

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 May 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

Who is this poster for? What is its aim?

- This poster is for all ABS staff. All senior staff and most others need to know what XML is and where it might be important
- This poster aims to provide "All you need to know about XML" for most staff
 - no technical knowledge is assumed
 - if you understand this you should know whether you need to know more
 - there will be more detailed and technical posters for those who need or want more

Technical Business

Specific General

4 Is XML important? To whom?

- XML is the best attempt yet at a universal data exchange format
 - it is standard, but 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 flows 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 or generate XML and there are lots of tools to manipulate and consume XML
 - using any other technique would tie 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, for dissemination, for general business activity
 - "intelligent" - easy for users (ABS or its clients) to have their systems interact with internal data or data from other organisations (when the data is in XML)
 - stylesheets allow flexibility in presentation and use
 - one XML document can have multiple stylesheets to serve multiple client uses

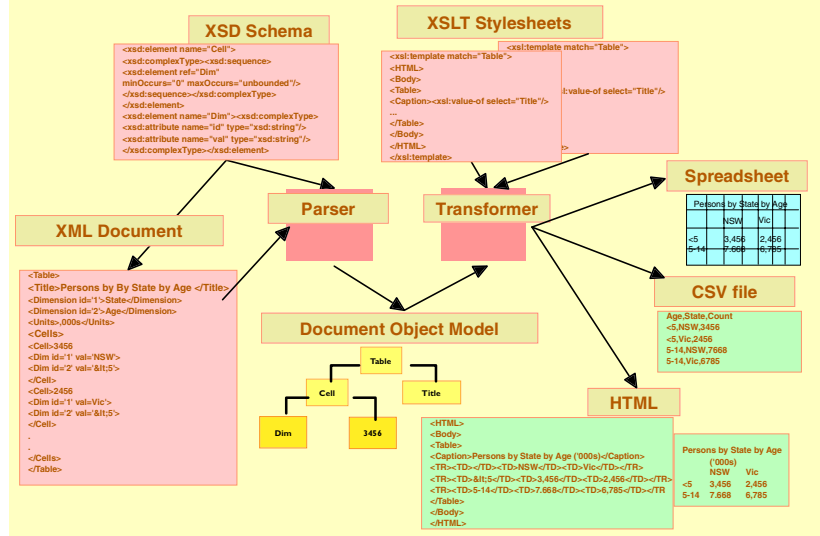
5 Opportunities for ABS

- 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 by-product data providers (registries, councils, etc) to agree schemas 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 Cubes, individual statistics
- Main Features, Explanatory Notes, entire Publications
- Classifications, Statistical metadata
- Stylesheets for graphs, maps, charts
- Forms and other collection instruments

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



ABS Enterprise Architecture

XML Overview

Ver 1.0

A poster in the "Everybody should know.." series

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 fashion that is based on standards and can be used anywhere
- Its origins are in SGML (Structured Generalised Markup Language), a metalanguage 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
 - it is heavy-weight and complex
- HTML (HyperText Markup Language - the original language of the internet) has the same origins
 - HTML is a markup language - it focusses on layout and does not preserve the separation of structure, content, and presentation
 - HTML was standardised "after the event" and implementations are sloppy - HTML is a mess!
- 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 (eg a Person type) and instances of the type (eg 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, eg XBRL
- XML involves only simple text, so that it is easy to transport and use anywhere
 - no machine, platform, or language incompatibilities
 - can be transmitted over almost any protocol - including http (standard web protocol) and https (secure web protocol)

XML is just "data"

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 "files" are usually called "Documents" - even though most do not actually describe something we might think of as a document
- XML uses "Tags" (eg <Person>, <Surname>) to identify parts of the data
 - "closing" Tags start with "/" - eg </Person>
 - "Elements", enclosed by opening and closing Tags, are the basic units of an XML document eg <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
 - and can have "Attributes" (eg. HeadPerson="True") for aspects of the data
 - the box at right shows nested elements and an attribute
- A "Schema" defines the rules for the XML for a data type (eg for Person) - sometimes called a "Vocabulary"
 - schemas are themselves written in XML using the XSD (XML Schema Definition) schema
 - ie XSD defines the rules for writing schema definitions
 - a schema defines an XML vocabulary for a particular business purpose
- XML documents can be "transformed" by applying "stylesheets"
 - stylesheets are written in XML using the XSLT (XML Style Language Transformations) schema
 - a program (called a "Transformer") applies a XSLT stylesheet to an XML document to transform the document(s)
- 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
 - eg //Person[@HeadPerson="True" and Surname="Smith"] would return the head person of all the Smith families
- An XML Parser reads a file containing XML, interprets it using the designated schema, and produces the internal DOM format
 - parsers and transformers are available for almost all platforms and languages (and are usually free!)
- SVG (Scalable Vector Graphics) is an XML schema for drawing diagrams including maps

Some simple XML

```
<Person HeadPerson="True">
  <FirstName>John</FirstName>
  <Surname>Smith</Surname>
</Person>
```

3 Stylesheets can transform a single XML document for many purposes

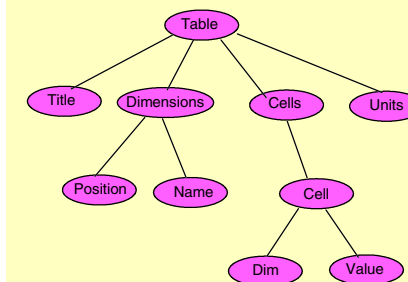
- perhaps into different XML (eg conforming to a different schema)
- perhaps into some other file format (eg CSV) or into some specific format required by a particular system
- perhaps into a presentation layout (eg as a table in HTML or in a spreadsheet)
- perhaps for a mobile phone display, for handicapped access, or into another language

Suggested order to read this poster?

- Start in the middle under the banner, then go down the left-hand side, then the right - follow the numbers!
- the middle panels provide the basic introduction, the others elaborate the theme
- but there is no "right" order - hop around if you wish

7 What does XML look like?

- A series of nested elements
 - nested to any depth
 - indented only for ease of reading
 - layout is not significant
 - repeated elements are valid
 - values (or attributes) must distinguish them
 - order of repeated elements is not significant
- Presents a textual representation of a complex hierarchy



```
<Table>
  <Title>Population</Title>
  <Dimensions>
    <Dimension>
      <Position>1</Position>
      <Name>State</Name>
    </Dimension>
    <Dimension>
      <Position>2</Position>
      <Name>Year</Name>
    </Dimension>
  </Dimensions>
  <Units>Thousands</Units>
  <Cells>
    <Cell>
      <Dim Id=1>NSW</Dim>
      <Dim Id=2>1991</Dim>
      <Value>5,195</Value>
    </Cell>
    <Cell>
      <Dim Id=1>NSW</Dim>
      <Dim Id=2>1996</Dim>
      <Value>6,024</Value>
    </Cell>
    ...
  </Cells>
</Table>
```

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 publicising agreed schemas and for looking at what has been done by others
- 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 interoperability)
 - W3C defines all the basics - XML syntax, XSD, XSL, XPATH
 - any organisation or group can define their own XML vocabulary by defining and agreeing on a schema
 - eg MathXML, XBRL (business reporting), ebXML (electronic business)
 - many industry groups have defined their own schemas - eg health care, transport
- XML is widely supported by the IT industry
 - IT vendors are also defining XML schemas for use by their products (eg Microsoft has an XML for Excel spreadsheets, Lotus has XML for Notes documents and design elements)
 - www.w3.org and www.w3.org/XML are World Wide Web Consortium sites
 - www.xml.org is "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 basic XMLsupport (parsers, transformers)
 - Microsoft's MSXML 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 to/from Notes
- XML support is starting to appear in Oracle and SAS
 - currently have limited support (eg to read XML)
- New products (such as VisualStudio .Net) provide comprehensive support