Title: XML Poster

Source: Australia

Author: Don Bartley, Australia

As part of the WG2 discussion on XML at the SC32 meeting Seoul on Friday 2 My 2002 Australia presented an XML poster produced by the Australian Bureau of Statistics (ABS). The ABS has created a series of posters to support the ABS Enterprise Architecture. The posters are living documents, expected to evolve in line with changing ABS needs.

As demonstrated in the meeting, the posters are supported on an internal portal. The portal provides further detail, easy access to text and diagrams for reuse in other documents, presentations and, for some, a streaming video presentation. Currently the posters include (those under development are marked *):

- ABS Enterprise Architecture
- IT Governance framework
- Business Process Taxonomy
- Maintaining and Using the Enterprise Architecture
- Applications Architecture
- IT Infrastructure “Ecosystem”
- Technologies and Toolsets
- Components and Services Interfaces
- Commercial Systems and Tools
- Security (*)
- Data Management (*)
- ABS DeveloperWorks (3 posters Context, Scenarios, Technology)
- ABS Input Data Warehouse (*)
- ABS Information Warehouse and Corporate Metadata Repository (*)
- Process Management
- Others………

The poster is part of the “Everybody should know” series within the ABS Enterprise Architecture poster set. This series includes a SOAP and Web Services Poster.

Attachment: XML Overview - ABS Poster
4 Is XML important? To whom?

- XML is the best attempt yet at a universal data exchange format
- It is standard, open and extensible, and widely used and supported
- It is important to any organisation that needs to exchange data internally or with other organisations or needs to be able to extract data from documents, amongst external organisations
- Increasingly all data flows between organisations will use XML
- It provides an easy method to agree on data formats
- Most commercial systems will handle generic XML and there are lots of tools to manipulate and consume XML
- Any other technical detail would be the organisations to particular systems or platforms
- It is just as relevant for use inside the organisation
- Availability of tools and flexibility to change and use in other systems is just as important
- Ideal archiving format - you can have a very high level of confidence about future usability
- XML is a key enabler of easy, reliable, "intelligent" data exchange
- For data capture, dissemination, for general business activity
- "Intelligent" - easy for users (ABS or its clients) to have their systems interact with internal or external systems (when the data is in XML)
- Stylesheets allow flexibility in presentation and use
- One XML document can have multiple stylesheets to serve multiple client users

5 Opportunities for ABS

- ABS should be promoting the development of a "StatsXML" and monitoring/influencing developments by other groups where we may have an interest
- We should be monitoring and understanding the XML vocabularies used in business and government and the data flows that they enable
- We might use them as data sources
- We should be liaising with our product-data providers (registrars, councils, etc) to agree standards for data delivery
- Even if multiple schemas are involved, stylesheets can make integration of data from multiple sources quite straightforward

6 What might be included in a "StatsXML"?

- Time Series, "Rich" Tables, Data Cuts, individual statistics
- Main Features, Explanatory Notes, entire publications
- Classifications, Statistical metadata
- Stylesheets for graphics, maps, charts
- Forms and other collection instruments

XML Lifecycle, showing schema, XML document, and multiple stylesheets for different outputs

1 What is XML?

- XML stands for eXtensible Markup Language
- The name, "Markup Language" is historical and comes from the printing terminology of "marking up a page"
- It is not very appropriate for XML, where describing layout is not the main focus
- XML is a language for describing structure, content, and layout of data in a format that is based on standards and can be used anywhere
- Its origins are in SGML (Structured Generalised Markup Language), a metalinguage to define different document types
- SGML was created by IBM and became an ISO standard in 1986 - permits "document processing by computer"
- A key feature is the separation of structure, content, and presentation
- XML is not a programming language
- But it does have some components (like stylesheets) that support reformatting and presentation of data
- XML is used to define both the rules for describing a particular data type (e.g. a Person type) and instances of the type (e.g. John Smith, Mary Jones)
- The definition of a data type is called a Schema and the XML language for defining types is called XSD - XML Schema Definition
- Many industry groups define schemas for data types they wish to exchange, e.g. XBRL
- XML involves only simple text, so that it is easy to transport and use anywhere
- No machine, platform, or language incompatibilities
- It is not a program! It does not "do" anything until some system (or person) reads it and acts on it

2 XML terminology and structure

- XML "Tags" are usually called "Documents" - even though most do not actually describe something we might think of as a document
- XML uses "Tags" (e.g. <Person>, "Surname") to identify parts of the data
- "closing" Tags start with "" - e.g. </Person>
- "Elements", enclosed by opening and closing Tags, are the basic units of an XML document e.g. <Person>John Smith</Person>
- Blanks between opening and closing tags are significant and part of the data
- Elements can be nested to describe a hierarchy
- Any node, machine, platform, or language incompatibilities
- Can be transmitted over almost any protocol - including http (standard web protocol) and https (secure web protocol)
- XML documents can be "transformed" by applying "stylesheets"
- Stylesheets are written in XSL (XSL), XSLT (XSL Transformations) schema
- A program (called a "Transformer") applies an XSLT stylesheet to an XML document to transform the document
- The in-memory representation for XML is called the Document Object Model (DOM)
- It is an object structure rather than a text stream and supports querying, updating, and navigation via a programmatic interface
- XPath is a language for querying and navigating the internal DOM document
- An XPath expression returns a list of matching elements
- <xsl:template match="Table"><HTML><Body><Table><Caption><xsl:value-of select="Title"/>

3 Stylesheets can transform a single XML document for many purposes

- Perhaps into different XML (conforming to a different schema)
- Perhaps into some other file format (e.g CSV) or into some specific format required by a particular system
- Perhaps into a presentation layout (eg as a table in HTML) or in a spreadsheet
- SVG (Scalable Vector Graphics) is an XML schema for drawing diagrams including maps

Suggested order to read this poster:

- A series of nested elements
- Nested to any depth
- Intended only for ease of reading
- Layout is not significant
- Repeated elements are valid
- Order of repeated elements is not significant
- Values (or attributes) must distinguish them
- Presents a textual representation of a complex hierarchy

7 What does XML look like?

| Table | <xsl:template match="Table">
| Cell | ... |
| Cell | ... |
| Cell | ... |

8 Getting value from XML is not primarily a technical matter

- The key is to develop and agree schemas for the data types of interest
- or to decide that existing schemas promoted by other groups are good enough to meet the requirement
- Developing good schemas for general use can only happen if the organisations in the community of interest agree on the concept to be described and the terminology to be used
- Any group of interested organisations can work together to develop and promote schemas
- W3C provides a forum for publicly agreed schemas and for looking at what has been agreed in different areas
- The only thing that can undermine the success of XML is the widespread development of competing and alternative schemas

9 XML is a standard

- XML is an open standard driven by the World Wide Web Consortium (W3C - a body that promotes internet interoperaibility)
- W3C defines all the basics - XML syntax, XSD, XML, XPath
- Any organisation or group can define their own XML vocabulary by defining and agreeing on a schema e.g. eqMutXML, XBRL (business reporting), eXSL (eXtensible business)
- Many industry groups have defined their own schemas - e.g. health care, transport
- XML is widely supported by the IT industry
- IT vendors are also defining XML schemas for use by their products (e.g. Microsoft has an XML for Excel spreadsheet, Lotus has XML for Notes documents and design elements)
- www.w3c.org and www.oasis.org/XML are World Wide Web Consortium sites
- www.iso.org "To the XML Industry Portal"
- A good site for looking at what industry groups are doing

10 Support for XML in the ABS environment

- The standard desktop includes XML support (parsers, transformers)
- Microsoft’s Internet Explorer, parses XML text and provides methods to create and manipulate the internal DOM
- XMLSpy is an XML editor and schema design tool
- Lotus Notes now contains design elements for handling XML with parsers and transformers to manipulate it
- Notes defines a Domino XML (DXL) for its own documents, items, and views
- DXL allows import and export of XML data from Notes
- XML support is starting to appear in Oracle and SAS
- Currently has limited support (eg to read XML)
- New products such as (VisualStudio.Net) provide comprehensive support