SEMIDE

EMWIS

SEMIDE

http://www.semide.org

EMWIS

http://www.emwis.org

An introduction to XML

F. Fedele & A. Pischedda

EMWIS Technical Unit

EUROMEDITERRANEAN PARTNERSHIP

SEMIDE
EMWIS

The XML

- (eXtensible Markup Language)

SEMIDE
EMWIS

The creation of XML

- A standard proposed by the World Wide Web Consortium (W3C)
- eXtensible Markup Language
  - Language – it is a language
  - Markup – instructions controlling the content and information aspect
  - eXtensible – users may define a set of tags according to their requirements
- A TAG based text format
  - Separates contents and presentation
  - It is responsible for data definition only

SEMIDE
EMWIS

The markup concept

- The electronic production of texts has been solved by the use of two different techniques:
  - word processors
  - markup languages
- Word processors are WYSIWYG (What You See Is What You Get) systems that are limited insofar as portability and re-use are concerned
- On the other hand, markup languages describe the structural and representative mechanisms of the text by means of real languages that often use standardised conventions, thus usable on a greater range of systems
- A markup language called SGML (Standard Generalized Markup Language) has existed for years
- A new markup language was created for the web: HTML (HyperText Markup Language) – the web was born and documents were “published”

SEMIDE
EMWIS

Separating the contents from the presentation

- Characteristics of HTML
  - guarantees the correct use of the information due to the distinction made between structure and presentation
  - easy to use
  - adapted to basic publishing
  - the instructions used do not require the browser to carry out a validation test on the hierarchical structure of the information
- Limits of HTML
  - created as a markup language for making documents available on line, it has become a technology assembler capable of meeting the requirements of the “net”
  - it is not extensible
  - it is display-centred
  - it is not, in general, re-usable
  - it has little or no semantic structure
  - it is rigid, supplying a single “view” of the data

SEMIDE
EMWIS

The evolution of the Web

The development poses problems of integration between heterogeneous data and data in various formats:

- management of integration based on the logical display of the data and not through its physical implementation
- describing, in a standardised manner, the logical display of the information available
- Creating interactions between the site and other sites (server/server), not only with the client (server/client)

SEMIDE
EMWIS

The need for a substitute for HTML

- New Internet requirements have generated the need to evolve away from HTML
  - SGML (Standard Generalized Markup Language) does exist, but it is too complex and was not designed for use on the web
  - HTML (HyperText Markup Language) is not data-oriented, it is only adapted to Web page presentations
- A tool capable of facilitating data exchange is required
  - a markup language with 80% of SGML's functionalities and 20% of its complexity
  - adapted to document processing and web publishing but also to the development of applications
  - similar to HTML, but more flexible and data-oriented

SEMIDE
EMWIS

The creation of XML

- It is a direct descendant of SGML
  - Simplified for the Web
  - HTML is also a SGML subset
- Rigorous, but easy to interpret
- It can represent:
  - traditional record-structured relational data
  - hierarchical data
  - unstructured data
- XML separates contents and presentation

SEMIDE
EMWIS

XML slightly resembles HTML but is not HTML

- Like HTML, XML uses tags (words comprised between the less than and greater than signs ('<' and '>') and attributes (in name format="value"), however, while HTML specifies the signification of each tag and attribute (and often what the text contained between them will look like in a browser), XML only uses tags to delimit parts of information, leaving the interpretation of that information entirely to the application reading it
- In other words, if a "<p>" tag appears in an XML file, it cannot be assumed that it represents a paragraph. Depending on the context it may be a price, a parameter, a person, a p….
  (by the way, who said that it should be a word starting with "p"?)

SEMIDE
EMWIS

From HTML to XML

```
<!--Example of HTML  -->
<h1>Invoice 01/00</h1>
<p>Effedue Consulting
<p>Date: 31 October 2000
<br><br>
<p>IBM SpA
<br>
<p>Amount: GBP 2,000
<p>VAT: 20%
<p>
<b>Total: GBP 2,400
```

```
<!-- Example of XML  -->
<Invoice>
<Number>01/00</Number>
<Emitted by>Effedue
    Consulting</Emitted by>
<Date year = '2000' month =
    '10' day = '31' />
<To>IBM</To>
<Amount value = 'GBP'>2000
    </Amount>
<%VAT>20</%VAT>
<Total value =
    'GBP'>2400</Total>
</Invoice>
```

SEMIDE
EMWIS

Structuring an XML document

- Since XML is an extensible language, the user must create the rules establishing
  - what data should be included in an XML document
  - what data is considered necessary or optional
  - how the data should be interpreted
- Rules are grouped together in an appropriate document
  - the conformity of the XML documents subjected to such a structure is verified
- Determining whether there is any missing data is easy
  - thus guaranteeing accuracy in a catalogue
  - imposing links to the format of the data is possible

SEMIDE
EMWIS

Two alternatives: DTD or XML Schema

Capable of defining grammatical specifications for applications

- DTD (Document Type Definition)
  - grammar for tag and attribute descriptions
- XML Schema
  - Richer grammar for XML-based descriptions

SEMIDE
EMWIS

DTD (Document Type Definition)

DTDs were defined at the time of SGML

- They are used to define the syntax of a document and define:
  - tags that may or must be present in the document
  - how the tags are to be nested
  - how often a given tag may be repeated in the document
  - tag attributes and their default values
  - all valid values for given attributes
- They may semantically define a group of data specifying the different possibilities allowed and when these are possible
- They are written in a specific language (not XML)

SEMIDE
EMWIS

XML Schema

- XMLSchema is a mode for the description of the structure, contents and type of the XML data
- It extends the functionalities of the DTDs and solves some of their problems
- It allows greater control of the XML file elements, their contents, the data types and their value
- Like DTD it may be used to validate XML
- Unlike the DTDs, XMLSchema uses an XML syntax

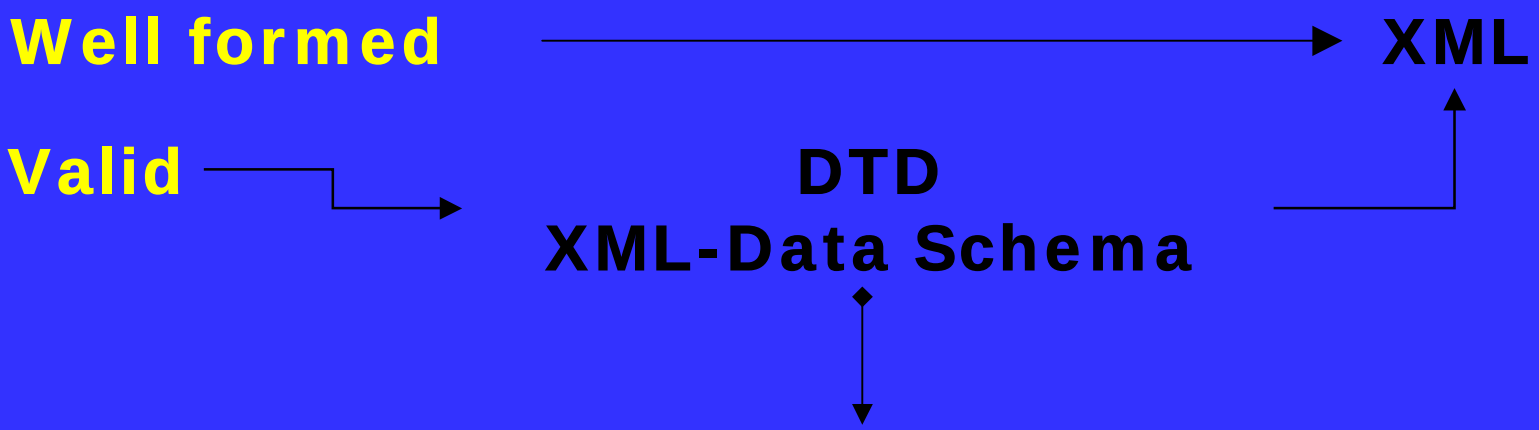
SEMIDE
EMWIS

Valid documents

- Well formed documents using a Document Type Definition and conforming to all DTD rules are also “valid” documents

SEMIDE
EMWIS

Parsing




```
<?XML version="1.0" encoding="UTF-8"?>
<!DOCTYPE book SYSTEM "automobile.dtd">
```

SEMIDE
EMWIS

Parser - base concepts

- XML Parsing
  - XML is a metalanguage particularly adapted to data exchange
  - it is possible to consider an XML document as a simple text document, and thus to process it as such
  - XML may represent:
    - relational data
    - hierarchical data
    - unstructured data
  - XML separates the contents from the presentation
  - It is not for use with the WEB only

SEMIDE
EMWIS

XSL

- eXtensible Stylesheet Language (XSL) is a W3C tool that is still in development aiming at creating a standard in the domain of XML document formatting
- XSL is an extension of the HTML style sheet concept (CSS)
- Includes XSL-FO for formatting specifications
- It allows the coding of the style rules, e.g. “all titles must be in bold type”. XSL thus guarantees homogeneous formatting of all XML documents of the same type. If the need to modify the formatting arises, all that is necessary is to modify the associated XSL

SEMIDE
EMWIS

XSLT

- XSL Transformations was, for a long time, only a part of XSL and has recently become an independent language allowing the transformation of an XML type document into a different type
  - XSL-T now constitutes the first part of the standard
  - the second part (XSL) defines the semantics of the formatting
- Allows the handling of a document
- Using XSLT it is possible to generate:
  - an HTML document
  - another XML document with a different structure
  - n audio document
  - … almost any other document type

SEMIDE
EMWIS

XML in the WEB

- http://www.w3c.org/xml
- Sites on XML
  - http://www.xml.org
  - http://www.xml.com
- Sites on initiatives linked to XML
  - www.biztalk.org
  - www.ebxml.org
  - www.goxml.com
  - www.ariba.com
  - www.xmledi.com
  - www.xmlglobal.com
- Software
  - www.xmlspy.com

SEMIDE
EMWIS

Tank you for your attention