Using XML-Schema to model data from present and future astronomical databases

Bernard Debray

41 bis, avenue de l'Observatoire
BP 1615
25010 Besançon Cedex
France

- XML, the *e**X**tensible **M**arkup **L**anguage*, which was released in February 1998 by the W3 Consortium, was designed to permit the exchange of formatted informations over the World Wide Web in providing a standardized framework for the description of both data and metadata.
- XML has since then been used in the scientific community for both :
  - the management of information *documents* such as publications, for instance to ease the process of handling them in different formats ;
  - the modelling and exchange of scientific *data* (see e.g. Shaya *et al.*, 1999 [6], Ochsenbein *et al.*, 1999 [3]) ;
- Originally, XML documents make use of a *DTD* (Document Type Definition) to define their structure and syntactic organization.

Limitations of DTDs

- DTDs make use of a non-XML syntax (SGML), which does not allow to process them using XML tools (such as parsers).
- They offer limited possibilities to insert documentation.
- They offer limited data types.
- They offer limited possibilities for specifying data formats and constraints on the data.
- Extra informations about e.g. data formats or constraints have to be included in dedicated comments to be possibly processed by user programs, following the XML parsing process ; examples of this can be found in the *Astrores* DTD (*http://cdsweb.u-strasbg.fr/proj/astrores.htx*).

Excerpt of the *Astrores* DTD

```
<!-- Definition of a field -->
<!ELEMENT FIELD (NAME?, TITLE?, DESCRIPTION?, VALUES*, LINK? ) >
   <!-- ID       To refer this element
        unit     Unit used for the field (see appendix)
        datatype Datatype of field value
                 F-float, D-double, I-integer, A-ascii
                 L-boolean (logical), E-exponential
        precision Precision of field value: number of significant digits
                 after the dot (ex: "3")
        width    number of characters          "8"
        format   indicate the format by a "%fmt" template (as "printf()"
                 in language C). Use for coordinate format, time and date
                 formats   (ex: "%RAh:%RAmd %DEd:%DEmd" - (see appendix)
        ref      Reference to the system for this field. For example it can be
                 a coordinate or a photometric system.
                 ex : #id("myJ2000")  -> The syntax comes from Xpointers
        name     Alternative to specify the name of the field if it uses
                 only ascii characters (see the NAME element for the
                 other possibility)
        ucd      Unified column descriptor    "POS_EQ_RA" (ESO/CDS work)
        type     To characterize the field
                 hidden   for fields used typically for server/client exchange
                 no_query for fields which specify some parameters
                          (e.g. the equinox of a coordinate system)
                 trigger  for fields which contain a parameter for an action

        actualvalues Reference to the range or list of actual values
                 (it is an alternative to the <VALUES> entity in order to
                 be able to factorize them)
        legalvalues Reference to the range or list of legal values
                 (it's an alternative to the <VALUES> entity in order to
                 be able to factorize them)
   -->
```

Advantages of XML Schema for the handling of data

In May 2001, *XML Schema* 1.0 was released as a full W3C Recommendation ([9], [2], [7], [1]). Using XML Schema, one can now reflect completely the data types, constraints and structure of a data set using the XML syntax itself (see e.g. Williams *et al.*, 2000 [8]).

XML Schema allow :

- to handle the data model itself through standard XML parsers and softwares,
- to use XML parsers in the handling of the data up to higher level than with DTDs, before handing over the processing of the data to user programs ;
- complex data types can be defined from basic data types and other data types.
- data structures can be inherited ;
- as an XML structure, the data model can be included within the XML document itself.

To highlight the usefulness of XML Schema for handling and modeling data, here are some of the built-in XML Schema data types defined in the standard :

Type	Examples
string	
token	
byte	
unsignedByte	
base64Binary	
hexBinary	
integer	
positiveInteger	
negativeInteger	
long	-1, 12678967543233
short	-1, 12678
decimal	-1.23, 0, 123.4, 1000.00
float	
double	
boolean	true, false, 1, 0
time	13:20:00.000, 13:20:00.000-05:00
dateTime	2002-06-14T13:20:00.000-05:00
duration	P1Y2M3DT10H30M12.3S
Name	
QName (XML Namespace)	
anyURI	http://www.example.com/doc.html#ID5
language	en-GB, en-US, fr
ID	
IDREF	
IDREFS	
ENTITY	
ENTITIES	
NOTATION	
NMTOKEN	
NMTOKENS	

a complete reference can be found in http://www.w3.org/TR/xmlschema-2/#built-in-datatypes

Moving from DTDs to XML Schema

- As a first step to move forward to XML Schemas, and to avoid the tedious work of defining the schema from scratch, one can use conversion tools to translate DTDs into XML Schema. The conversion process must anyhow try to preserve as much as possible the information available in the DTD.

- The *dtd2xs* software was especially designed to preserve informations present in DTD comments, during the conversion process (Schweiger *et al.*, 2001 [5]).

- The resulting schema should nevertheless still be modified to draw all the potential benefit out of XML Schema.

Towards the use of XML Schema in an astronomical database

- At the Besançon Observatory, an XML output to data base queries following the Astrores DTD was implemented in the *BDB* binary stars data base (Oblak *et al.* [4]).
- To concretely assess the use of XML Schema, we are now testing the production of XML Schema validated XML output.
- As a first step, we converted the Astrores DTD into XML Schema using the *dtd2xs* software ; an excerpt of the resulting XML Schema file can be found below ; the complete file can be found at http://bdb.obs-besancon.fr/xml/astrores.xsd
- The *VOTable* format for representing a data table in XML was recently released in the framework of the developments for the Virtual Observatory initiatives (see *http://cdsweb.u-strasbg.fr/doc/VOTable/*) ; a draft version of an XML Schema description of the format was included in the document. We therefore intend to turn our developments into this format from now on.

Excerpt of the *Astrores* format for the description of tabular data converted into XML Schema

```
<xs:element name="FIELD">
    <xs:annotation>
      <xs:documentation xml:lang="en"> Note: table may be empty .. thus ================ FIELD*
     </xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:sequence>
        <xs:element minOccurs="0" ref="NAME" />
        <xs:element minOccurs="0" ref="TITLE" />
        <xs:element minOccurs="0" ref="DESCRIPTION" />
        <xs:element minOccurs="0" maxOccurs="unbounded" ref="VALUES" />
        <xs:element minOccurs="0" ref="LINK" />
      </xs:sequence>
      <xs:attribute name="ID" type="xs:ID" />
      <xs:attribute name="unit" type="xs:string" />
      <xs:attribute name="datatype">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="F" />
            <xs:enumeration value="I" />
            <xs:enumeration value="D" />
            <xs:enumeration value="E" />
            <xs:enumeration value="A" />
            <xs:enumeration value="L" />
          </xs:restriction>
        </xs:simpleType>
      </xs:attribute>
      <xs:attribute name="precision" type="xs:string" />
      <xs:attribute name="width" type="xs:string" />
      <xs:attribute name="format" type="xs:string" />
      <xs:attribute name="ref" type="xs:IDREF" />
      <xs:attribute name="name" type="xs:string" />
      <xs:attribute name="ucd" type="xs:string" />
```

```
    <xs:attribute name="type">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="hidden" />
          <xs:enumeration value="no_query" />
          <xs:enumeration value="trigger" />
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
  </xs:complexType>
</xs:element>
<xs:element name="VALUES">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="MIN" />
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="MAX" />
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="OPTION" />
      <xs:element minOccurs="0" ref="NULL" />
    </xs:sequence>
    <xs:attribute name="ID" type="xs:ID" />
    <xs:attribute name="multiple" default="no">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="yes" />
          <xs:enumeration value="no" />
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="type" default="legal">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:enumeration value="legal" />
          <xs:enumeration value="actual" />
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
  </xs:complexType>
</xs:element>
```

Conclusions

- XML Schema is now a standard and software tools handling it are rapidly emerging.
- Although they are “heavier” than DTDs in terms of overhead, XML Schema offer a standard way within the XML framework to formulate the models of scientific data, their types and constraints of values.
- The standardization of data formats inside the XML framework makes XML Schema and inescapable tool in the process of data validation when interweaving scientific databases ; it therefore appears as a useful standard for the development of the Astrophysical Virtual Observatory.

References

[1] Biron, P., V., Malhotra, A., *XML Schema W3C Recommendation, Part 2: Datatypes*, http://www.w3c.org/TR/xmlschema-2/ , May 2, 2001

[2] Fallside, D.C., *XML Schema W3C Recommendation, Part 0: Primer*, http://www.w3c.org/TR/xmlschema-0/ , May 2, 2001

[3] Ochsenbein, F., Albrecht, M., Brighton, A., Fernique, P., Guillaume, D., Hanisch, R. J., Shaya, E., & Wicenec, A., in ASP Conf. Ser., Vol. 216, *Astronomical Data Analysis Software and Systems IX*, eds. N. Manset, C. Veillet, D. Crabtree (San Francisco: ASP), p. 83, 2000

[4] Oblak, O., Debray, B., Kundera, T., Lastennet, E., Proceedings of SF2A annual meeting, EDP Science, 2002, *to be published*

[5] Schweiger, R., Hoelzer, S., Heitmann, K.U., Dudeck, D., *DTDs go XML Schema—a tools perspective*, Medical Informatics and the Internet in Medicine, Volume 26 (4), p. 297, October 1, 2001

[6] Shaya, E., Gass, J., Blackwell, J., Thomas, B., Holmes, B., & Cheung, C. Y., in ASP Conf. Ser., Vol. 216, *Astronomical Data Analysis Software and Systems IX*, eds. N. Manset, C. Veillet, D. Crabtree (San Francisco: ASP), p. 87, 2000

[7] Thompson, H., S., Beech, B., Maloney, M., Mendelsohn, N., *XML Schema W3C Recommendation, Part 1: Structures*, http://www.w3c.org/TR/xmlschema-1/ , May 2, 2001

[8] Williams, K., Brundage, M., Dengler, P., Gabriel, J., Hoskinson, A., Kay, M., Maxwell, T., Ochoa, M., Papa, J., Vanmane, M., *Professional XML Databases*, Wrox Press, 2000

[9] World Wide Web Consortium Press Release, Issues XML Schema as a W3C Recommendation, 2 May 2001, http://www.w3.org/2001/05/xml-schema-pressrelease