our XML definitions.

4. The *astrores* Definitions

Figure 1 is an incomplete illustration of the basic layout of an *astrores* document; the full set of definitions and related DTD can be accessed at http://vizier. u-strasbg.fr/doc/astrores.htx . The basic tags describe *resources*, each resource being made of one or several *tables*, a table being made of a set of *fields* and their corresponding values. Several formats can be used in the actual data part, the one shown in Figure 1 being the character-separated-value *(CSV)* which has the advantage of being very compact, and quite easy to produce from a spreadsheet utility.

```
<ASTRO>
   <RESOURCE type="results|meta">
   <TITLE>The HST Guide Star Catalog, Version 1.2</TITLE>
      <TABLE>
         <FIELD name="_r"  unit="arcmin"  datatype="F" width="7"/>
         <FIELD name="GSC"  datatype="I"   width="10" />
         <LINK href="..."   content-role="query">${GSC}</LINK>
         <FIELD name="..."  unit="...">
         <DATA><CSV colsep="|"   headlines="2" ><![CDATA[
            _r |  GSC-id  |RA2000    |DE2000     |Err|Bmag|Err |m
         arcmin|----------|deg       |deg       |arcsec|mag|mag|
         0.0146|0430201297|003.25378|+72.522223|3.6|8.59|0.20|0
         0.9704|0430200545|003.20863|+72.513339|0.2|12.18|0.34|0
         ]]>
         </CSV></DATA>
      </TABLE>
   </RESOURCE>
</ASTRO>
```

Figure 1. Illustration of the basic *astrores* definitions.

A key point in the *astrores* definitions is that it is possible to ask for just a *description* of a resource in terms of tables, fields, queryable parameters, and

[8]http://fits.gsfc.nasa.gov/

the way to actually query it. The application receiving this description can complete the parameters described in the <`FIELD`> tags by itself, or can generate a *query form* to be filled. This facility is marked by the *meta* attribute in the <`RESOURCE`> tag. The <`LINK`> tag, which specifies how to actually submit a query or how to get more details about some incomplete result, is another important piece in the *astrores* definitions. It has the structure

<`LINK content-role="[query|hints|doc]" [href|action]="..." ...`>

where `content-role` indicates the possible meanings of the link:

- `query`: to generate a new query,
- `hints`: to learn more about the *metadata* or how to formulate a query—typically to get details about the valid *domains* of some parameter,
- `doc`: to get human-readable documentation.

and `href` may contain references to the actual contents of <`DATA`> with the `$` symbol, e.g. `href="program/center=${RA}${Dec}"`

5. Futures

A VizieR engine[9] is already producing XML results conforming to these *astrores* definitions; which can readily be used in the Aladin applet (Fernique & Bonnarel 2000); it is hoped that *astrores* or similar definitions could be adopted for other data services, allowing an immediate improvement of the interoperability of data servers through generic Java applications (see also the *Jsky* BoF).

Interoperability between many partners from different disciplines is far from easy (see the ISAIA[10] exercise): it means sharing common metadata, and adopting a common vocabulary. XML with its flexibility and growing usage is however a very good candidate for such developments.

A potential problem is however that, once the data can easily be used and reprocessed, their *origin* may completely disappear in the applications; the visibility of the original data servers, at least for minimal quality assessments, should be ensured in the various applications making use of the data.

References

Fernique, P., & Bonnarel, F. 2000, this volume, 71

Fernique, P., Ochsenbein, F., & Wenger, M. 1998, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 466

Hanisch, R. 2000, this volume, 201

McGlynn, T., & White, N. 1998, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 481

---

[9] `http://vizier.u-strasbg.fr/viz-bin/asu-xml`

[10] `http://heasarc.gsfc.nasa.gov/isaia/`